UNIVERSITAT DE LES ILLES BALEARS

FACULTAD DE PSICOLOGÍA

TESIS DOCTORAL

EVALUACIÓN EN EL MODELADO DE RESPUESTAS DE RECUENTO

NOELIA LLORENS ALEIXANDRE

DIRECTOR: DR. ALFONSO PALMER POL

PALMA DE MALLORCA, ABRIL 2005

Este trabajo esta dedicado a Miguel J. Perelló.
Mi apoyo y fuente de inspiración.

"Los hechos no dejan de existir aunque se los ignore."

Aldous Huxley

INDICE

Prólogo

La presente Tesis Doctoral ha sido realizada en el Área de Metodología de las Ciencias del Comportamiento, perteneciente a la Facultad de Psicología de la Universidad de las Islas Baleares.

Dentro de las líneas de investigación del Área encontramos la que concierne al estudio del Modelo Lineal Generalizado y al análisis de datos en el campo de las conductas adictivas.

La motivación inicial de los trabajos de investigación fue el estudio de los datos de recuento. La escasez de información en la evaluación de modelos adecuados para datos de recuento así como el uso incorrecto del análisis que se aplica, en el campo sustantivo, a este tipo de datos, nos llevó al planteamiento de la presente tesis.

Planteamos el estudio metodológico de la etapa de evaluación en datos de recuento, con especial énfasis en la evaluación de la equidispersión, y su aplicación en datos reales obtenidos en nuestras investigaciones en el campo de las conductas adictivas..

Como fruto de esta labor investigadora desarrollada por nuestro equipo, se han publicado, o se está en vías de publicación, 7 trabajos en diversas revistas científicas y se han presentado diferentes trabajos en dos congresos de Metodología. Esta tesis trata de aglutinar los logros alcanzados en este conjunto de trabajos, en los que se pone de manifiesto la utilidad de la etapa de evaluación para un correcto análisis de los datos.

Noelia Llorens Aleixandre

Universidad de las Islas Baleares, Facultad de Psicología

Palma de Mallorca, Abril de 2005.

I.- INTRODUCCIÓN

La presente tesis se centra en el estudio de la etapa de evaluación en el modelado de datos de recuento. El modelo de referencia para este tipo de datos es el modelo de regresión de Poisson (MRP) y un aspecto fundamental es el supuesto de equidispersión, un incumplimiento del cual desemboca en sobredispersión (en menos ocasiones en infradispersión). Es por ello que la etapa de evaluación en datos de recuento hace de la comprobación de la equidispersión un pilar entorno al cual girar. Esta ha sido la principal motivación del compendio de publicaciones que aquí se presentan.

Para forjar el entorno adecuado a la etapa de evaluación, en la introducción describiremos en primer lugar el Modelo Lineal Generalizado (MLG) con sus etapas y el modelo de regresión de Poisson. En la etapa de evaluación estudiaremos los residuales, los gráficos y la sobredispersión.

La etapa de evaluación recoge mucha información desde diferentes ángulos. Cabe decir que al utilizar los modelos de datos de recuento en los artículos de carácter más aplicado en el campo de las conductas adictivas, nos encontramos con el problema de que los *referees* de dichas revistas “aconsejaban” que se disminuyera, o incluso que se eliminara, gran parte de las explicaciones referentes a los recursos metodológicos utilizados. Esto obligaba a resumir en tres frases todo el proceso de evaluación. Por ello planteamos aquí realizar una introducción en la que se expusieran la mayoría de las pruebas, índices diagnósticos, residuales, etc., de forma que constituyera una taxonomía a partir de la cual entender lo que se puede y se debe hacer en esta fase tan importante de los trabajos prácticos.

Como anexo se adjunta un estudio en el que se comprueba, a modo de ejemplo, el funcionamiento de los residuales, gráficos e índices diagnósticos, en la etapa de evaluación.

En cuanto a los trabajos que componen la presente tesis, son los siguientes:

Estudio 1: “Overdispersion diagnostics in count data analysis models” (Vives, Losilla y Llorens, en revisión) que constituye un primer paso en el estudio de la etapa de evaluación como es el conocer aquellas pruebas que nos informan de la sobredispersión. Este estudio pretende recoger y evaluar las diferentes pruebas diagnósticas de sobredispersión.

Ante la situación de sobredispersión disponemos de dos opciones principales a seguir, o bien corregir los errores estándar o bien modelar los datos con otros modelos.

En el caso de querer corregir o estimar el error estándar, nos encontramos con la necesidad de reunir en un mismo estudio todos aquellos índices y estimadores de los errores estándar utilizados para corregir el problema de la sobredispersión. La literatura nos mostraba gran variedad de índices y estimadores pero en ningún momento se comparaban todos bajo las mismas circunstancias. Con el estudio 2: “Ajuste y estimación de los errores estándar de los parámetros del modelo de regresión de Poisson en presencia de sobredispersión” (Llorens, Palmer y Losilla, 2004) nos planteamos la realización de un estudio de simulación en el que pudiéramos comparar los principales índices y estimadores en diferentes situaciones de sobredispersión.

Ante la similitud de comportamiento de estos índices y estimadores consideramos de interés la ampliación de los estimadores incluidos en la investigación previa, así como el análisis específico de éstos. Para ello realizamos el estudio 3 “Overdispersion in the Poisson regression model: A comparative simulation study” (Llorens, Palmer, Losilla y Vives, en revisión).

En lugar de corregir los errores estándar o estimarlos, se puede optar por la vía del modelado. En este sentido es importante comprobar las diferencias que se producen al ajustar diferentes modelos estadísticos a datos de recuento con sobredispersión. El 4º estudio: “Modelado del número de días de consumo de

cannabis" (Palmer, Llorens y Perelló, en prensa), compara las diferencias de ajuste de cuatro modelos: el modelo de regresión lineal, el modelo de regresión de Poisson, el modelo de regresión de la Binomial Negativa y el modelo de regresión de ceros aumentados (ZIP).

Cuando disponíamos de soluciones para la sobredispersión, decidimos aplicar el modelo de regresión de Poisson a datos de recuento obtenidos en el campo de las conductas adictivas. Posteriormente estudiaríamos la etapa de evaluación, en la que podríamos evaluar la adecuación del modelo aplicado. En el 5º estudio "Las estrategias de afrontamiento: factores de protección en el consumo de alcohol, tabaco y cannabis" (Llorens, Perelló y Palmer, 2004) se comprobó el cumplimiento de la equidispersión en cuatro modelos, un modelo cumplió el supuesto, por ello se le aplicó el MRP. Los otros tres modelos no cumplieron el supuesto de equidispersión por ello se procedió a la corrección del EE.

En el estudio 6º, "Activity levels and drug use in a sample of Spanish adolescents" (Llorens, Perelló y Palmer, 2005) después de comprobar la existencia de sobredispersión se aplicó el MRBN, por ser el que presentaba un mejor ajuste en función de las características de los datos.

En el estudio 7º "Características de personalidad en adolescentes como predictores de la conducta de consumo de sustancias psicoactivas" (Llorens, Palmer y Perelló, 2005), nos encontramos con datos de recuento con sobredispersión en los que se encontraban un exceso de ceros debido a la existencia de un doble proceso de aparición de los ceros. La evaluación del modelo señala al modelo ZIP como modelo de elección.

1. El Modelo Lineal Generalizado (MLG)

La teoría del Modelo Lineal Generalizado fue desarrollada y presentada por Nelder y Wedderburn (1972). Ellos descubrieron la conexión de cierto tipo de modelos de regresión, modelos cuyas variables de respuesta eran miembros de la familia exponencial de distribuciones. Incluyeron en esta familia distribuciones tales como la Gausiana o Normal, Gamma, Poisson, Geométrica y Binomial Negativa.

Estos autores mostraron que si se relajaban las asunciones del modelo lineal general, se podían desarrollar modelos más generales. Así, reestructurando la relación entre el predictor lineal y el valor ajustado, se podía modelar relaciones lineales que previamente eran consideradas no lineales. Como señalan Hardin y Hilbe (2001) a los modelos que permitían este tipo de reestructuración se acordó llamarlos “Modelos Lineales Generalizados”. Nelder y Wedderburn linealizaron cada uno de los miembros de la familia de MLG por medio de una función de enlace.

1.1. El modelado estadístico

El modelado estadístico es una herramienta fundamental en el estudio de la variabilidad de un conjunto de datos observados, una formalización de la variabilidad observada en la que se distinguen dos elementos, la variabilidad sistemática y la variabilidad aleatoria (Lindsey, 1995) y responden, por tanto, a la siguiente expresión:

Variable de respuesta = componente sistemático + componente aleatorio

El componente sistemático resume cómo la variabilidad en la respuesta es explicada por los valores de ciertas variables o niveles de ciertos factores y es descrita, generalmente, mediante un modelo de regresión. El componente sistemático del modelo describe una respuesta «ideal» y, por tanto, para

considerar las fluctuaciones en la respuesta debemos incluir un componente probabilístico en el modelo, denominado componente aleatorio o residual que mediante una distribución de probabilidad describe en qué medida la variable de respuesta observada se desvía de la respuesta esperada a partir de la parte sistemática del modelo. Una forma paralela a la anterior de especificar la relación entre componente sistemático y aleatorio es mediante la ecuación:

$$\text{Datos} = \text{Modelo} + \text{Error} \qquad (1)$$

donde el modelo se corresponde con la variabilidad de los datos explicada por el componente sistemático y el error es la variabilidad no explicada o componente aleatorio del modelo, y representa la discrepancia observada entre los datos observados y los predichos por el componente sistemático del modelo.

La definición estructural de los MLG se puede expresar como:

$$g(y) = f(x) + \varepsilon \qquad (2)$$

siendo *g(y)* una función aplicada sobre las variables respuesta, denominada *función de enlace*; *f(x)* una función aplicada sobre las variables explicativas con el objeto de establecer una relación estadística con la/s variable/s respuesta/s (componente sistemático); ε el término de error, o diferencia entre las dos funciones establecidas anteriormente (componente aleatorio).

1.2. Etapas del modelado

1.2.1. Especificación del modelo

El objetivo de esta etapa es seleccionar, de entre el conjunto de modelos posibles, aquellos más relevantes para describir las principales características de las variables de respuesta. Este proceso implica tomar decisiones que conciernen a la formulación del componente sistemático, los supuestos sobre el componente aleatorio y cómo los dos componentes son combinados en el modelo. La

especificación del modelo está sustentada tanto en la teoría sustantiva como en la observación de los datos. A este respecto Box, Hunter y Hunter (1988) señalan que «la identificación es un proceso informal en el que la construcción de gráficos, el análisis preliminar de los datos y la reflexión sobre las relaciones entre los elementos fundamentales del sistema a modelar se emplean para llegar a una clase de modelos que valga la pena considerar más profundamente».

1.2.2. Estimación y ajuste

Tras la especificación de un modelo particular se requiere estimar los parámetros del componente sistemático del modelo y valorar la discrepancia entre los datos observados y los ajustados por el modelo. Como señala Ato y López (1996) «el proceso de ajuste de un modelo puede considerarse como una forma de comparar un conjunto de valores empíricos observados con un conjunto de valores ajustados de un modelo que implica usualmente un número menor de parámetros». A menudo varios modelos estadísticos compiten por explicar los datos, y la comparación de modelos anidados es el procedimiento utilizado para obtener el modelo más parsimonioso que reproduzca mejor los datos observados. Por último, en esta fase también es fundamental la estimación de intervalos de confianza para los parámetros del modelo con el objetivo de clarificar la eficacia predictiva e interpretabilidad del mismo.

1.2.3. Evaluación del modelo

La evaluación del modelo supone valorar si el modelo ajustado en la etapa anterior es un modelo válido, más allá de que presente un ajuste adecuado a los datos. La evaluación o diagnóstico del modelo se refiere a la adecuación de los aspectos implicados en la etapa de especificación. En este sentido se han de evaluar posibles errores de especificación del componente sistemático, de la distribución de probabilidad del componente aleatorio y de la relación asumida entre ambos componentes del modelo en la fase de especificación. Por otra parte se requiere evaluar la presencia de observaciones extremas (*outliers*) o influyentes

(*influentials*), así como el conjunto completo de elementos que forman los supuestos bajo los cuales se valida el modelo.

1.2.4. Interpretación

En esta etapa, que cierra el proceso de modelado, una vez seleccionado el modelo óptimo en función de los criterios de bondad de ajuste y parsimonia, y una vez contrastado que el modelo es válido, se ha de proceder a su interpretación e integración en el marco teórico desde el que fue propuesto, esto es, se requiere retornar al nivel teórico-conceptual del proceso metodológico.

2. El modelo de regresión de Poisson

La distribución de Poisson debe su nombre al matemático francés Siméon Denis Poisson, quien publicó en 1937 un trabajo de investigación en que presentaba una nueva distribución para el cálculo de probabilidades aplicado al ámbito penal. *«En recherchés sur la probabilité des jugements…»* (Poisson, 1837). Posteriormente y a raíz del desarrollo de los modelos lineales generalizados, el modelo de regresión de Poisson apareció como un caso especial de estos modelos, descrito por Nelder y Wedderburn (1972) y detallado en MCullagh y Nelder (1989). Contribuyeron a su construcción los trabajos de Gourieroux, Monfort y Trognon (1984 a, b) y de Hausman, Hall y Griliches (1984).

La distribución de Poisson es el modelo de referencia para datos de recuento (Cameron y Trivedi, 1986, 1990; Gurmu, 1991; Lee, 1986; Lindsey, 1998). La ley de eventos raros establece que el número total de eventos seguirá, aproximadamente, una distribución de Poisson si un evento puede ocurrir en cualquier punto del tiempo o del espacio bajo observación, pero la probabilidad de ocurrencia en un punto determinado es pequeña (Cameron y Trivedi, 1998). Es decir, los datos de recuento de fenómenos con una baja probabilidad de ocurrencia (sucesos raros) siguen una distribución de probabilidad conocida, denominada distribución de Poisson.

La distribución de Poisson permite obtener la probabilidad de que se produzca un número determinado k de ocurrencias de un evento:

$$\pi_i = P(y_i = k) = \frac{exp(-\mu)\mu^{y_i}}{y_i!} \qquad (3)$$

donde $\mu > 0$ es el parámetro media de la distribución, que coincide con el valor de la variancia, lo que define la propiedad de “equidispersión”.

El modelo de regresión de Poisson (MRP) presenta una estructura simple y puede ser fácilmente estimado (Greene, 2000; Lee, 1986). Sin embargo, esta simplicidad

es el resultado, como señala entre otros Sturman (1999), de algunas limitaciones en sus asunciones, el incumplimiento de las cuales tienen efectos sustanciales en la eficiencia de los coeficientes del modelo. La crítica más notable al modelo de regresión de Poisson es la asunción de que la media de y_i es igual a su variancia, este supuesto en la mayoría de las ocasiones no es realista. Si esta condición no se satisface, aparecen en general datos sobredispersos (overdispersed data) aunque podrían ser también datos infradispersos (underdispersed data). En esta situación, al ajustar el modelo de regresión de Poisson se obtiene una infraestimación de la matriz de covariancias de los parámetros de regresión produciendo una sobreestimación de los valores de la prueba de conformidad de los parámetros y por tanto de su significación (Liao, 1994).

A lo largo de los años han ido apareciendo gran cantidad de pruebas, como señalan Karlis y Xekalaki (2000), cuya función es comprobar las asunciones de la distribución de Poisson (tabla 1 y 2).

Tabla 1. Pruebas de evaluación de las asunciones de la distribución de Poisson.

Nombre de la prueba	**Referencia**	**Región Crítica**	**Alternativa**	**Prueba estadística en tabla 2**
VT	Cochran (1954)	Depende de la alternativa	Sobredispersión o infradispesión	1
Böhning	Böhning (1994)	Depende de la alternativa	Sobredispersión o infradispesión	2
Zelterman	Zelterman (1988)	Depende de la alternativa	Mezclas	3
2nd cumulant	Gart y Pettigrew (1970)	Depende de la alternativa	Sobredispersión o infradispesión	4
3rd cumulant	Gart y Pettigrew (1970)	Ambas colas	No Poisson	5
4^{th} cumulant	Gart y Pettigrew (1970)	Ambas colas	No Poisson	6
Kocherlakota with t=-0.05	Kocherlakota y Kocherlakota (1986)	Ambas colas	No Poisson	7
Kocherlakota with t= 0.05	Kocherlakota y Kocherlakota (1986)	Ambas colas	No Poisson	7
Kocherlakota with t=0.125	Kocherlakota y Kocherlakota (1986)	Ambas colas	No Poisson	7
Nass	Nass (1959)	Ambas colas	No Poisson	8

Tabla 1(continuación). Pruebas de evaluación de las asunciones de la distribución de Poisson.

Gupta	Gupta , Mori y Szekely (1994)	Cola izquierda	Distribución divisible infinita	9
Baringhaus	Baringhaus y Henze (1992)	Cola derecha	No Poisson	10
Nakamura	Nakamura y Perez-Abreu (1993)	Cola derecha	No Poisson	11
Rueda	Rueda , Perez-Abreu y O'Reilly (1991)	Cola derecha	No Poisson	12
KS	Campbell y Oprian (1979)	Cola derecha	No Poisson	13
McIntyre	Rayner y McIntyre (1985)	Cola derecha	No Poisson	14
χ^2	Cochran (1954)	Cola derecha	No Poisson	15
Power divergent family, $\lambda = 2/3$	Read y Cressie (1988)	Cola derecha	No Poisson	16
LRT	Titterington, Smith y Markov (1985)	Cola derecha	Mezcla de Poisson	17
Hellinger deviance test	Karlis y Xekalaki (1998)	Cola derecha	Mezcla de Poisson	18
Rayner	Rayner y Best (1988)	Cola derecha	No Poisson	9
Crámer-von Mises 1	Spinelli and Stephens (1997)	Cola derecha	No Poisson	19
Crámer-von Mises 2	Spinelli and Stephens (1997)	Cola derecha	No Poisson	20
Crámer-von Mises 3	Spinelli and Stephens (1997)	Cola derecha	No Poisson	21
Crámer-von Mises 4	Henze (1996)	Cola derecha	No Poisson	22
Rescaled VT	Henze y Klar (1996)	Cola derecha	Sobredispersión o infradispesión	23
Empirical Integrated distribution function	Klar (1999)	Cola derecha	No Poisson	24
Efron double-exp.l family	Lee (1998)	Cola derecha	No Poisson	25

Tabla 2. Pruebas estadísticas de las pruebas de evaluación de la tabla 1.

1	VT= $\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\overline{X}}=(n-1)\frac{S^2}{\overline{X}}$
2	$O_2=\sqrt{\left(\frac{n-1}{2}\right)\left(\frac{S^2}{\overline{X}}-1\right)}=\frac{\sum_i(X_i-\overline{X})^2}{\overline{X}\sqrt{\{2(n-1)\}}}-\frac{\sqrt{(n-1)}}{\sqrt{2}}$
3	Z= $\frac{\sum_i(X_i-\overline{X})^2}{\overline{X}\sqrt{(2n)}}-\sqrt{\left(\frac{n}{2}\right)}$
4	$Z_2=\frac{S^2-\overline{X}}{\sqrt{\{2n\overline{X}(n\overline{X}-1)\}}}n\sqrt{(n-1)}$
5	$Z_3=\frac{m_3-\overline{X}}{\sqrt{\left\{6n\overline{X}(n\overline{X}-1)(3+\frac{n\overline{X}-2}{n-2}\right\}}}n\sqrt{(n-1)}$
6	$Z_4=\frac{m_4-3S^2-\overline{X}}{\sqrt{\left[2n\overline{X}(n\overline{X}-1)\left\{49+\frac{108(n\overline{X}-2)}{n-2}+\frac{12(n+1)(n\overline{X}-2)(n\overline{X}-3)}{n(n-2)(n-3)}\right\}\right]}}n\sqrt{(n-1)}$
7	K= $\sqrt{n\frac{\phi_n(t)-exp\{\overline{x}(t-1)\}}{exp\{\overline{x}(t^2-1)\}-exp\{2\overline{x}(t-1)\}\{1+\overline{x}(t-1)^2\}}}$
8	N= $\frac{\sum_{i=0}^{m}\frac{O_i^2}{E_i}-n-(m-1)}{\sqrt{\left[\frac{m-1}{m}\left\{2m-\frac{(m+1)^2+2m}{n}+\sum_{i=0}^{m}\frac{1}{E_i}\right\}\right]}}$
9	$S_k=\sum_{i=2}^{K}V_i^2$
10	$T_n=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\overline{X}^2}{X_i+X_j+1}+\frac{X_iX_j}{X_iX_j-1}\right)-(n-f_0)\overline{X}$

Tabla 2 (continuación). Pruebas estadísticas de las pruebas de evaluación de la tabla 1.

11	$NP_n = \frac{1}{n^3} \frac{\sum_{i,j,k,l=1}^{n} X(X_i - X_j - 1)X_k(X_K - X_l - 1)I_{(X_i + X_j = X_K + X_l)}}{\overline{X}^{1.45}}$
12	$d_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{X_i + X_j + 1} - 2exp(-\lambda)\sum_{i=1}^{n} T(X_i, \lambda) + n\frac{1 - exp(-2\lambda)}{2\lambda}$
13	$d = max\lvert H(x) - S_N(x)\rvert$
14	$W = \left(\frac{S^2}{\overline{X}} - 1\right)^2 \frac{n}{2}$
15	$\chi^2 = \sum_{x=0}^{k} \frac{(O_x - E_x)^2}{E_x}$
16	$I^{\lambda} = \frac{1}{\lambda(\lambda+1)}\sum_{i=0}^{m} E_i \left\{\left(\frac{O_i}{E_i}\right)^{\lambda+1} - 1\right\}$
17	$L = 2(L_1 - L_0)$
18	$HDT = 4n(HD_0 - HD_1)$
19	$CVM_1 = n^{-1}\sum_{j=0}^{M} Z_j^2 p_j$
20	$CVM_2 = n^{-1}\sum_{j=0}^{M} \frac{Z_j^2 p_j}{H(j)\{1 - H(j)\}}$
21	$CVM_3 = n^{-1}\sum_{j=0}^{M} Z_j^2$

Tabla 2 (continuación). Pruebas estadísticas de las pruebas de evaluación de la tabla 1.

22	$CVM_4=n^{-2}\sum_{j=0}^{M} Z_j^2 O_j$
23	$S^* = \frac{\overline{X}(VT-n)^2}{\sum_{i=1}^{n}\{(X_i-\overline{X})^2 - X_i\}}$
24	$EIDF = \sup_{t\geq 0}\left\{\lvert\Psi(t)-\hat{\Psi}_n(t)\rvert\sqrt{n}\right\}$
25	$EF = 2\sum_{i=1}^{n}\left\{\overline{X}-X_i+X_i\, ln\left(\frac{X_i}{\overline{X}}\right)\right\} = 2\sum_{i=1}^{n} X_i\, ln\left(\frac{X_i}{\overline{X}}\right)$

2.1. Componentes del modelo

Los tres componentes del MRP son:

- Componente sistemático: El predictor lineal $\eta_i = \beta_0 + \beta_1 X_1$ expresa la combinación lineal de las variables explicativas y proporciona el valor predicho.

- Componente aleatorio: el componente aleatorio ε, recoge la variabilidad de Y no explicada por el predictor lineal η

- Función de enlace: En el modelo de regresión de Poisson, la función que enlaza el componente sistemático η con el valor esperado μ es la función logarítmica, ya que $\eta = log(\mu)$.

2.2. Variable de exposición

En ocasiones es necesario incluir un término adicional al modelo, es la llamada "variable de exposición" o, también "multiplicador de tasa", que se simboliza por t. En aquellos casos en los que los recuentos de observaciones se basan en periodos de tiempo, tamaños poblacionales o tamaños espaciales no homogéneos, es aconsejable incluir en el modelo este término adicional (Kleinbaum, Kupper y Muller, 1988; Lunneborg, 1994; Winkelmann, 2000).

$$E(t_i) = m_i = t_i \exp(b_0 + b_1 X_i) \quad (4)$$

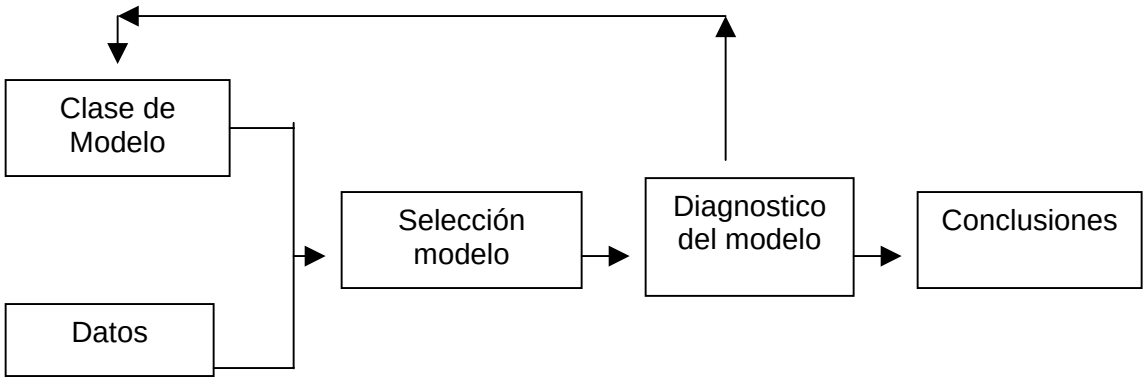
$$\eta_i = log(t_i) + b_0 + b_1 X_i \quad (5)$$

Como indica Liao (1994) la variable t debe ser como mínimo de 10 a 100 veces mayor que el dato de recuento para que los supuestos de la distribución de Poisson se cumplan, si no es así, debemos utilizar modelos log-lineales.

3. Diagnóstico del modelo

Cualquier investigador debe seleccionar el modelo a aplicar cuidadosamente, prestando mucha atención al tipo y a la estructura de los datos. Así, en datos de recuento, los valores ajustados por el modelo deben limitarse a valores no negativos, por la naturaleza misma de los datos. Los residuales y los gráficos de residuales representan un papel principal en la comprobación de la adecuación del modelo. En regresión lineal los residuales están distribuidos normalmente y pueden ser estandarizados teniendo variancias iguales. En situaciones de regresión no normales, los residuales están lejos de la normalidad y de tener variancias iguales. Un problema particular aparece cuando la variable de respuesta es discreta y toma un número pequeño de valores distintos, como en Poisson, cuya media esta cercana a cero. En estas situaciones los residuales se acercan a líneas paralelas que corresponden a valores de respuesta distintos. Estas curvas impiden obtener cualquier significado del gráfico del residual.

El diagnóstico del modelo permite determinar si el ajuste del modelo de regresión representa los datos adecuadamente. En ocasiones una elección cuidada del modelo no exime de error, por ello la comprobación de la adecuación del modelo, introduce un bucle en todo el proceso de modelado. Como señalan McCullagh y Nelder (1989, pp. 392) el proceso es como sigue:



Hay dos métodos para comprobar la adecuación del modelo: métodos formales e informales:

Informales: si podemos detectar patrones en los residuales, eso supone que podemos encontrar un modelo más adecuado a los datos.

Formales: comprueba el ajuste del modelo al introducir o eliminar un parámetro en el modelo más amplio.

A continuación presentamos los diferentes índices y test diagnósticos de los que disponemos para evaluar la adecuación del modelo.

3.1. Índices numéricos y test diagnósticos

3.1.1. Valores de influencia

El valor de influencia (hat value) para MLG puede tomarse directamente de la última iteración de los procedimientos de IWLS para ajustar el modelo, y tiene la interpretación usual. Sólo que, a diferencia del modelo lineal, el valor de influencia en MLG depende de y, así como en la configuración de las xs.

Los elementos de la diagonal principal de la matriz H constituyen los h_{ii} o valores de influencia. La observación que tenga un valor influyente alto se dice que tiene influencia. El valor de influencia será menor a medida que haya más observaciones.

La traza de la matriz H es igual a p, el número de parámetros en el modelo, el valor de p/n es el "leverage" promedio, que se emplea como valor de referencia. A partir de él se considera que una observación será influyente si su valor es, como señalan Cameron y Trivedi (1998):

$$h_{ii} > \frac{3p}{n} \qquad (6)$$

En el anexo se presenta un ejemplo de este índice y se muestra la utilidad en la etapa de evaluación.

3.1.2. Distancia de Mahalanobis

En un modelo de regresión con k variables explicativas se define la distancia de Mahalanobis (MD) de una observación i como la distancia de ese punto al centroide.

$$MD = (v_i - \bar{v})\, C^{-1}\, (v_i - \bar{v})^1 \qquad (7)$$

Donde v es el vector fila de valores de las k variables explicativas, $\bar{v}$ es el centroide y C es la matriz de covariancias. Este valor se compara con valores de la distribución χ^2 con k grados de libertad, siendo k el número de variables explicativas del modelo de regresión sin la constante. La relación entre MD y el valor h_{ii} viene dada por:

$$MD_i = (n-1)\left[h_{ii} - \frac{1}{n}\right] \qquad (8)$$

De forma que la DM será grande cuando la influencia también lo sea.

3.1.3. Distancia de Cook

La distancia de Cook (Cook, 1977) evalúa el cambio que se produce en la estimación del parámetro, cuando se elimina cada observación, es decir, evalúa la influencia de una observación sobre la estimación de los coeficientes de regresión. La estrategia que sigue es obtener la estimación de los parámetros del modelo con y sin esa observación.

Aquellas que presenten un gran imparto sobre el modelo ajustado se denominan observaciones influyentes. La adaptación de este estadístico al MLG viene dada por:

$$D_i = \frac{(\theta - \theta_i)' X'WX(\theta - \theta_{(i)})}{p\hat{\phi}} \qquad (9)$$

donde θ es el vector de estimaciones de los p parámetros, $\theta_{(i)}$ es el vector cuando la observación i ha sido eliminada, y $\hat{\phi}$ es el parámetro de escala estimado. Así pues, un valor alto de D_i indica que la observación i tiene influencia sobre la estimación de los parámetros. Cook (1977) sugiere comparar D_i con la distribución F con p y (n-p) grados de libertad. Fox (1991)sugiere utilizar como valor de referencia:

$$D_i > \frac{4}{n - p - 1} \qquad (10)$$

donde n es el tamaño de la muestra y p el número de variables del modelo. En el anexo se presenta la utilización de este índice, comprobándose la capacidad de detectar la influencia de las observaciones.

3.1.4. Índice DFFITS

Otro índice introducido por Bersley, Kuh y Welsch (1980) denominado $DFFITS_I$ mide la influencia sobre la predicción de la eliminación de la observación i.

Viene dado por:

$$DFFITS_i = \frac{r_i\sqrt{h_{ii}}}{s_i 1 - h_{ii}} \qquad (11)$$

donde r_i es el residual ordinario de la observación i, s_i es la desviación estándar de la variable de respuesta del modelo sin considerar la observación i. Valores absolutos altos de F_i indican observaciones influyentes. Un valor de corte general considerado es el 2, un punto recomendado de corte ajustado al tamaño es de

$$2\sqrt{p/n}\ . \qquad (12)$$

El punto de corte en este índice determina las observaciones que deben estudiarse más detenidamente.

3.1.5. Covratio

Mide el efecto de las observaciones en la matriz de covariancias de la estimación de los parámetros.

$$C_i = \frac{\left|\hat{\phi}_{(i)}(X'_{(i)}W_{(i)}X_{(i)})^{-1}\right|}{\left|\hat{\phi}(X'WX)^{-1}\right|} \qquad (13)$$

donde $W_{(1)}$ es la matriz W sin la observación i, $W=W_0$ cuando la matriz Hessiana completa sea usada y $W=W_e$ cuando se usa el método Fisher Scoring.

Valores de C_i cercanos a 1 indican que la observación tiene un efecto pequeño en la precisión de la estimación. Observaciones con $|C_i - 1| \geq 3p/n$ sugiere la necesidad de más investigación.

3.1.6. Indice Dfbetas

Es una medida normalizada del efecto de las observaciones en la estimación de los coeficientes de regresión. Pueden obtenerse directamente desde la iteración final del procedimiento IWLS.

$$B_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{\hat{\phi}_{(i)}(X'WX)^{-1}_{jj}}} \qquad (14)$$

donde $W=W_0$ cuando la matriz Hessiana completa sea usada y $W=W_e$ cuando se usa el método Fisher Scoring.

3.2. Análisis de residuales

Los residuales son una medida de acuerdo entre el valor observado y el valor ajustado por el modelo y permiten identificar las observaciones que no han sido ajustadas por el modelo. Como señala Cameron y Trivedi (1998) los residuales pueden usarse para detectar valores alejados, observaciones influyentes, observaciones con un gran impacto en el modelo ajustado u observaciones con un pobre ajuste.

En los modelos lineales los residuales son claramente definidos como las diferencias entre los valores actuales y los ajustados. Para modelos no lineales, como el presentado, no hay una sola definición de residuales.

Pierce y Schafer (1986) y Cox y Snell (1968) dan una excelente visión de varias definiciones de diferentes residuales propuestos para MLG. Para unificar la literatura tendremos en cuenta que:

- El nombre de un residual dependerá del estadístico del cual deriva o del autor que lo propuso.

- El adjetivo *modificado* significa que el residual ha sido modificado por una estimación de la variancia de y. El residual base se ha multiplicado por un factor $(k/w_i)^{-1/2}$, donde k es el parámetro de escala.

- El adjetivo *estandarizado* significa que la variancia del residual ha sido estandarizada para tener en cuenta la correlación entre y y $\hat{\mu}$. El residual base ha sido multiplicado por el factor $(1-h)^{-1/2}$. Tiene media 0 y desviación estándar 1.

- El adjetivo *estudentizado* significa que el residual ha sido escalado por una estimación de un parámetro de escala desconocido. El residual base ha sido multiplicado por el factor $\hat{\phi}^{-1/2}$. Tiene media 0 y desviación estándar 1. En algunas

ocasiones se estudentiza un residual estandarizado, en estos casos también se habla de residuales estudentizados en lugar de residual estandarizado estudentizado. El cálculo exacto requiere literalmente el reajuste del modelo, eliminando cada observación y calculando la discrepancia.

- El adjetivo *ajustado* significa que el residual ha sido ajustado (por la función de variancia de la familia) desde la definición original. Este adjetivo parece utilizarse únicamente con el residual de discrepancia (Pierce y Schafer,1986).

3.2.1. Residuales de respuesta ("response residual")

En el Modelo Lineal General estos residuales no son únicamente fáciles de calcular sino que además juegan un papel central en determinar el ajuste de un modelo. Se obtienen de la diferencia entre la respuesta observada y su valor estimado esperado.

$$r_{RES} = y_i - \widehat{\mu}_i \qquad (15)$$

En datos de recuento, este residual es heterocedástico y asimétrico.

En ocasiones encontramos altos outliers con o sin influencia (su valor individual no causa cambios importantes en la estimación de la curva). En la mayoría de estas situaciones las conclusiones no se ven afectadas por estos valores. Además, la normalidad asintótica de los residuales se puede lograr, en una situación más general, utilizando la variante Lindeberg-Feller del teorema central del límite. Este teorema relaja la asunción de independencia a favor de la condición de que ningún término domine la suma. Sin embargo, es más típico en los modelos lineales generalizados producir residuales que se desvían sustancialmente, en lugar de ligeramente, de las condiciones básicas. En estos casos, estos residuales no dan información. Una opción alternativa es el residual estandarizado de Pearson.

3.2.2. Residual de Pearson

En datos de recuento, como señalan Cameron y Trivedi (1998), no hay ningún residual que tenga media cero, variancia constante y distribución simétrica. Para compararlos se deben convertir a la misma escala, lo que se consigue dividiendo el residual por la desviación estándar.

$$r_i^P = \frac{r_i}{\sqrt{\text{var}(\hat{\mu}_i)}} \tag{17}$$

Con muestras grandes el residual será cero y homocedastico con variancia igual a 1, pero asimétricamente distribuido. Este residual es una versión reescalada de los residuales de trabajo. Se puede utilizar este residual para comprobar el ajuste de cada observación en los MLG. Para detectar outliers podemos hacer un gráfico del residual de Pearson versus el número de observación.

En el anexo comprobamos como este residual detecta claramente un outlier importante en el modelo, ya el estudio de la influencia y la distancia de Cook, lo presenta como un valor influyente.

El nombre viene del hecho de que para la distribución de Poisson, el residual de Pearson es justo la raíz cuadrada de la χ^2 de Pearson

$$\Sigma r_I^P = \chi^2 \tag{18}$$

Pudiendo usarse este estadístico como una medida de variación residual. Valores altos (en valor absoluto) de este residual, indican un fallo del modelo en el ajuste de una observación particular.

3.2.2.1. Residual de Pearson estudentizado

Se utilizan para detectar observaciones anómalas (outliers). Cualquier residual estudentizado cuyo valor absoluto sea superior a dos, debería ser evaluado, aunque ello no indica que sea un “outlier”.

Este residual presenta un problema de asimetría, por lo que es conveniente transformarlo.

$$r_i^{PT} = \frac{r_{P_i}}{\sqrt{\hat{\phi}_{(i)}(1-h_i)}} \qquad (19)$$

donde $\hat{\phi}_{(i)}$ es una aproximación por pasos de ϕ después de excluir la observación i.

3.2.2.2. Residual de Pearson estandarizado

$$r^{PS}{}_i = \frac{r_{P_i}}{\sqrt{\hat{\phi}(1-h_i)}} \qquad (20)$$

donde $\hat{\phi}$ es la estimación del parámetro de dispersión ϕ.

3.2.3. Residual de trabajo (“working residual”)

Como señala Gill (2000) en el proceso de ajuste de los modelos lineales generalizados, los programas utilizan el algoritmo IWLS. Un grupo de “pesos de trabajo” se calculan en cada paso de la estimación hasta que la derivada esta suficientemente cercana a cero. Ocasionalmente se utiliza la cantidad obtenida del último paso del proceso iterativo: la diferencia entre la respuesta de trabajo y el predictor lineal. Se define como:

$$r_i^W = (y_i - \hat{\mu}_i)\left(\frac{\partial \eta}{\partial \mu}\right)_i \qquad (16)$$

Este residual se utiliza como diagnóstico para la evaluación de la convergencia así como indicador del ajuste del modelo en ese punto. El gráfico "componente+residual" utiliza estos residuales para definir los residuales parciales.

3.2.4. Residual de Anscombe

Para distribuciones de respuesta no normal en MLG, la distribución de los residuales de Pearson es a menudo asimétrica. Anscombe propuso un residual que utilizaba una función A(y) en lugar de y en la derivación de residuales (Anscombe 1953, McCullagh y Nelder, 1989). La función A(y) es elegida para hacer la distribución de A(y) lo más normal posible y viene dada por:

$$\text{A(y)} = \int \frac{d\mu}{V^{1/3}(\mu)} \qquad (21)$$

Donde el residual es:

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}} \qquad (22)$$

Este tipo de residual es especialmente usado en los casos donde los residuales de discrepancia ajustados son inapropiados. Para la distribución de la respuesta Poisson, los residuales de Anscombe son:

$$r_i^A = \frac{3}{2}\left(y_i^{2/3}\hat{\mu}_i^{-1/6} - \hat{\mu}_i^{1/2}\right) \qquad (23)$$

Los residuales de Anscombe toman una gran variedad de formas en la literatura (McCullagh y Nelder, 1989, p.38; Fahrmeir y Tutz, 2001; Pierce y Schafer, 1986, p.978; Cox y Snell, 1968, pp.258-261). Como señala Gill (2000) una posible explicación de este hecho, es el deseo de los autores de acercar la estimación a la media añadiendo o eliminando una constante.

3.2.4.1. Residual de Anscombe estandarizado

Algunas versiones incluyen $\sqrt{1-h_i}$ en el denominador de la definición del residual Anscombe. Esta transformación daría el residual de Anscombe estantarizado con media 0 y variancia 1.

$$r_i^{AS} = \frac{r_i^A}{\sqrt{\hat{\phi}(1-h_i)}} \qquad (24)$$

donde $\hat{\phi}$ es la estimación del parámetro de dispersión ϕ.

3.2.3.2. Residual de Anscombe estudentizado

$$r_i^{AT} = \frac{r_i^A}{\sqrt{\hat{\phi}_{(1)}(1-h_i)}} \qquad (25)$$

donde $\hat{\phi}_{(i)}$ es una aproximación por pasos de ϕ después de excluir la observación i.

3.2.5. Residual de discrepancia ("deviance residual")

El residual más utilizado en MLG es el residual de discrepancia. Basado en la contribución a la discrepancia global aportada por cada observación, permite ver la contribución de cada observación a la discrepancia, de forma similar a los residuales del modelo lineal. Así la discrepancia juega un papel clave en las derivaciones del MLG y en las inferencias de los resultados.

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{\hat{d}_i^{\,2}} \qquad (26)$$

donde d_i es la contribución individual a la discrepancia. En Poisson el cálculo de d_i^2 es:

$d_i^2 = 2\hat{\mu}_i$ si $y_i = 0$

$d_i^2 = 2\left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$ en otro caso

Este tipo de residuales pueden usarse para comprobar el ajuste de cada observación en MLG.

Estos residuales son a menudo estandarizados, estudentizados o ambos. Este residual como señala Hardin y Hilbe (2001) es preferido sobre los residuales de Pearson, para la evaluación del modelo, porque sus características distribucionales están cercanas a las que aparecen en los modelos de regresión lineal. Tienden a una distribución normal con media cero y desviación estándar uno, independientemente del tipo de MLG utilizado.

3.2.5.1. Residual de discrepancia estandarizado

$$r_i^{DS} = \frac{r_i^D}{\sqrt{\hat{\phi}(1-h_i)}} \qquad (27)$$

donde $\hat{\phi}$ es la estimación del parámetro de dispersión ϕ.

3.2.5.2. Residual de discrepancia estudentizado

Se aproximan mejor a la distribución normal que los residuales de discrepancia.

$$r_i^{DS} = \frac{r_i^D}{\sqrt{\hat{\phi}_{(i)}(1-h_i)}} \quad (28)$$

donde $\hat{\phi}_{(i)}$ es una aproximación por pasos de ϕ después de excluir la observación i.

3.2.5.3. Residual de discrepancia ajustado

Pierce y Schafer (1986), estudiaron este residual en detalle y recomendaron una corrección para mejorar la normalidad. El residual de discrepancia se ajusta haciendo la convergencia a la distribución normal.

$$r_i^{D_a} = r_i^D + \frac{1}{6}\rho_3(\theta) \quad (29)$$

donde $\rho_3(\theta)$ es definido para cada familia de distribuciones. Para Poisson viene dado por:

$$\frac{1}{\sqrt{\hat{\mu}_i}} \quad (30)$$

3.2.6. Residual de Puntuaciones "Score Residual"

Estos son los residuales usados al calcular la estimación sandwich de la variancia. Las puntuaciones están relacionadas con la "Score Function" o ecuación de estimación que se optimice:

$$r_i^S = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)}\left(\frac{\vartheta\eta}{\vartheta\mu}\right)_i^{-1} \quad (31)$$

3.2.7. Residuales de verosimilitud

Analizan la contribución de cada observación en el ajuste del modelo, obteniendo la discrepancia para el conjunto de observaciones y comparándola con la discrepancia del modelo cuando se elimina la observación i.

$$r_i^L = sign(y_i - \hat{\mu}_i)\sqrt{h_i(r_i^{PS})^2 + (1-h_{ii})(r_i^{DS})^2} \qquad (32)$$

Permite obtener una distancia de Cook modificada. Estos residuales son una combinación de los residuales de Pearson estandarizados y los residuales de discrepancia estandarizados. Cox y Snell (1968) los llaman raw residual o crude residual.

3.2.8. Residuales "Jackknife" o parciales

También llamados "residuales de validación cruzada" y "residuales estudentizados eliminados". Se utilizan para comprobar la forma del predictor y se calculan, por tanto, para cada predictor. Hines y Carter (1993) presentan el uso gráfico de estos residuales para comprobar el ajuste del modelo. Son apropiados para señalar las observaciones atípicas, inconsistentes.

$$r_{ki}^T = (y_i - \hat{\mu}_i)\left(\frac{\partial \eta}{\partial \mu}\right)_i + \left(x_{ij}\hat{\beta}_K\right) \qquad (33)$$

donde k = 1, ..., p; p es el número de predictores y $\left(x_{ij}\hat{\beta}_K\right)$ se refiere a la observación i del predictor k en los k coeficientes ajustados.

Se distribuyen según una distribución t con (n-p-1) grados de libertad. Su cuadrado corresponde al valor de la prueba F de comparación entre dos modelos: modelo de trabajo y modelo outlier (en el modelo de trabajo se ha añadido un parámetro específico para la observación i, incluye una variable x que toma el

valor 1 para esta observación y 0 para el resto) Ato, Losilla, Navarro, Palmer y Rodrigo (2000 a)

3.2.9. "Quantile Residuals"

Son los residuales de elección para los modelos lineales generalizados en situaciones con mucha dispersión cuando el residual de Pearson y el de la discrepancia pueden ser aproximadamente no normales. Pueden ser los únicos residuales apropiados para binomial y Poisson cuando las respuestas pueden tomar solo un pequeño número de valores distintos (Dunn y Smyth ,1996).

$$r_i^q = \phi^{-1}\{1 - \exp(y_i / \widehat{\mu}_i\} \qquad (34)$$

Estos residuales son una transformación de los residuales de Cox y Snell (1968).

Comprobar las características distribucionales de los residuales puede ayudarnos al encontrarnos en una situación de evaluación. Los residuales pueden ofrecernos una valiosa información por ellos mismos, porque, aunque los residuales en MLG no se les requiere que sean normales, patrones sistemáticos de la distribución puede indicarnos un mal ajuste o una mala medida. Sin embargo la mejor forma de evaluar la distribución de los residuales es mediante gráficos. A continuación presentamos los diferentes gráficos que podríamos utilizar para hacerlo y en el anexo 1 se presenta varios ejemplos aplicados en datos del campo de las conductas adictivas.

3.3. Gráficos

Como señalan Cameron y Trivedi (1998) quizás la opción más fructífera de los residuales es su visualización a través de gráficos. Estos gráficos pueden incluir residuales frente a valores de identificación, frente a valores predichos de la variable dependiente, frente a factores incluidos en el modelo o factores no incluidos para valorar su inclusión.

Como señalan Cameron y Trivedi (1998) en datos de recuento no tiene sentido hacer un gráfico de residuales frente al valor de la variable dependiente. Pero existe una gran variedad de gráficos que si se pueden hacer y que son interpretables en este tipo de datos. Gráfico de residuales frente la media predicha, gráfico del valor actual de y sobre el valor predicho (este gráfico es difícil de interpretar si la variable y toma pocos valores).

3.3.1. Grafico de probabilidad normal (Q-Q plot)

Permite estudiar si la distribución de los residuales es normal. Para ello se ordenan en el eje de las abcisas los residuales estandarizados y en el eje de las ordenadas se sitúan los correspondientes valores esperados bajo la distribución Normal, obtenidos mediante la función inversa de la distribución Normal acumulada, y dados por:

$$\phi^{-1}\left[\frac{i-3/8}{n+1/4}\right] \qquad (35)$$

Si el modelo ajusta bien, se obtendrá una recta de 45º por el origen. Si los residuales son muy asimétricos la recta no pasará por el origen, mientras que si la distribución es muy alargada se dibujará una línea curva. En el anexo se muestra la utilización de este gráfico en el estudio de la distribución de los residuales.

Este gráfico también se puede encontrar con el nombre de “normal score plot”.

$$\text{Rnorm}_i = \bar{r} + s_r\phi^{-1}((i-.5)/n) \qquad (36)$$

i=1,.....,n, donde s_r es la desviación estándar de la muestra de r y ϕ^{-1} es la inversa de la función de distribución acumulada.

3.3.2. Gráfico índice (Index plot)

Este gráfico sitúa un determinado índice en las ordenadas frente al número de observación, permitiendo así la detección de múltiples anomalías en el modelo (observaciones alejadas, influyentes, predictor lineal mal especificado...). Es uno de los gráficos más utilizados en la etapa de evaluación, tanto por la sencillez como por su fácil interpretación.

3.3.3. Residuales frente a valores del predictor lineal (Residual plot)

Este gráfico sitúa valores residuales en las ordenadas frente a valores resultantes del predictor lineal del modelo, permitiendo así la detección de errores de especificación en el componente sistemático debidos a la omisión de alguna variable relevante o la necesidad de transformar alguna de las variables incluidas.

3.3.4. Gráfico de variable añadida (Added variable plot)

Este gráfico permite detectar si se debe incluir o no una variable en un modelo, en el que están presentes otras variables. La variable evaluada puede ser una variables nueva o la potencia de una variable ya incluida.

Este gráfico se construye situando los residuales de Pearson frente a los residuales de la variable añadida, obtenidos mediante $(I - H)\ W^{-1/2}\ u$, donde u simboliza la nueva variable.

Un gráfico de variable añadida con tendencia lineal indica que la nueva variable debe ser añadida al modelo.

Para ayudar a la dificultad de examinar residuales cuando las variables de respuesta son discretas, podemos ajustar en este gráfico un “nonresistant scatterplot smoother”.

3.3.5. Gráfico de residuales parciales (Partial residual plot)

Un residual parcial de una variable explicativa es aquel que se obtiene después de haber eliminado de la variable de respuesta la influencia modelada de todas las demás variables incluidas en el modelo. Un gráfico de residuales parciales debería

ser una recta si la variable no necesita transformación. Un gráfico no lineal indica que la variable debe ser transformada y la forma del gráfico proporciona una guía del tipo de transformación que se debe utilizar.

Un gráfico de residual parcial, es un gráfico de $r_i + b_k x_{ik}$ versus $x_i k$ donde r_i es el residual de la observación i, x_{ik} es la observación i del predictor k y b_k es el coeficiente de regresión estimado para el predictor k.

3.3.6. Gráfico de variable construida (Constructed variable plot)

Este gráfico permite averiguar si una variable X debe ser sustituida por alguna potencia de X. Para ello, se basa en la transformación de Box y Cox (1964) buscando el valor de λ más adecuado por medio del siguiente esquema:

$$x^{\lambda} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} \lambda \neq 0 \\ \log x \lambda = 0 \end{cases} \qquad (37)$$

Si la variable no requiere transformación entonces λ=1. Para averiguar si X_i necesita transformación se define la variable construida Z_j que viene dada por $Z_j = \hat{\beta}_j X_j \log X_j$ siendo $\hat{\beta}_j$ el coeficiente de variación de la variable en el modelo ajustado.

3.3.7. Grafico de residuales vs. escala de información

Si la variabilidad de los residuales no es constante, es posible que haya una mala especificación de la distribución del componente aleatorio. El gráfico de residuales vs. escala de información, representa los residuales de Pearson, frente a la escala de información constante de la distribución del error en Poisson dada por $2\sqrt{\mu_i}$.

3.4. El problema de la sobredispersión

Con el término “sobredispersión” entendemos que la variancia de la variable respuesta Y excede la variancia nominal. Es muy común en la práctica. La

incidencia y el grado de sobredispersión encontrado dependen mucho del campo de aplicación. Son muchos los autores que desde diferentes disciplinas han tratado este tema: Hauer (2001); Osgood (2000).

El fallo de la equidispersión tiene similares consecuencias cualitativas que el fallo de la asunción de la homocedasticidad en el modelo de regresión lineal.

3.4.1. Causas de la sobredispersión

- **Fuentes de especificación errónea en el MRP**

El modelo de regresión de Poisson se basa en tres asunciones:

1. $\mu_i | X_i \approx Poisson(\mu)$

2. μ_i = exp($X_i \beta$)

3. $\mu_i | X_i, i = 1,2,.....,n$ están distribuidas independientemente.

o ***Función media incorrecta***

Tal como hemos visto anteriormente la función media del MRP es:

$$E(y_i \, I \, x_i) = \mu_i = \exp(x_i \beta) \qquad (38)$$

Si denotamos la función media verdadera como μ_0 y el valor esperado respecto a la densidad verdadera como E_0 :

$$E_0(y_i \, I \, x_i) = \mu_0 = f(x_{0,} \beta_0) \qquad (39)$$

La función media está especificada erróneamente si no existe ninguna β que cumpla μ_i = f($x_{i,} \beta_0$) de forma que $\mu_i \neq \mu_0$.

Los errores de especificación, según Winkelmann (2000) pueden ser debidos a:

- Omisión de variables explicativas, siendo éstas no independientes de X.

- El predictor no es lineal en β.

- Las variables explicativas entran en el predictor a través de alguna transformación f(x) en lugar de linealmente.

- Error de especificación en la función de enlace.

o ***Heterogeneidad no observada***

Esta situación surge cuando las variables explicativas no explican la heterogeneidad individual, es decir, las observaciones difieren aleatoriamente de una forma que no es recogida exhaustivamente por las variables explicativas del modelo. Esta situación puede ser interpretada como un defecto de la función media, resultado de la omisión de algunas variables explicativas (Cameron y Trivedi, 1998; Lindsey, 1998)

El MRP no tiene término de error aditivo, de forma que la heterogeneidad de Y es modelada a través de una función determinística de las variables explicativas. Así un modelo con heterogeneidad no observada no puede seguir una distribución de Poisson.

o ***Proceso dependiente***

El MRP presenta errores de especificación cuando la probabilidad de ocurrencia de un evento entre t y $t + \Delta$ depende de eventos pretéritos. En esta situación, el proceso de Poisson es un proceso dependiente y generador de sobredispersión.

Las diferentes aproximaciones para modelar los procesos dependientes han consistido en:

- La derivación de un Modelo de Regresión Binomial Negativa (MRBN).

- Un modelo de dependencia de duración. Utiliza la distribución gamma y admite la infradispersión.

o ***Selectividad***

Cuando los datos son generados de tal forma que el investigador no observa el rango entero de valores de la variable de recuento y*, sino una selección de los mismos, estamos ante una situación de selectividad. Esta fuente de error está presente si los datos están truncados, censurados o si la información esta parcializada.

La diferencia entre una situación de censura y una de truncamiento radica en que en la primera, la variable dependiente es observada para un rango determinado de valores, mientras que en el caso del truncamiento, ciertas observaciones son omitidas completamente de la muestra.

Si ignoramos el efecto de la selectividad, las estimaciones del MRP resultan, en general, inconsistentes. Es más, en caso de truncamiento, las estimaciones resultan, además ineficientes.

Se pueden distinguir dos tipos de selectividad:

- Las observaciones son truncadas o censuradas dependiendo del resultado en y*.

- Las observaciones pueden ser censuradas o truncadas dependiendo del resultado de otra variable c, que puede ser dependiente o independiente de y*. A esta situación se le denomina truncamiento incidental o selectividad endógena.

o ***Información parcializada***

Sea y* un número total (real) de eventos e y el número de recuentos reportados; en una situación de información parcializada se tiene que $y \leq y^*$.

Winkelmann (2000) señala que la estructura formal que presenta es:

$$y = \sum_{i=1}^{y^*} B_i \qquad (40)$$

Donde B_i es una variable indicadora que toma el valor 1 si el evento es reportado o 0 en caso contrario.

Existen tres tipos diferentes de información parcializada, cada uno de los cuales genera un modelo diferente:

- Información parcializada aleatoria: Las B_i son i.i.d. y siguen una distribución de Bernoulli.

- Información parcializada logística: la probabilidad de informar Pr(B=1) es una función logística.

- Modelo de recuento total: Los eventos son registrados sólo si se sobrepasa un umbral determinado.

o ***Exceso de ceros***

El exceso de ceros es una de las fuentes de especificación errónea más frecuente, que consiste en la presencia de un exceso de valores de recuento 0 con relación a la probabilidad predicha por la distribución de Poisson.

Una de las primeras aplicaciones de un MRP con una distribución de recuentos con exceso de 0 fue llevada a cabo por Lambert (1992), quien introdujo un MRP de ceros modificados en el cual con una probabilidad ω, el único resultado posible es 0, y con una probabilidad 1-ω, el resultado es una variable aleatoria de Poisson (λ). Tanto ω como λ pueden depender de las variables explicativas.

o ***Función variancia incorrecta***

La ausencia de variancia nominal de Poisson implica un incumplimiento del supuesto distribucional.

Como hemos visto anteriormente la función variancia de referencia es:

$$\text{Var}\,(y_i \mid x_i) = E(y_i \mid x_i) = \exp(x_i \beta) \qquad (41)$$

Situación que se denomina “equidispersión”, mientras que cuando no se da la relación de identidad anterior, se produce sobredispersión caracterizada por:

$$\text{Var}\,(y_i \mid x_i) > E(y_i \mid x_i) \qquad (42)$$

O bien infradispersión:

$$\text{Var}\,(y_i \mid x_i) < E(y_i \mid x_i) \qquad (43)$$

Cuando la función que relaciona la media condicional con la variancia condicional no es la función identidad aparece la sobredispersión o la infradispersión. En general, puede tratarse de una función arbitraria que recoge variables explicativas adicionales u_i, de forma que la función variancia puede definirse como (Winkelmann, 2000; Winkelmann y Zimmermann, 1995) :

$$Var(y_i \mid x_i, u_i) = f[\exp(x_i \beta), u_i] = \exp(x_i \beta)\, u_i \qquad (44)$$

Esta modificación en la función de la variancia ha sido incorporada en diversos modelos de los cuales el aplicado con mayor frecuencia es el MRBN, que constituye un caso particular del hipermodelo denominado Negbin k, (Cameron y Trivedi, 1998; Winkelmann y Zimmermann, 1995). La función variancia de Negbin k es, siguiendo la notación de Cameron y Trivedi (op. cit.):

$$Var(y_i \mid x_i, u_i) = f[\exp(x_i \beta), u_i] \qquad (45)$$

3.4.2. Detección de la sobredispersión

Las pruebas que vamos a ver, se clasifican en tres grandes bloques:

- Pruebas para modelos anidados: La detección de sobredispersión se basa en la comparación de la variancia poissoniana con una función variancia generalizada, en la que queda anidada la primera. La función variancia más habitual es la que viene dada por la distribución binomial negativa.

- Pruebas para modelos no anidados: La teoría estadística proporciona un amplio abanico de herramientas para el contraste de hipótesis que se presentan en forma de restricciones paramétricas. Evaluar una restricción implica comparar un modelo restringido con uno más general o ampliado donde el primero está anidado en el segundo. Una implicación directa es que el modelo restringido nunca puede ser mejor que el modelo ampliado, medido en términos de verosimilitud (Winkelmann, 2000). Sin embargo, en algunas situaciones se requiere la evaluación de dos modelos no anidados, en cuyo caso es necesaria la aplicación de pruebas de detección de sobredispersión para este tipo de modelos. Uno de los errores de especificación más frecuente proviene de la presencia excesiva de ceros. En esta situación, puede ser adecuado aplicar modelos específicos como los modelos de recuento de umbral o los modelos de recuentos con ceros aumentados (Shankar, Milton y Mannering, 1997), que son extensiones de los modelos de referencia como pueden ser el MRP o MRBN. Dichos modelos son, con respecto a sus modelos de referencia, ejemplos de modelos no anidados.

- Pruebas basadas en la regresión: Los residuales de Poisson pueden indicar un incumplimiento del supuesto de equidispersión (Winkelmann, 2000). El análisis de residuales puede llevarse a cabo gráficamente o mediante regresiones auxiliares.

3.4.2.1. Pruebas para modelos anidados

Cuando existe un modelo alternativo al MRP que contempla una función variancia más general que la de Poisson y, al mismo tiempo, la función variancia de Poisson queda anidada en esa función variancia más general a través de alguna restricción paramétrica, son aplicables las pruebas clásicas de sobredispersión. En este caso en que un modelo queda anidado dentro de un modelo mas general se habla,

genéricamente, de modelos anidados. En este sentido, el MRP y el MRBN son modelos anidados que frecuentemente se comparan en presencia de sobredispersión. Concretamente, el MRP queda anidado dentro del MRBN si se cumple la restricción: $H_0: \alpha = 0$ puesto que se obtiene:

$$\text{Var}(y_i | x_i) = \mu_i \qquad (46)$$

Vamos a ver tres pruebas clásicas en la evaluación de sobredispersión:

- *Prueba de razón de verosimilitud (LR)*
- *Prueba de Wald*
- *Prueba multiplicador de Lagrange (LM)*

Cuando H_0 es verdadera, las pruebas LR, Wald y LM son asintóticamente equivalentes (Rodríguez, 2002). A medida que la n aumenta, la distribución muestral de las tres pruebas converge en la misma distribución χ^2 con grados de libertad igual al número de restricciones evaluadas.

- **Prueba de razón de verosimilitud (LR)**

Sea ℓ_r el valor de la función log-verosímil evaluada en las estimaciones de la máxima verosimilitud restringida (p.ej. El MRP), y ℓ_{nr} el valor de la función log-verosímil evaluada en las estimaciones de la máxima verosimilitud no restringida (p.ej. el MRBN), y sea k el número de restricciones (k=1 en el caso de una prueba de MRP contra MRBN). Entonces, bajo H_0 (si la restricción es correcta) (Winkelmann, 2000):

$$LR = -2(\ell_r - \ell_{nr}) \approx \chi^2_{(K)} \qquad (47)$$

Sin embargo, la distribución de este estadístico no es estándar, debido a la restricción que α no puede ser negativa. Cameron y Trivedi (1998) exponen una

solución. La distribución asintótica de la prueba LR tiene una probabilidad de masa de 0.5 en 0 y una distribución 0.5 $\chi^2_{(1)}$ para valores estrictamente positivos.

Esto significa que si el contraste se fija para el nivel α, se rechaza H_0 si la prueba estadística supera $\chi^2_{1-2\alpha}$ en lugar de $\chi^2_{1-\alpha}$.

La prueba LR requiere el uso de la misma muestra para todos los modelos empleados. Puesto que la estimación máximo-verosímil excluye todos los casos con datos faltantes, es frecuente que el tamaño muestral cambie al incluir o excluir una variable. Para asegurar una constancia en el tamaño de la muestra, Long (1997) recomienda excluir de la matriz de datos aquellas observaciones que presenten datos faltantes en las variables que formaran parte de los modelos evaluados.

- **Prueba de Wald**

Esta prueba como indica Breslow (1990) se basa en la comparación de los coeficientes de estimación con sus errores estándar.

El punto de partida de la prueba de Wald es una distribución asintótica del estimador máximo-verosímil del modelo no restringido. En contraste con la prueba LR, es suficiente la estimación de un solo modelo.

El estadístico de Wald viene dado por:

$$W = [R\hat{\theta} - q]' \, [R\hat{V}ar(\hat{\theta})R']^{-1} \, [R\hat{\theta} - q] \qquad (48)$$

donde W sigue una distribución χ^2 con grados de libertad igual al número de restricciones, si la hipótesis nula es correcta. Si el número de restricciones es 1, el estadístico queda reducido al cuadrado del estadístico t. Si dividimos W por sus grados de libertad obtenemos el estadístico F.

La prueba de Wald presenta dos componentes (Long, 1997):

- [R$\hat{\theta} - q$] al principio y al final de la fórmula, mide la distancia entre el valor estimado y el hipotetizado.

- [R$\hat{Var}(\hat{\theta})R'$]$^{-1}$ refleja la variabilidad en el estimador o, alternativamente, la curvatura de la función de verosimilitud.

Por ejemplo, si asumimos que la estimación del MRBN produce una estimación $\hat{\alpha}$ con una variancia asintótica estimada $\hat{Var}(\hat{\alpha})$. El MRP requiere $\alpha = 0.$ Así, la prueba de Wald aplicada a la H_0: Poisson(μ) contra H_1: binomial negativa con media μ y variancia $\mu + \alpha\mu^2$, se basa en el estadístico t:

$$\frac{(\hat{\alpha} - 0)}{\sqrt{\hat{Var}(\hat{\alpha})}} \qquad (49)$$

De hecho, tal como indican Cameron y Trivedi (1998), el test Wald se implementa habitualmente como una prueba t, que aquí tiene una masa de 0.5 en 0 y una distribución normal para valores estrictamente positivos. En este caso se aplica el valor crítico habitual de contraste de hipótesis unilateral $z_{1-\alpha}$.

- **Prueba multiplicador de Lagrange (LM)**

La prueba multiplicador de Lagrange (LM), también conocida como la prueba de puntuaciones (Score test), estima sólo el modelo restringido y evalúa la pendiente de la función log-verosímil en la restricción. Tal como indica Long (1997), si la hipótesis es cierta, la pendiente (conocida como puntuación) evaluada en la restricción a través de:

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_r} \qquad (50)$$

debe estar próxima a 0.

Así, la restricción es rechazada si la puntuación esta alejada de cero. Si la hipótesis nula es verdadera, esto es $\hat{\theta}_r = \theta_0$, el vector de puntuaciones sigue asintóticamente una distribución normal con media cero y matriz de variancias-covariancias igual a la matriz de información de Fisher (Winkelmann, 2000).

La prueba estadística LM viene dada por (Winkelmann, 2000):

$$\text{LM}=\sqrt{\left[\sum_{i=1}^{n}\frac{1}{2\hat{\mu}_i^2}\right]}\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2-y_i \qquad (51)$$

Bajo H_0, la prueba LM sigue asintóticamente una distribución normal (puesto que es la raíz cuadrada de una distribución χ^2 con un grado de libertad) y la prueba de sobredispersión es una prueba unilateral con el valor crítico z_α.

3.4.2.2. Pruebas para modelos no anidados

Cuando nos encontramos frente a modelos no anidados, no son aplicables las pruebas clásicas de sobredispersión. En este caso disponemos de otras pruebas para evaluar la sobredispersión:

- **Prueba de Vuong**

La prueba de Vuong (Vuong,1989) es una extensión de la prueba Razón de Verosimilitud para evaluar modelos no anidados (Winkelmann, 2000):

$$\text{LR}_{\text{NA}} = \frac{\frac{1}{\sqrt{n}}\left[l_f(\hat{\alpha})-l_g(\hat{\beta})\right]}{\omega} \qquad (52)$$

donde:

$$\omega^2=\frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i|x_i,\hat{\alpha})-l_g(y_i|x_i,\hat{\beta})\right]^2-\left[\frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i|x_i,\hat{\alpha}\right]-l_g(y_i|x_i,\hat{\beta})\right]^2 \qquad (53)$$

El objetivo de la prueba es seleccionar el modelo más cercano a la distribución condicional verdadera. La hipótesis nula es que los dos modelos son equivalentes:

$$H_0 = E_0 [l_f (\hat{\alpha}) - l_g(\hat{\beta})] = 0 \qquad (54)$$

Bajo la hipótesis nula, el estadístico LR_{NA} converge en una distribución normal.

Si el estadístico $LR_{NA} > c$, siendo c el valor crítico para el nivel de significación, se rechaza la hipótesis de igualdad entre modelos, y se selecciona el modelo f en lugar del modelo g. Por el contrario, un valor $LR_{NA} < (-c)$ es indicador de que el modelo g es mejor modelo que el f. Por otro lado si $|LR_{NA}| < c$, no se puede discriminar entre ambos modelos, de forma que no se rechaza la hipótesis nula (Shankar, Milton y Mannering, 1997; Winkelmann, 2000).

- **Anidamiento artificial**

Un método alternativo para evaluar modelos no anidados es la construcción de hipermodelos. En general, los hipermodelos contienen un parámetro adicional y una prueba entre dos modelos que se convierte en una prueba de restricción en el hiperparámetro (Winkelmann, 2000). Un ejemplo es el modelo Negbin k, (Cameron y Trivedi, 1986; Winkelmann y Zimmermann, 1991, 1995) que es equivalente al modelo generalizado de binomial negativa (Saha y Dong, 1997) y constituye un hipermodelo para los modelos no anidados Negbin I y Negbin II. En el modelo Negbin K

$$\text{Var}(y \mid x) = \mu + \alpha\mu^{2-k} \qquad (55)$$

Concretamente, los modelos Negbin I y Negbin II quedan anidados en Negbin k a través de las restricciones paramétricas k=0 y k=1, respectivamente:

-> Si k=1, entonces Var (y l x) = $\mu + \alpha\mu$ = $\text{Var}_{\text{NegbinI}}$

-> Si k=0, entonces Var (y l x) = $\mu + \alpha\mu^2$ = $\text{Var}_{\text{NegbinII}}$

3.4.2.3. Pruebas basadas en la regresión

Siguiendo a Palmer, Losilla, Llorens, Sesé, Montaño Jiménez y Cajal (2002) se puede llevar a cabo la contrastación de la componente de variancia del modelo de regresión de Poisson frente a los modelos Negbin I y Negbin II, mediante una regresión lineal por el origen (Cameron y Trivedi, 1990)

La novedad de estas pruebas, radica en la no-asunción de la distribución.

Las hipótesis a contrastar son:

H_0 : Var (y_i) = μ_i

H_1 : Var (y_i) = $\mu_i + \alpha g(\mu_i)$

donde $g(\mu_i) = \mu_i$ para el modelo BN1 y $g(\mu_i) = \mu_i^2$ para el modelo BN2

Para ello se forman los valores esperados m_i = exp($X\beta$) y se lleva a cabo una regresión lineal OLS por el origen de:

$$\frac{(y_i - m_i)^2 - y_i}{\sqrt{2m_i}} = b\frac{m_i^2}{\sqrt{2m_i}} + u_i \qquad (56)$$

donde u_i es el término de error. En este modelo, el test de α es asintóticamente normal bajo la hipótesis nula de no sobredispersión frente a la alternativa de sobredispersión del tipo BN2.

3.4.3. Corrección de los errores estándar en presencia de sobredispersión

Diferentes soluciones se pueden aplicar para corregir los errores estándar infraestimados, debido a la presencia de sobredispersión. Por un lado existe la posibilidad, como señalan entre otros Palmer et al (2002), de corregir directamente el error estándar de los coeficientes del MRP mediante su producto por diferentes

índices, por otro lado se puede estimar los errores estándar a través de técnicas de remuestreo como el Bootstrap y el Jackknife o a través del estimador robusto de la variancia o estimador Sandwich (Cameron y Trivedi, 1998).

En los siguientes apartados se expondrán las diferentes opciones de forma extendida.

3.4.3.1. Modificación de los errores estándar

Wedderburn (1974), así como Cox y Snell (1989), sugiere utilizar un factor de inflación de la variancia:

$$C = \sqrt{\frac{\chi^2}{gl}} \qquad (57)$$

La estimación de este factor se debe obtener a partir del modelo ampliado.

En el caso de independencia y homogeneidad este valor será igual a la unidad, en caso de dependencia o heterogeneidad, c>1. Se recomiendan dos consideraciones antes de usar este valor: Conocer la naturaleza que lleva a sospechar de la sobredispersión y que el valor c obtenido sea igual o mayor a 1'3 (considerando también que la "significación" del valor p de la prueba de bondad de ajuste de Pearson sea < 0'2). Si los grados de libertad son menores de 5, entonces se puede establecer un valor de p más pequeño, (entre 0'05 y 0'01).

Si, basándonos en los dos criterios anteriores, aparece la sobredispersión debemos multiplicar la estimación de la variancia y covariancia por el factor c. Así, la variancia de la muestra será $c \cdot var(\hat{\theta})$. Como señalan Anderson, Burnham y White(1994), si el factor de inflación de la variancia estimada esta cercano a 1, el efecto sobre la estimación del error estándar es muy pequeño.

En los casos en los que la sobredispersión está muy marcada, entonces la inflación de la variancia y la covariancia es importante. Se debe esperar que $1 \leq c \leq 4$ (ver Eberhardt, 1978). Si la sobredispersión se ha identificado usando la

prueba de bondad de ajuste y sus grados de libertad del modelo global, entonces el modelo de selección debe estar basado en QAIC o $QAIC_c$ como:

$$QAIC = -\left[2\,log(L(\hat{\theta}))/\hat{c}\right] + 2K \qquad (58)$$

$$QAIC_c = -\left[2\,log(L(\hat{\theta}))/\hat{c}\right] + 2K + \frac{2K(K+1)}{n-K-1} \qquad (59)$$

Estos índices basados en la teoría de Cuasi-verosimilitud (Wedderburn, 1974) modifican los índices AIC y AIC_c, basados en la teoría de Verosimilitud de Fisher cuando estamos ante datos sobredispersos. Si no existe sobredispersión, -2log[L ($\hat{\theta}$)] es una medida de la pérdida de ajuste.

Otro factor de inflación de la variancia es:

$$\sqrt{\frac{D}{gl}} \qquad (60)$$

donde D es la discrepancia

La estimación de este factor se obtiene desde el cálculo de la discrepancia del modelo ampliado y sus grados de libertad.

$$D = -2 \times \sum_{i=1}^{n} y_i \times log\left(\frac{m_i}{y_i}\right) + (y_i - m_i) \qquad (61)$$

Una versión diferente del mismo factor de inflación de la variancia que podemos encontrar es:

$$\sqrt{\phi} \qquad (62)$$

donde ϕ es el parámetro de dispersión

El parámetro de dispersión, también llamado parámetro de escala se obtiene de:

$$\phi = \frac{D}{gl} \qquad (63)$$

De esta forma, como indican Ato, Losilla, Navarro, Palmer y Rodrigo (2000 b) ,para corregir la infraestimación de los errores estándar de los parámetros del predictor lineal producida por la sobredispersión, se pueden multiplicar éstos por la raíz cuadrada del parámetro de escala.

También es posible la estimación del error estándar a través de técnicas relacionadas con la simulación como (Cameron y Trivedi, 1998): Sandwich, Jackknife, Bootstrap.

3.4.3.2. Estimación del error estándar

- **Estimador Sandwich**

El estimador Sandwich fue propuesto inicialmente por Huber (1967), Eicker (1967) y White (1980). En las últimas décadas este método ha sido muy usado dentro del contexto de las ecuaciones de estimación generalizadas para datos longitudinales, Liang y Zeger (1986) y Zeger y Liang (1986)

El estimador sandwich, también conocido como estimador robusto de matrices de covariancias, tiene como objetivo la estimación consistente de la variancia. Cabe señalar que la estimación eficiente de parámetros requiere la especificación de la estructura de correlaciones entre las observaciones, la cual normalmente, es desconocida. Por consiguiente, la llamada matriz de covariancias de trabajo se usa en la estimación por pasos, así la estimación de la variancia se combina con su versión empírica correspondiente, en forma de sandwich. Es decir, la información intercala la covariancia empírica del vector de puntuaciones. Esta aproximación da estimaciones consistentes de la matriz de covariancias, tanto si trabajamos con covariancias no especificadas como si trabajamos bajo heterocedasticidad de errores. (Kauermann y Carroll, 2001)

El principal argumento a favor del estimador sandwich es que la normalidad asintótica y la cobertura asintótica de los intervalos de confianza requieren únicamente una estimación consistente de la variancia. Por otro lado uno de los problemas que parece presentar, según Kauermann y Carroll (2001) es un incremento de la variabilidad, presentando variancias mayores que las estimadas por métodos clásicos de estimación.

$$\text{Var}_s\left(\hat{\beta}\right) = \mathrm{I}^{-1}G\mathrm{I}^{-1} = \mathrm{I}^{-1}\left[\sum_{i=1}^{n} w_i^2\left(\frac{y_i-\mu_i}{v(\mu_i)}\right)^2 x_i x_i^t\right]\mathrm{I}^{-1} \qquad (64)$$

Donde I es la matriz de información de Fisher sin el factor de escala.

Breslow (1996) señala que el sandwich con tamaños de muestra moderados, a menudo puede infraestimar la variabilidad real.

HC0 es la forma más común del sandwich, conocida también como *White estimator* y *Huber estimator*. Como ha demostrado White (1980) entre otros, HC0 es un estimador consistente de la variancia en presencia de sobredispersión o heterocedasticidad, que se obtiene por medio de:

$$\text{HC0} = (X'X)^{-1}X'\hat{\Phi}X\,(X'X)^{-1} = (X'X)^{-1}X'\text{diag}\,(e^2_i)X\,(X'X)^{-1} \qquad (65)$$

donde $\hat{\Phi}$ = diag { e_1^2,.....e_n^2}. Esto es, $\hat{\Phi}$ es la diagonal de la matriz del vector de errores mínimo cuadrados. Ver White (1980). Sin embargo, puede no funcionar correctamente en muestras pequeñas; ver Cribari-Neto (2004); Cribari-Neto y Zarkos (1999, 2001) y MacKinnon y White (1985).

El estimador de la variancia sandwich se obtiene mediante la expresión:

$$\text{Var}_s\,(\hat{\beta}) = \mathrm{I}^{-1}\left[\sum_{i=1}^{n} w_i^2\left(\frac{y_i-\mu_1}{v(\mu_i)}\right)^2 x_i x_i^t\right]\mathrm{I}^{-1} \qquad (66)$$

donde I es la matriz de información de Fisher

MacKninnon y White (1985) proponen tres estimadores alternativos diseñados para mejorar las características de HC0 en muestras pequeñas.

El ajuste más simple, sugerido por Hinkley (1977) obtiene una corrección utilizando los grados de libertad, concretamente ajusta aplicando el factor de escala n/(n-K). Esta versión es conocida como HC1.

$$HC1 = (n/(n-p))\ (X'X)^{-1}X'\text{diag}(e^2_i)X\ (X'X)^{-1} = (n/n(n-p))HC0 \qquad (67)$$

Donde n es el número de observaciones en la variable dependiente y p es el numero de predictores.

Un segundo ajuste propuesto por Belsley, Kuh y Welsch (1980) y Wu(1986) introduce un factor de escala $1/(1-h_i)$ basándose en el análisis de valores influyentes y valores alejados. Esta versión es conocida como HC2:

$$HC2 = (X'X)^{-1}X'\text{diag}\left(\frac{e_i^2}{1-h_{ii}}\right) X\ (X'X)^{-1} \qquad (68)$$

donde h_i es el elemento i de la diagonal de la "matriz sombrero" $H= X(X'X)^{-1}X'$, $i =1,\ldots\ldots n$.

Un tercer ajuste, sugerido también por Mackninnon y White (1985) aproxima el sandwich al estimador Jackknife de Efron (1982). Incluye el parámetro de escala $1/(1-h_i)^2$.

$$HC3 = (X'X)^{-1}X'\text{diag}\left(\frac{e_i^2}{(1-h_{ii})^2}\right) X\ (X'X)^{-1} \qquad (69)$$

Cribari-Neto (2004) propone un cuarto ajuste que, según este autor, funcionará mejor que HC2 y HC3, cuando aparezcan valores alejados y valores influyentes.

$$\text{HC4= } (X'X)^{-1}X'\text{diag}\left(\frac{e_{in}^2}{(1-h_{in})^{\delta_{in}}}\right)X\,(X'X)^{-1} \qquad (70)$$

Donde:

$$\delta_{in} = \min\left\{4, \frac{h_i}{\bar{h}}\right\} \qquad (71)$$

podemos usar h_i como una medida de influencia de las ith observaciones.

- **Estimador Jackknife**

El remuestreo Jackknife es una técnica ideada por Maurice Quenouille (1949,1956) y perfeccionada por John W. Tukey (1958), quien la denominó con este nombre por su carácter de aplicabilidad general para la estimación del sesgo y del error estándar de un estadístico. El nombre alude a esas navajas multiuso que son a la vez sacacorchos, abrelatas, destornillador..., buenas para todo si no se tiene algo mejor a mano.

Este método, como señala Losilla(2002) consiste básicamente en, definida una muestra de observaciones de tamaño n, suprimir cada vez un conjunto de observaciones g y calcular sobre el conjunto (n-g) restante de datos el estadístico de interés. La aplicación más generalizada se basa en excluir cada vez una única observación, debido a que como indica Miller (1974) se evita sí la arbitrariedad en la formación de subgrupos y parece haberse mostrado como la forma óptima de aplicación.

Una vez obtenidas las n estimaciones del estadístico para cada una de las muestras Jackknife, se puede calcular una estimación del sesgo asociado a la muestra original, así como una estimación no paramétrica del error estándar del estadístico.

Calculamos ahora el parámetro de interés ϑ en la muestra total de tamaño *N*; en este caso las diferentes muestras se van construyendo eliminado cada vez una de las observaciones X_i y en ellas se calcula de nuevo el valor del parámetro de interés ϑ_i, repitiendo el proceso para *i=1* hasta *N*. Se obtiene lo que se denominan pseudo valores:

$$J_i = N \bullet \vartheta - (N-1) \bullet \vartheta_i$$

Se llama estimador Jackknife a la media de esos pseudo valores y a partir de la distribución obtenida se calcula también su variancia y un intervalo de confianza para el mismo.

Estimación Jackknife del error estándar:

$$\hat{\sigma}_j = \sqrt{\frac{(n-1)}{n}\sum_{i=1}^{n}\left[\hat{\theta}_{-i} - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}\right)\right]^2} \qquad (72)$$

$$\text{Var}_{\text{jackknife}} = \frac{n-p}{n}\sum_{i=1}^{n}\left(\hat{\beta}^{*}_{(i)} - \hat{\beta}_{(.)}\right)\left(\hat{\beta}^{*}_{(i)} - \hat{\beta}_{(.)}\right)^{t} \qquad (73)$$

$$\hat{\beta}^{*}_{(i)} = \hat{\beta} - \frac{\left(X^{T}\hat{V}X\right)^{-1} x_i r_i^{R}}{1-\hat{h}_i} \qquad (74)$$

donde $\hat{V} = \text{diag}(V-(\hat{\mu}))$ es una matriz (n x n). X es una (n x p) matriz de covariadas, $\hat{h}_i$ es la ith diagonal de la matriz hat. $\hat{\beta}$ es el vector de estimación de coeficientes usando todas las puntuaciones y $\hat{r}_i^{R}$ es la estimación de respuesta residual.

- **Estimador Bootstrap**

En el año 1979, a partir de un profundo análisis teórico y aplicado sobre el origen y evolución de los métodos estadísticos, Bradley Efron desarrolla el Bootstrap. Este término, que significa “levantarse tirando hacia arriba de las propias correas de las botas”, refleja el principal aspecto de esta técnica: autosuficiencia.

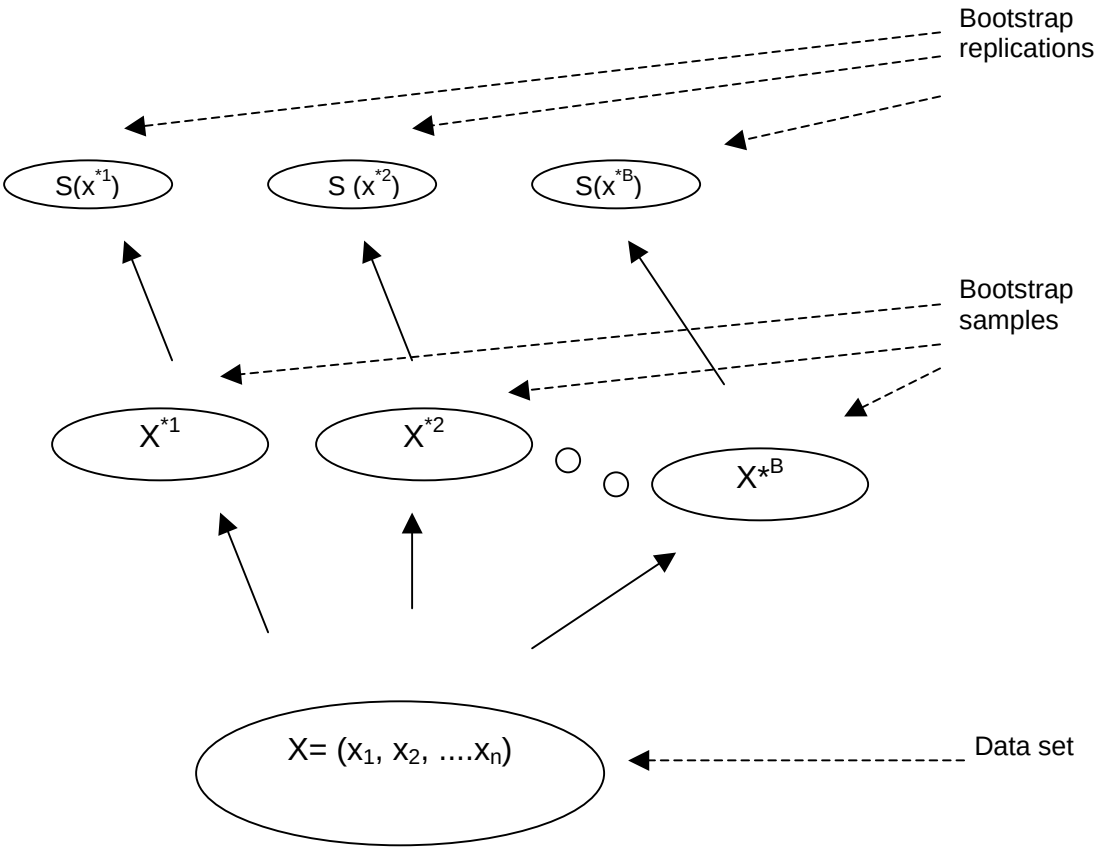


Figura 1. Proceso esquemático del funcionamiento del Bootstrap. Reproducido de Efron y Tibshirani, (1993, p.13).

El método Bootstrap aunque ya usado anteriormente, fue descrito de forma sistemática por Efron en 1979. El nombre alude al cordón de los zapatos, recordando la imagen de alguien intentando salir del barro tirando del cordón de sus propios zapatos, y consiste, si tenemos una muestra de tamaño *N*, en generar un gran número de muestras de tamaño *N* efectuando un muestreo con reemplazamiento de esos valores. Es como si metiésemos los valores en una urna, extraemos una papeleta, anotamos el resultado, y volvemos a colocarlo en la urna, y así hasta obtener *N* valores. En esa muestra calculamos el valor del parámetro que estamos estimando. Y así repetimos el proceso un gran número *B* de veces (por ejemplo 10000 o más), con lo que obtenemos una distribución de valores para el parámetro en la que podemos calcular su dispersión (análogo del error estándar) y determinar unos límites de confianza utilizando esa distribución.

Los procedimientos Bootstrap constituyen, junto con las pruebas de aleatorización el grupo más importante de técnicas de remuestreo. El Bootstrap se basa en remuestrear con reposición a partir de la muestra original para obtener nuevas muestras.

Originariamente Efron describe el Bootstrap como una técnica útil para evaluar el error estadístico y para realizar estimaciones de parámetros; sin embargo, posteriormente se realizan adaptaciones y desarrollos que van desde la obtención del grado de significación en el contraste de hipótesis (Lunneborg, 1985, 1999; Wilcox y Muska, 1999) hasta la estimación y corrección del error estándar de los coeficientes, etc...

Los métodos Bootstrap dependen de la noción de *muestra Bootstrap.* Tenemos que $\hat{F}$ es una distribución empírica, con una probabilidad de 1/n en cada uno de los valores observados x_i, i = 1,2,3,.....,n. Una *muestra Bootstrap* es definida como una muestra aleatoria de tamaño n obtenida de $\hat{F}$, decimos:

$$x^* = (x^*_1, x^*_2, ... x^*_n)$$

$$\hat{F} \rightarrow (x^*_1, x^*_2, ... x^*_n)$$

Los asteriscos indican que se tratan de muestras extraídas a partir de x, es decir, son muestras aleatorias de tamaño n obtenidas con reemplazamiento desde la población de n objetos (x_1, x_2, x_3....., x_n), con idéntica función de distribución F, es la estimación máximo verosímil no paramétrica de F. En este sentido, Bickel y Freedman (1981) en su estudio sobre las propiedades asintóticas del Bootstrap demuestran el postulado anterior a partir del teorema de Glivenko-Cantelli, principio que permite establecer una convergencia casi segura y monótona, cuando $n \rightarrow \infty$, entre las distribuciones F y $\hat{F}$. Consecuentemente, se puede considerar que si $\hat{F} \cong$ F, entonces $\hat{J} \cong$ J (distribución del estadístico $\hat{\theta}$) proposición que sustenta el "Bootstrap no paramétrico".

El algoritmo de la estimación Bootstrap no paramétrica trabaja obteniendo muestras Bootstrap y calculando el error estándar correspondiente para cada una de las muestras. El resultado se llama estimación Bootstrap del error estándar, representado por $\hat{se}_B$, donde B representa el número de muestras Bootstrap usadas. Como señala Losilla (2002) la estimación Bootstrap es independiente de $\hat{\theta}$, y los estudios teóricos demuestran que coincide con el verdadero error estándar σ cuando B $\rightarrow \infty$ (Efron, 1982 ;Efron y Tibshirani, 1986). En este sentido establecen que el número B de muestras Bootstrap adecuado para la estimación del error estándar puede oscilar entre 50 y 200 en la mayor parte de los casos

En el caso de que conozcamos la función de distribución F_θ a partir de la cual se han extraído los datos, pero no conozcamos su parámetro θ, podemos recurrir a la estimación de este parámetro considerando que $\hat{\theta}$ es una buena estimación de θ, obtenida a partir de los datos muestrales y, por tanto, que $F_\theta \approx F_{\hat{\theta}}$. En esta situación en lugar de extraer muestras aleatorias a partir de la muestra original X_0, se pueden extraer directamente por muestreo Monte Carlo sobre la base del modelo de distribución $F_{\hat{\theta}}$, este proceso se corresponde con el llamado "Bootstrap paramétrico" cuyo algoritmo general es igual al del Bootstrap no paramétrico, excepto en el punto de generación de muestras.

3.4.4. Solución mediante modelado

Otra forma de acabar con la sobredispersión es el modelado de los datos con otros modelos que tengan en cuenta las características de éstos. Así nos encontramos con una gran variedad de modelos, algunos cuyo objetivo es el abordaje de las causas de la especificación errónea y otros que pretenden abordar las consecuencias.

• Modelos no específicos:

- Modelos Mixtos: Son modelos que parten del hecho de que la causa más frecuente de sobredispersión es que el mecanismo generador de datos no es Poisson.

A) Paramétricos: Adoptan una función de densidad compuesta no poissoniana, como la binomial negativa (resultado de la mezcla de Poisson y su conjugada gamma), la distribución mixta gaussiana inversa (Dean, Lawless y Willmot, 1989), o la distribución mixta log-normal (Aitchison y Ho, 1989; Winkelmann, 2000). La principal ventaja de los modelos mixtos paramétricos es el incremento en la eficiencia, pero su principal desventaja es la potencial pérdida de consistencia si la asunción paramétrica es incorrecta (Winkelmann, 2000). El modelo de regresión binomial negativa (MRBN) es, sin duda, el miembro de esta categoría que se aplica con mayor frecuencia.

B) Semi-paramétricos: Otra solución pasa por la estimación semiparamétrica, que utiliza métodos matemáticos como polinomios de Laguerre (Gurmu, Rilstone y Stern, 1998) y aproximaciones discretas (Brännäs y Rosenqvist, 1994) a la distribución de la heterogeneidad no observada (*u).* La principal ventaja de estos modelos es la ganancia en robustez debida al hecho de que parten de asunciones menos restrictivas, aunque, por otra parte, presentan una cierta pérdida de eficiencia.

- Modelos con variancia generalizada. Son modelos genéricos en cuanto al origen de la especificación errónea. El objetivo de estos modelos es abordar las consecuencias de los errores de especificación, esto es, sobredispersión.

• Modelos específicos:

Son modelos diseñados para fuentes de especificación errónea concretas, de forma que su ámbito de aplicación es, en principio, más reducido que los otros dos tipos de modelos. Estos modelos, denominados “extensiones” en muchos casos, son modificaciones de un modelo de referencia que, habitualmente, son el MRP y el MRBN.

• Modelos de Clase Latente

Estos modelos son más flexibles que los modelos mixtos ya que permiten ser mixtos tanto respecto a los ceros como a los valores positivos. El MBN está anidado en los modelos mixtos y en los modelos de clase latente (MCL), pero los modelos mixtos y los modelos de clase latente, entre ellos, están relacionados pero no anidados. Como señalan Deb y Trivedi (2002) no hay una clara prioridad de que modelo trabaja mejor empíricamente ya que en diferentes trabajos, Deb y Trivedi (1997), Cameron y Trivedi (1998) llegaron a resultados contrarios. Tanto el MCL como los modelos mixtos requieren que el investigador especifique la distribución de probabilidad de los datos. Esto es una fuente de especificación errónea en ambos casos, aunque su impacto es menor en los MCL (Deb y Trivedi, 2002), por ser más flexibles.

Un modelo mixto puede ser simplemente una forma de flexibilizar y facilitar el modelado de los datos, donde cada componente mixto da una aproximación a alguna parte de una distribución real. Por otra parte los Modelos de Clase Latente pueden ajustar los datos mejor, simplemente porque los datos contienen valores extremos y valores influyentes. Los MCL captan este hecho a través de componentes mixtos adicionales.

- **Modelo de regresión binomial negativa (MRBN)**

Una forma de relajar la conocida restricción de igualdad media-variancia del MRP es especificar una distribución que permita un modelado más flexible de la variancia. En este sentido, el modelo paramétrico estándar para datos de recuento con presencia de sobredispersión es el modelo de regresión binomial negativa (MRBN) (Cameron y Trivedi, 1998; Nakashima, 1997). Aunque el MRBN puede ser derivado de diversas formas y con diversos objetivos, consideraremos aquí la situación más común que es la caracterizada por un conjunto de datos distribuidos según una distribución de Poisson, cuya media está especificada de forma incompleta debido a una situación de heterogeneidad no observada, y esta media es considerada como una variable aleatoria que en la población sigue una distribución gamma (Cameron y Trivedi, 1998; McCullagh y Nelder, 1989); a continuación, se desarrolla esta última.

Mientras que en el MRP la media condicional de *y* es $\mu_i = \exp(x_i \beta)$, en el MRBN, la media μ es reemplazada por la variable aleatoria $\tilde{\mu}$ (Long, 1997), de forma que se obtiene la siguiente ecuación estocástica:

$$\tilde{\mu}_i = \exp(x_i \beta + \varepsilon) \qquad (75)$$

donde se asume que ε no está correlacionado con *x.* El término de error ε puede ser el resultado del efecto conjunto de variables no incluidas en el modelo (Gourieroux et al., 1984a) o bien una fuente de aleatoriedad intrínseca (Hausman et al., 1984). Sea cual sea su origen, ε representa la heterogeneidad no observada. En el MRP, la variación en μ es introducida a través de la *heterogeneidad observada*, de forma que diferentes valores de *x* resultan en diferentes valores de μ. Así, todos los individuos con el valor x_i tienen la misma μ_i. En el MRBN, la variación en $\tilde{\mu}$ es debida tanto a la variación en x_i entre los individuos, como a la *heterogeneidad no observada* introducida a través de ε. Para una combinación de valores en las variables independientes, existe una distribución de diversas $\tilde{\mu}$ en lugar de una μ única.

La relación entre la $\tilde{\mu}$ y la μ "original" es:

$$\tilde{\mu} = \exp(x_i\beta)\exp(\varepsilon_i) = \mu_i\exp(\varepsilon_i) = \mu_i\delta_i \qquad (76)$$

donde δ_i se define como equivalente a $\exp(\varepsilon_i)$. La corrección del MRBN depende de la especificación de una asunción acerca de la media del término de error. La asunción más conveniente es que (Long, 1997):

$$E(\delta_i) = 1 \qquad (77)$$

Esta asunción implica que el recuento esperado después de añadir la nueva fuente de variación es el mismo que para el MRP:

$$E(\tilde{\mu}_i) = E(\mu_i\delta_i) = \mu_i E(\delta_i)\ \mu_i\delta_i) = \mu_i \qquad (78)$$

Por otro lado, la distribución de las observaciones dados x y δ es también Poisson:

$$\Pr(y_i|x_i,\delta_i) = \frac{\exp(-\tilde{\mu}_i)\tilde{\mu}_i^{Y_i}}{y_i!} = \frac{\exp(-\mu_i\delta_i)(\mu_i\delta_i)}{y_i!} \qquad (79)$$

Sin embargo, puesto que δ es desconocido no podemos calcular Pr(y | xδ). Para calcular Pr(y | x) sin tener en cuenta δ, promediamos Pr(y | $x\delta$) por la probabilidad de cada valor de δ. Si g es la función de densidad de probabilidad de δ, entonces la densidad marginal de y_i puede ser obtenida integrando con respecto a δ_i (Cameron y Trivedi, 1986; Long, 1997):

$$\Pr(y_i \mid x_i) = \int_0^\infty \left[\Pr(y_i|x_i,\delta_i)\right] g(\delta_i)d\delta_i = \int_0^\infty \frac{e^{-\exp(x_i\beta+\delta_i)}\exp(x_i\beta+\delta_i)^{y_i}}{y_i!} g(\delta_i)d\delta_i \qquad (80)$$

Esta expresión define la distribución de Poisson compuesta (Cameron y Trivedi, 1986). Tal como indican estos mismos autores, las distribuciones de Poisson compuestas proporcionan una generalización natural de los modelos de Poisson básicos y, su aplicación obedece generalmente a una necesidad de mayor flexibilidad, especialmente en situaciones de sobredispersión.

La ecuación de la distribución de Poisson compuesta (80) calcula la probabilidad de *y* como una mezcla de dos distribuciones de probabilidad (Long, 1997). Asimismo, la forma de (80) depende de la selección de g(δ_i), es decir, de la función de densidad de probabilidad que se asuma para δ_i. En este sentido, Long (1997) afirma que la asunción más común es que δ_i sigue una distribución gamma con el parámetro v_i:

$$g(\delta_i) = \frac{v_i^{v_i}}{\Gamma(v_i)} \delta_i^{v_i - 1} \exp(-\delta_i v_i) \qquad (81)$$

para $v_i > 0$

donde la función gamma se define como $\Gamma(v) = \int_0^\infty t^{v-1} e^{-t} dt$. Cuando se asume que g($\delta_i$) sigue una distribución gamma, la integración de la ecuación de la regresión de Poisson compuesta conduce a una distribución binomial negativa.

Tal como indica Poortema (1999), la distribución binomial negativa es la distribución compuesta resultante si la conjugada de la distribución de Poisson, la distribución gamma, es utilizada para la composición. Johnson, Kotz y Balakrishnan (1994) demuestran que si δ_i sigue una distribución gamma, entonces *E*(δ_i) = 1, ecuación que coincide con la asunción del MRBN expuesta anteriormente, y Var(δ_i) = 1/v_i. El parámetro *v* también afecta a la forma de la distribución, de manera que a medida que *v* aumenta la distribución se va aproximando a una distribución normal centrada alrededor de 1 (Long, 1997). La

distribución de probabilidad binomial negativa se define como (Long, 1997; Nakashima, 1997):

$$\Pr(yi\ Ixi) = \frac{\Gamma(y_i + v_i)}{\Gamma(y+1)\Gamma(v_i)}\left(\frac{v_i}{v_i + \mu_i}\right)^{v_i}\left(\frac{\mu_i}{v_i + \mu_i}\right) \qquad (82)$$

para $y_i = 0,1,2,3.....$

El valor esperado de *y* para la distribución binomial negativa es el mismo que para la distribución de Poisson:

$$E(y_i \mid x_i) = \exp(x_i\beta) = \mu_i \qquad (83)$$

Sin embargo, la variancia condicional sí difiere con relación a la de la distribución de Poisson:

$$\Pr(y_i\ Ix_i) = \mu_i\left(1 + \frac{\mu_i}{v_{ii}}\right)^{v_i} = \exp(x_i\beta)\left(1 + \frac{\exp(x_i\beta}{v_{ii}}\right) \qquad (84)$$

Puesto que $\mu > 0$ y $v > 0$, la variancia condicional de *y* en el MRBN será mayor que la media condicional $\exp(x_i\beta)$ (Cameron y Trivedi, 1986; Long, 1997). Obsérvese que a medida que *v* aumenta, la distribución tiende a la equidispersión puesto que $Var(y \mid x) \rightarrow \mu$. Por otro lado, una variancia condicional elevada en *y* incrementa la frecuencia relativa de valores de recuento altos y bajos. De esta forma, en una situación de sobredispersión, la distribución binomial negativa corrige, especialmente, la probabilidad asociada a valores bajos de recuento que, habitualmente presentan un ajuste deficiente a través del MRP (Long, 1997).

El problema de (84) es que si *v* varía entre individuos, entonces existen más parámetros que observaciones. La solución más común pasa por asumir que *v* es común para todos los individuos (Long, 1997):

$$V_i = \alpha^{-1} \tag{85}$$

para $\alpha > 0$

De esta forma, la densidad queda reexpresada como:

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \tag{86}$$

Por otro lado, la asunción (85) implica que la variancia de δ es constante. Al hacer efectiva la igualdad $v_i = \alpha^{-1}$, se hace evidente que al incrementar α, que es conocido como el parámetro de dispersión, aumenta la variancia condicional de *y*:

$$\mathrm{Var}(y_i \mid x) = \mu_i\left(1+\frac{\mu_i}{\alpha^{-1}_i}\right)^{\alpha^{-1}} = \mu_i(1+\alpha\mu_i) = \mu_i + \alpha\mu_i^2 \tag{87}$$

Obsérvese que si el parámetro de dispersión α = 0, habría equidispersión o variancia de Poisson, puesto que $\mathrm{Var}(y \mid x) = \mu+\alpha\mu^2 = \mu$.

La densidad (86) y la variancia (87) caracterizan la especificación estándar de un MRBN, que corresponde al denominado modelo Negbin II (Cameron y Trivedi, 1986).

El modelo Negbin I tiene una distribución de probabilidad,

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1}\mu_i)}{\Gamma(y+1)\Gamma(\alpha^{-1}\mu_i)}\left(\frac{\alpha^{-1}\mu_i}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}\mu_i+\mu_i}\right)^{y_i} \tag{88}$$

y su variancia condicional es:

$$\mathrm{Var}(y_i \mid x) = \mu_i + \alpha\mu_i = (1+\alpha)\mu_i \tag{89}$$

Además de los modelos Negbin I y Negbin II considerados anteriormente, algunos autores como Cameron y Trivedi (1986) o Winkelmann y Zimmerman (1991) proponen el denominado hipermodelo Negbin *k*, en el cual Var(*y* | *x*) = μ + αμ^{2-k}

Estimación

La función de verosimilitud del MRBN estándar, esto es, Negbin II, es (Long, 1997):

$$L(\beta;Y.X) = \prod_{i=1}^{n} \Pr(y_i|x_i) = \prod_{i=1}^{n} \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \qquad (90)$$

y= 0,1,2....

donde $\alpha \geq 0$, y $\mu = \exp(x_i\beta)$. Después de tomar los logaritmos, se obtiene la función de log-verosimilitud (Cameron y Trivedi, 1998):

$$\ln L\ (\beta;\ Y,\ X) \qquad (91)$$

La función log-verosímil para el modelo Negbin I es (Cameron y Trivedi, 1998):

$$\ln L\ (\beta;\ Y,\ X) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1}\mu_i\right) - \ln y_i! - (y_i+\alpha^{-1}\mu_i)\ln(1+\alpha) + y_i\ln\alpha\right\} \qquad (92)$$

Interpretación

Los métodos de interpretación basados en el recuento esperado *E*(y | x) son idénticos a los usados en el MRP, puesto que las estructuras de la media son las mismas. Los cálculos de las probabilidades predichas están basados en

$$\widehat{\Pr}(y|x) = \frac{\Gamma(y_i+\widehat{\alpha}^{-1})}{\Gamma(y+1)\Gamma(\widehat{\alpha}^{-1})}\left(\frac{\widehat{\alpha}^{-1}}{\widehat{\alpha}^{-1}+\widehat{\mu}_i}\right)^{\widehat{\alpha}^{-1}}\left(\frac{\widehat{\mu}_i}{\widehat{\alpha}^{-1}+\widehat{\mu}_i}\right)^{y_i} \qquad (93)$$

para el modelo Negbin II, y en

$$\widehat{\Pr}(y|x) = \frac{\Gamma\left(y_i + \widehat{\alpha}^{-1}\widehat{\mu}\right)}{\Gamma(y+1)\Gamma\left(\widehat{\alpha}^{-1}\widehat{\mu}\right)}\left(\frac{\widehat{\alpha}^{-1}\widehat{\mu}}{\widehat{\alpha}^{-1}+\widehat{\mu}_i}\right)^{\widehat{\alpha}^{-1}}\left(\frac{\widehat{\mu}_i}{\widehat{\alpha}^{-1}\widehat{\mu}+\widehat{\mu}_i}\right)^{y_i} \quad (94)$$

para el modelo Negbin I.

- **Modelos con variancia generalizada**

Una característica común a los modelos generalizados de datos de recuento que se presentan a continuación, es su falta de especificidad con respecto al origen del error de especificación que produce la sobredispersión. Por otro lado, un aspecto importante de estos modelos es que admiten tanto sobredispersión como infradispersión.

Partiendo del hecho de que la mayor parte de los errores de especificación producen una violación del supuesto de equidispersión, una estrategia consiste en evitar tal restrictividad impuesta por el MRP haciendo uso de una función variancia generalizada e incorporando dicha función de variancia en un modelo de datos de recuento.

Uno de los modelos de variancia generalizada es el modelo de recuento de eventos generalizado (*Generalizad Event Count, GECk*) introducido por King (1989). Este modelo no será expuesto ya que, según señala Winkelmann (2000), su única ventaja con respecto al modelo Negbin *k*, es que admite tanto sobredispersión como infradispersión.

o **Regresión de Poisson generalizada**

La regresión de Poisson generalizada (RPG) (Consul y Famoye, 1992; Famoye, 1993; Wang y Famoye, 1997) deriva de la distribución de Poisson generalizada (Consul, 1989) y constituye una alternativa al modelo de recuento de eventos generalizado puesto que además de admitir tanto sobredispersión como

infradispersión, anida al MRP como un caso especial Bae, Famoye, Wulu, Bartolucci y Singh (2005).

Su función de densidad es:

$$f(y_i) = \left(\frac{\mu_i}{1+a\mu_i}\right)^{y_i} \frac{(1+a\mu_i)^{y_i-1}}{y_i!} \exp\left(-\frac{\mu_i(1+ay_1}{1+a\mu_i}\right) \qquad (95)$$

La media condicional de la RPG es

$$E(y_i \mid x_i) = \mu_i \qquad (96)$$

y su variancia condicional

$$Var(y_i \mid x_i) = \mu_i (1+a\mu_i)^2 \qquad (97)$$

a actúa como un parámetro de dispersión, de forma que:

• a < 0 indica infradispersión,

• a > 0 indica sobredispersión, y

• a = 0 indica equidispersión, en cuyo caso queda reducida al MRP.

La función de log-verosimilitud es (Winkelmann, 2000):

$$\ell = (a, \beta, y_i) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\mu_i}{1+a\mu_i}\right) + (y_i - 1)\ln(1+ay_1) - \frac{\mu_i(1+ay_i)}{1+a\mu_1} - \ln(y_i!) \right\} \qquad (98)$$

o **Regresión de Poisson robusta**

En el MRP la estructura de la media está relacionada con la estructura de la variancia. Si esta relación no se reproduce en los datos, una de las soluciones más utilizadas se basa en la introducción de un parámetro de dispersión que sea

capaz de modelar variación no poissoniana, de forma que Var (y_i | x_i) = $\varphi\mu_i$. De esta forma, la estimación pasa a ser semiparamétrica o robusta, en la cual no se asume un conocimiento exhaustivo de la distribución de los datos, de forma que las asunciones paramétricas son menos restrictivas (Poortema, 1999). Concretamente, sólo deben ser especificados el primer y segundo momentos de la distribución (Fahrmeir y Tutz, 2001).

Dentro de los métodos de estimación semiparamétrica, los aplicados con mayor frecuencia son la estimación quasi máximo-verosímil («*quasi-maximum likelihood*», *QML*), también denominada «quasi-Poisson», y la pseudo máximoverosímil («*pseudo-maximum likelihood*», *PML*). Winkelmann (2000, p. 84) advierte que la estimación QML es «*en general inconsistente e ineficiente*». Sin embargo, Gourieroux et. al.(1984b) indican que si la media está especificada correctamente y el modelo forma parte de la familia exponencial lineal, como es el caso de la distribución de Poisson o la binomial negativa, el estimador QML es consistente. Gourieroux et al. (op.cit.) denominan a este estimador pseudo máximo-verosímil (PML). Así, PML se considera un caso particular de QML en el que el error de especificación consiste en una función media correctamente especificada y una estimación basada en la familia exponencial de distribuciones.

La estimación PML se basa en el hecho de que, dada la pertenencia de la distribución de Poisson a la familia exponencial de distribuciones, las desviaciones de la función variancia estándar no afectan a la consistencia de los parámetros estimados, mientras la media esté especificada correctamente. De esta forma, se asume una media de Poisson mientras que se relaja la restricción poissoniana de equidispersión (Cameron y Trivedi, 1998). El único efecto del error de especificación de la función variancia es que la matriz de variancias estimada bajo la asunción máximo-verosímil resulta inadecuada y debe ser ajustada (Winkelmann, 2000). Es decir, los errores estándar de los parámetros MRP deben ser ajustados en presencia de sobredispersión (Winkelmann y Zimmermann, 1995).

Winkelmann y Zimmermann (1995) proponen la siguiente estrategia: partir del supuesto de consistencia de las estimaciones de los parámetros y calcular (asintóticamente) errores estándar válidos. Estos autores (op. cit) denominan a esta estrategia regresión de Poisson robusta. En realidad, tal como señala Winkelmann (2000), este método es equivalente a la estimación PML. La estimación PML de Poisson se define (Cameron y Trivedi, 1998) como la solución a:

$$\sum_{i=1}^{n}\left[y_i - \exp(x_i'\beta\right]x_i = 0 \qquad (99)$$

Si se cumple $E(y_i \mid x_i) = \mu_i = \exp(x_i\beta)$, entonces

$$\hat{\beta}^a \sim N\left[\beta, Var_{PLM}\left(\hat{\beta}\right)\right] \qquad (100)$$

donde

$$\text{Var}_{PLM}(\hat{\beta}) = \left(\sum_{i=1}^{n}\mu_i x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n}\omega_i x_i x_i'\right)\left(\sum_{i=1}^{n}\mu_i x_i x_i'\right) \qquad (101)$$

$$\omega_i = \hat{Var}(y_i|x_i) \qquad (102)$$

Un punto crucial es la evaluación del término ω_i. Se pueden distinguir tres asunciones en relación a la función variancia (Winkelmann, 2000):

En ausencia de asunción (Breslow, 1990), la variancia estimada es:

$$\hat{Var}(y_i|x_i) = (y_i - \hat{\mu}_i)^2 \qquad (103)$$

Si se asume una función variancia lineal (McCullagh y Nelder, 1989):

$$\widehat{Var}(y_i|x_i) = \widehat{\delta}^2 \widehat{\mu}_i \quad (104)$$

donde $\widehat{\delta}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$

Si se asume una función variancia cuadrática (Gourieroux et al., 1984a)

$$\widehat{Var}(y_i|x_i) = \widehat{\mu}_i + \widehat{\delta}^2 \widehat{\mu}_i^2 \quad (105)$$

A través de regresión auxiliar puede obtenerse un estimador de δ^2

$$(y_i - \widehat{\mu}_i)^2 - \widehat{\mu}_i = \widehat{\delta}^2 \widehat{\mu}_i^2 + v_i \quad (106)$$

Mientras que en las dos primeras asunciones, la estimación PML de la distribución de Poisson usa la información disponible eficientemente, en la función variancia cuadrática no es así. Gourieroux et al (1984a) muestran que si se incorpora esta información relativa a la variancia en la estimación, se consigue una mayor eficiencia, y denominan a este procedimiento estimación pseudo máximo-verosímil cuasi-generalizada(«*quasi-generalized pseudo maximum likelihood estimation*», *QGPML*). En el caso en que la densidad verdadera sea miembro de la familia exponencial de distribuciones, la estimación QGPML es eficiente puesto que resulta asintóticamente equivalente a la estimación ML (Fahrmeir y Tutz, 2001; Winkelmann, 2000).

- **Modelos de datos de recuento truncados**

Las muestras con ceros truncados ocurren cuando las observaciones entran a formar parte de la muestra únicamente cuando ocurre el primer recuento. Las distribuciones de datos de recuento truncados, también denominadas *distribuciones positivas de Poisson* (Gurmu, 1991, p. 215), son aquellas en las cuales no se puede observar todo el rango de enteros positivos. Tal como indican

Gurmu y Trivedi (1992), esta situación suele ser debida a las características de la muestra seleccionada.

Los datos de recuento truncados pueden ser modelados como un proceso bimodal. La primera parte consiste en una distribución latente para *Y** y la segunda parte consiste en una variable indicadora binaria *c*, de forma que la distribución observada para *X* es truncada si *c* = 0, y no truncada si *c* = 1.

Como señala Winkelmann, (2000), se puede dar un conjunto de enteros positivos con exclusión del 0 {1,...,.}, hablaríamos así de una distribución de recuentos con *ceros truncados*, o se puede excluir un entero positivo, hablando así de recuentos con *truncamiento superior*.

Así, aunque el truncamiento puede ocurrir en cualquier valor de recuento, el truncamiento en el valor 0 aparece como el más frecuente (Cameron y Trivedi, 1998; Gurmu, 1991). Gurmu y Trivedi (1992) presentan una prueba de sobredispersión en modelos de recuento truncados.

Existen dos modelos para datos de recuento con ceros truncados, siendo cada uno de ellos una derivación o extensión del MRP o bien del MRBN.

o **MRP de ceros truncados**

En un MRP las probabilidades de valores de recuento 0 y positivos son, respectivamente:

$\Pr(y_i | x_i) = \exp(-\mu_i)$ para $y_i = 0$

$\Pr(y_i | x_i) = 1 - \exp(-\mu_i)$ para $y_i > 0$

La distribución de probabilidad para el MRP de ceros truncados se define como

(Long, 1997; Winkelmann, 2000):

$$\Pr(y|y_i>0, x_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!(1-\exp(-\mu_i))'} \quad (107)$$

Puesto que los recuentos con valor 0 están excluidos, el valor esperado se incrementa por la inversa de la probabilidad de un recuento positivo. De esta forma el valor esperado es (Cameron y Trivedi, 1998):

$$E(y|y_i > 0, x) = \frac{\mu_i}{1-\exp(-\mu_i)} \quad (108)$$

Por otro lado, la variancia es menor que en la distribución de Poisson sin truncar (Long, 1997):

$$\begin{aligned} \text{Var}(y \mid y_i > 0, x) &= \text{E}(y \mid y_i > 0, x_i)\,[1-\Pr(y_i = 0 \mid x_i)\,\text{E}(y \mid y_i > 0, x_i)] \\ &= \frac{\mu_i}{1-\exp(-\mu_i)}\left(1-\frac{\mu_i}{\exp(\mu_i)-1}\right) \end{aligned} \quad (109)$$

Puesto que μ (la media de la distribución no truncada) es mayor que cero, 0 < exp(–μ) < 1, de forma que el modelo truncado se desplaza a la derecha. Winkelmann y Zimmerman (1995) afirman que el MRP de ceros truncados tiende a presentar infradispersión con relación al MRP no truncado puesto que 0 < 1 - μ(exp(μ) - 1) < 1. Sin embargo, Grogger y Carson (1991) afirman que un modelo de recuento de ceros truncados también puede presentar sobredispersión, en cuyo caso las estimaciones de β resultan sesgadas e inconsistentes.

Famoye y Wang (2004) desarrollan el modelo de regresión de Poisson generalizado para datos truncados que puede usarse para modelar datos de recuento que exhiben sobre o infra dispersión, sin necesidad de saber el tipo de dispersión que presentan.

o **MRBN de ceros truncados**

En el caso del MRBN, las probabilidades de valores de recuento 0 y positivos son, respectivamente (Long, 1997):

$$\Pr(y_i \mid x_i) = (1+\alpha\mu_i)^{-\alpha^{-1}} \quad 1 \qquad \text{para } y_i = 0$$
$$\Pr(y_i \mid x_i) = 1 - (1+\alpha\mu_i)^{-\alpha^{-1}} \qquad \text{para } y_i > 0 \tag{110}$$

La distribución de probabilidad para el MRBN de ceros truncados se define como (Long, 1997):

$$\Pr(y \mid yi > 0, xi) = \frac{\dfrac{\Gamma\left(y_i + \alpha^{-1}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\alpha^{-1}\right)}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} \tag{111}$$

y=1,2,3.....

La media condicional y la variancia condicional se definen, respectivamente, como (Cameron y Trivedi, 1998; Grogger y Carson, 1991):

$$E(y \mid y_i > 0, x) = \frac{\mu_i}{1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}}} \tag{112}$$

$$\mathrm{Var}(y \mid y_i > 0, x) = \frac{\mu_i}{1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}}} \times \left(1 - (1+\alpha\mu_i)^{-\frac{1}{\alpha}} \frac{\mu_i}{1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}}}\right) \tag{113}$$

o **Modelos de datos de recuento con ceros modificados**

Los datos de recuento con ceros modificados son aquellos que presentan un exceso o bien una falta de ceros. El exceso de ceros es el fenómeno más frecuente y el que produce sobredispersión, la escasez de ceros, por su parte, produce infradispersión (Winkelmann y Zimmermann, 1995). Tanto los modelos de

recuento con ceros aumentados como los modelos de datos de recuento de umbral, expuestos más adelante, forman parte de un grupo de extensiones denominado modelos de ceros modificados.

A diferencia del MRBN que responde a la infrapredicción de ceros, en el MRP incrementando la variancia condicional y manteniendo intacta la media condicional, los modelos de recuentos de ceros modificados modifican la estructura de la media para modelar explícitamente la producción de valores de recuento 0. Esto es posible asumiendo que los valores 0 pueden ser generados por un proceso diferente al de los recuentos estrictamente positivos (Cameron y Trivedi, 1998; Long, 1997).

En los dos apartados siguientes se presentan los modelos específicos para datos de recuento con exceso de ceros: el modelo de datos de recuento de umbral y el modelo de recuento con ceros aumentados.

- **Modelo de datos de recuento de umbral**

Los *modelos de datos de recuento de umbral* («*Hurdle Count Data Models*») (Mullahy, 1986), contemplan la diferenciación sistemática en el proceso estadístico que gobierna las observaciones, en función de sí el valor de tales observaciones supera o no cierto umbral.

El modelo con datos de recuento de umbral más frecuente, denominado modelo *con ceros* («*With Zeros model*», WZ) (Melkersson y Rooth, 2000; Mullahy,1986), es el que fija el umbral a 0.

El modelo WZ parece tener, en la literatura, dos acepciones. En ambos casos se considera que el modelo WZ implica hacer uso de un modelo de Poisson compuesto. Tal composición implica dos distribuciones, las cuales describen dos tipos de observaciones separadas por un umbral. El valor de este umbral es también común (el valor 0); es después de cruzar el umbral cuando aparecen las diferencias entre ambas acepciones.

• *Acepción A* (Winkelmann, 2000; Winkelmann y Zimmermann, 1995; Yen, 1999). Al sobrepasar el valor de umbral se obtiene una distribución truncada en 0 y, por tanto, con valores estrictamente positivos.

• *Acepción B* (Cameron y Trivedi, 1998; Long, 1997; Mullahy, 1997). La distribución resultante de cruzar el umbral es una distribución de referencia («*parental distribution*»), como puede ser la distribución de Poisson o la binomial negativa, de forma que el valor 0 es posible y su probabilidad viene determinada por dicha distribución de referencia. Según indica Long (1997) el concepto básico de modelo WZ es retomado y ampliado en modelos de uso mucho más extendido denominados *modelos de ceros aumentados*.

Acepción A

El modelo WZ se basa en la diferencia sistemática entre los procesos estadísticos que gobiernan, por un lado las observaciones con valores de recuento 0 y, por otro, las observaciones con valores de recuento estrictamente positivos. Ello se consigue combinando el modelo dicotómico que rige el resultado binario de que un recuento sea 0 o superior a 0, y un modelo de Poisson truncado en 0 para valores estrictamente positivos (Winkelmann y Zimmermann, 1995).

Para una formulación del modelo de umbral, asúmase que $f1$ y $f2$ son funciones de distribución de probabilidad para enteros no negativos. Si $f1$ gobierna la parte del umbral (valor 0) y $f2$ el proceso una vez sobrepasado el umbral (>0) la función de probabilidad del modelo de umbral viene dada por (Winkelmann, 2000):

$$\begin{aligned} \Pr(y = 0) &= f1(0) \\ \Pr(y = k) &= \Phi \ f1(k) \text{ para } k = 1, 2, \ldots \end{aligned} \qquad (114)$$

$$\text{siendo } \Phi = \frac{1 - f_1(0)}{1 - f_1(0)}$$

De lo cual sigue que si *f*1 = *f*2, Φ = 1 y el modelo de umbral se convierte en el modelo de referencia (p.ej. la distribución de Poisson).

El valor esperado viene dado por:

$$E_h(y) = \sum_{K=1}^{\infty} k f_2(k) \Phi \tag{115}$$

Este valor difiere del valor esperado por el modelo de referencia en un factor de Φ. Si la probabilidad de cruzar el umbral es mayor que la suma de las probabilidades de recuentos positivos en el modelo de referencia, Φ será superior a 1, incrementando de esta forma el valor esperado del modelo de umbral con respecto al valor esperado en el modelo de referencia. La variancia es (Winkelmann, 2000):

$$\mathrm{Var_h}(y) = \sum_{K=1}^{\infty} k^2 f_2(k) \Phi - \left[\Phi \sum_{K=1}^{\infty} k f_2(k) \right]^2 \tag{116}$$

La razón variancia-media es (Winkelmann, 2000):

$$\frac{Var_h(y)}{E_h(y)} = \frac{\sum_{K=1}^{\infty} k^2 f_2(k) \Phi - \left[\Phi \sum_{K=1}^{\infty} k f_2(k) \right]^2}{\sum_{K=1}^{\infty} k f_2(k)} \tag{117}$$

Tal como se ha indicado, si Φ = 1 la razón media-variancia queda reducida a la del modelo de referencia. En este caso, es posible aplicar una prueba de sobredispersión para modelos anidados como, p. ej., la prueba LR (Winkelmann y Zimmermann, 1995). Así, si f_2 es una función de distribución de Poisson, existe equidispersión si Var(*y*) / E(*y*) = 1. Por otro lado, si f_2 es una función de distribución de Poisson y $\Phi \neq 1$, (114) define el modelo de umbral: Si 0 < Φ < 1 existe sobredispersión.

La función de log-verosimilitud es:

$$\ell = \sum_{y=0} \ln f_1(0) + \sum_{y>0} \ln\left[1 - f_1(0)\right] + \sum_{y>0} \left\{\ln f_2(y) - \ln\left[1 - f_2(0)\right]\right\} \quad (118)$$

Acepción B

El modelo WZ asume que la población consiste en dos grupos. La probabilidad de que una observación se encuentre en el grupo 1, que es el que presenta únicamente recuentos con valor 0, es ψ, y la probabilidad de que se encuentre en el grupo 2, donde se encuentran el resto de valores de recuento, es 1 – ψ. En el grupo 2 los recuentos se distribuyen según la distribución de Poisson (o según la distribución binomial negativa) y la probabilidad de un recuento igual a 0 viene determinada por dicha distribución. Es decir, en el grupo 2, los recuentos siguen una distribución de Poisson o una binomial negativa, de forma que, por ejemplo, en el caso de aplicar un MRP, la probabilidad de un recuento determinado viene dado por:

$$\Pr(Y=y_i | \mu_i) = \frac{\exp(-\mu_i)\mu_i^{Y_i}}{y_i!} \quad (119)$$

Donde $\mu = \exp(x_i\beta)$. En este grupo, los recuentos con valor 0 ocurren, por tanto, con probabilidad $\Pr(y = 0|\mu) = \exp(-\mu)$.

La probabilidad total de valores de recuento 0 es el resultado de una combinación de las probabilidades de 0 recuentos en ambos grupos, ponderada por la probabilidad de un individuo de pertenecer a ese grupo (Long, 1997; Mullahy, 1997):

$$\Pr(y_i = 0 \mid x_i) = \psi + (1-\psi)\exp(-\mu_i) \quad (120)$$

Puesto que el proceso de Poisson es aplicable sólo a una proporción 1–ψ de la muestra, la probabilidad de recuentos positivos debe ser ajustada (Long, 1997):

$$\Pr(y_i|x_i) = (1-\Psi)\frac{\exp(-\mu_i)\mu_i^{Y_i}}{y_i!} \qquad (121)$$

para $y > 0$

Winkelmann (2000) indica, con relación a la primera de las acepciones del modelo WZ expuesta anteriormente, la diferencia entre el modelo de datos de recuento con ceros aumentados y el modelo WZ es que en el primero $y_i = y_i^*$ para el rango entero de y_i^* y no sólo para los valores estrictamente positivos cuando el umbral se ha sobrepasado. De esta forma, en el modelo de datos de recuento con ceros aumentados se obtienen dos tipos de ceros: parte de ellos –la mayoría-, provienen de $c_i = 1$ y el resto cuando se dan las condiciones $c_i = 0$ y $y_i^* = 0$.

- **Modelos de datos de recuento con ceros aumentados**

El primer concepto de una distribución de ceros aumentados se originó en el trabajo de Rider (1961 a, b) y Cohen (1960, 1963), los cuales examinaban las características de una distribución de Poisson Mixta.

Estos modelos asumen la existencia de dos grupos latentes en los datos. Un grupo de “Siempre 0” que presenta un resultado de 0 con una probabilidad de 1, y otro grupo de “No siempre 0” que puede o no presentar un 0.

Una probabilidad binomial gobierna el resultado binario de que el recuento sea 0 o positivo (Cameron y Trivedi, 1998).

Sea c_i una variable de selección binaria que permite el tratamiento separado de valores de recuento 0 y valores de recuento estrictamente positivos, de forma que:

$$\begin{aligned} y_i &= 0 \quad \text{si } c_i = 1 \\ y_i &= y_i^* \quad \text{si } c_i = 0 \end{aligned} \qquad (122)$$

Si la probabilidad de que $c_i = 1$ es representada por ω_i, la función de probabilidad de y_i es (Winkelmann, 2000):

$$f(y_i) = \omega_i(1-d_i) + (1-\omega_i)g(y_i) \qquad (123)$$

$y_i = 0,1,2 \ldots$

donde $d_i = 1- c_i = \min \{y_i, 1\}$ y $g(y_i)$ es un modelo de recuento habitual como el MRP o el MRBN.

o ***Modelo de Poisson de ceros aumentados (ZIP)***

El modelo ZIP representa una forma de manejar datos con un exceso de ceros, desde su introducción formal por Lambert (1992) (quien amplió un trabajo de Johnson y Kotz (1969)), el uso de estos modelos ha crecido y actualmente puede encontrarse en diferentes campos: economía (Green, 1994), epidemiología (Heilbron, 1994), en ciencias políticas (Zorn, 1996), entre otros.

En esencia, como señalan Dalrymple, Hudson y Ford(2003) los modelos ZIP se caracterizan por ser un proceso dual, un caso especial de un modelo mixto de dos componentes, sin covariables en las probabilidades mixtas (Wang, Puterman y Cockburn(1998). Un componente es tomado como una distribución degenerativa, con una masa de 1 en y=0, mientras que el otro componente es un modelo de regresión de Poisson. El modelo ZIP es más restrictivo que el modelo mixto, en el sentido que únicamente permite ser mixto respecto a los ceros. La variable de respuesta es modelada como una combinación de la distribución de Bernoulli y la distribución de Poisson.

Combinando el modelo de recuento de Poisson con el proceso binario para el modelo ZIP (Long, 1997), las probabilidades de diferentes valores de recuento vienen dadas por:

$$\Pr(y_i \mid x_i) = \omega_i + (1-\omega_i)\exp(-\mu_i) \qquad \text{para } y_i = 0$$

$$\Pr(y_i \mid x_i) = (1-\omega_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \qquad \text{para } yi \geq 1 \tag{124}$$

La función de log-verosimilitud para el modelo ZIP es (Cameron y Trivedi, 1998):

$$\ell(\beta,\gamma) = \sum_{d_i=1} \ln(\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta)) + \sum_{d_i=1} y_i x_i'\beta - \exp(x_i'\beta) - \ln(y!) - \sum_{i=1}^{n} \ln(1+\exp(z_i'\gamma)) \tag{125}$$

o **Modelo de binomial negativa de ceros aumentados (ZINB)**

Como una extensión natural del ZIP, Heilbron (1994) y Green (1994) propusieron el uso del modelo ZINB. Este modelo es similar en su estructura al ZIP, las probabilidades de diferentes valores de recuento se obtienen a partir de:

$$\Pr(y_i \mid x_i) = \omega_i + (1-\omega_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \qquad \text{para } y_i = 0$$

$$\Pr(y_i \mid x_i) = (1-\hat{\omega})\,\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \qquad \text{para } y_i \geq 1 \tag{126}$$

4. Objetivos de la investigación

Como se ha señalado previamente, la evaluación del modelo se refiere a valorar si el modelo ajustado en la etapa de estimación y ajuste, es un modelo válido, más allá de que presente un ajuste adecuado a los datos. Se refiere a la adecuación de los aspectos implicados en la etapa de especificación. En este sentido se han de evaluar posibles errores de especificación del componente sistemático, de la distribución de probabilidad del componente aleatorio y de la relación asumida entre ambos componentes del modelo en la fase de especificación. Por otra parte se requiere evaluar la presencia de observaciones extremas (*outliers*) o influyentes (*influentials*), así como el conjunto completo de elementos que forman los supuestos bajo los cuales se valida el modelo.

El objetivo principal de la presente tesis fue el estudio de la etapa de evaluación en datos de recuento. Para ello se siguieron una serie de pasos.

Para la consecución del objetivo propuesto se siguieron los siguientes pasos:

- Primero queríamos demostrar como la aplicación de un modelo no adecuado a los datos, desemboca en falsos resultados.

- Tras aplicar el modelo de regresión Poisson, que en principio es el modelo adecuado a este tipo de datos, debíamos comprobar los supuestos de aplicación y corregir su incumplimiento.

- Puesto en que la práctica no existe consenso de aplicación de qué índice o estimador es adecuado para corregir el EE en caso de sobredispersión, nos planteamos el objetivo de reunir en un mismo escenario los principales índices y estimadores, realizando un estudio comparativo de la eficiencia y precisión de las principales pruebas que habitualmente se utilizan para este menester, bajo diferentes tamaños muestrales y diferentes grados de sobredispersión, generados desde distintos modelos estocásticos representativos de las causas habituales de

sobredispersión. Esta comparación se realizó a partir de los resultados de dos experimentos de simulación Monte Carlo en el entorno R.

- Tampoco parece estar claro el modelo más adecuado a aplicar cuando la elección es el modelado con otro modelo que de cuenta de la sobredispersión, por ello el siguiente objetivo que nos fijamos en la parte empírica de nuestro trabajo consistió en valorar los efectos de la aplicación de diferentes modelos. Para ello se aplicaron diferentes modelos adecuados a los datos, y se evaluó la adecuación de cada uno de ellos.

- Por último nos planteamos realizar un estudio de residuales, de gráficos evaluativos, valores influyentes y extremos. Puesto que en el ámbito del artículo aplicado no es posible desarrollar todos estos aspectos, ya que la propia revista limita la presentación de estos análisis, nos hemos referido a ellos en la parte de presentación de la Tesis.

II.- PUBLICACIONES

Overdispersion diagnostics in count data analysis models.

Overdispersion diagnostics in count data analysis models [1]

Jaume Vives *, Josep-Maria Losilla *, and Noelia Llorens **

*Universitat Autònoma de Barcelona, Spain.

**Universitat de les Illes Balears, Spain.

Summary

Count data are assumed to have a Poisson distribution, especially when there is no diagnostic procedure for checking this assumption. However count data rarely fits the restrictive assumptions of the Poisson distribution. The violation of much of such assumptions commonly results in overdispersion which invalidates the Poisson distribution. Undetected overdispersion may entail important misleading inferences so that its diagnostic becomes essential. In this study we evaluate different overdispersion diagnostic tests through two simulation experiments. In Experiment 1, we compare the nominal error rate under different sample size and λ conditions. Results show a remarkable performance of the χ^2 test. In Experiment 2, we compare the statistical power under different sample size, λ and overdispersion conditions. χ^2 and *LR* tests provided the higher statistical power.

Keywords: overdispersion; overdispersion diagnostic tests; count data; Poisson distribution; generalized linear model;

[1] This research was supported by Grants BSO2001-2518 and BSO2002-2513 from the Spanish Ministry of Science and Technology.

Counts can be defined as the number of events that occur on the same observation unit during a temporal or spatial interval (Lindsey, 1997).

The law of rare events states that the total number of events will fit, approximately, the Poisson distribution if an event may occur in any point of the time or the space interval under observation but the probability of occurrence at any given point is small (Cameron & Trivedi, 1998). Therefore, count data coming from small probability phenomena fit a Poisson distribution.

A Poisson process is characterized by two basic assumptions (Winkelmann, 2000):

a) Independence: the occurrence of an event at a particular moment is independent of the number of events that have already taken place, so that:

$$\Pr(Y = a \mid b) = \Pr(Y = a \mid 1\text{–}b)$$

b) Stationarity: the occurrence of an event at a particular moment $t + \Delta$ is independent of previous time t, so that:

$$\Pr(Y = a)_t = \Pr(Y = a)_{t+\Delta}$$

It can be derived that in a Poisson process the probability of at least one occurrence of an event is proportional to either (Lunneborg, 1994):

- the amount of time the count data observation extends over,
- the size of the space the count data observation extends over, or
- the size of the population.

This proportionality factor is a constant independent of time and is taken into account in the Poisson density function through the inclusion of an exposure variable or offset (Cameron & Trivedi, 1998; Rodríguez, 1998):

$$\Pr(Y = y_i \mid \lambda_i) = \frac{\exp(-\lambda_i t_i)(\lambda_i t_i)^{y_i}}{y_i!}$$

for $y_i = 0, 1, 2 \ldots, \infty$; $\lambda_i \in \Re^+$,

where $\lambda_i = \mu_i$ is the expected value of Y, E(Y), and t_i is the offset.

This one can be considered the generic expression of the Poisson density function so that when $t = 1$ the density function takes the simplified form:

$$\Pr(Y = y_i \mid \lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

The benchmark model to explain count response variables from a set of predictors is the Poisson regression model, which is generalized linear model whose random component is a Poisson distribution.

The Poisson regression equation that provides the expected counts is:

$$\mu_i = t_i \times \exp(x_i\beta)$$

This equation shows one of the assumptions of the Poisson regression model: the expected counts value is a deterministic function of the explanatory variables so that no unexplained randomness, i.e. an additive stochastic error representing additional heterogeneity, is allowed for.

The other assumptions of the Poisson regression model are derived from its distributional assumptions which, as stated before, are independence and stability of events.

Non-Poissonness

The violation of any of the assumptions of the Poisson distribution generally invalidates it so that count data lead to distributions other than Poisson (Winkelmann, 2000). Possible causes are:

- Occurrence dependence: Count data show occurrence dependence when all of the units of the sample initially have the same probability of occurrence but it is changed as the number of occurrences of events grows up. This "form of dynamic dependence between the occurrence of successive events" (Cameron & Trivedi, 1986, p. 31) is referred to as *occurrence dependence*, *state dependence* or *contagion*, being the latter the usual term used in Health Sciences (Ezell, Land, & Cohen, 2003; Lindsey, 1995; Navarro, Utzet, Puig, Caminal, & Martín, 2001). Contagion can be:

  - Positive, if prior occurrence of an event increases the probability of further occurrences.

  - Negative, if prior occurrences decrease the probability of further occurrences

  One of the persistent problems in count data modeling is to differentiate between true and false contagion in cross-section data: if a population is heterogeneous the probability of occurrence of an event may be constant over time but may differ across individuals. When this inter-individuals heterogeneity is unobserved (due to the omission from the Poisson regression model of explanatory variables) it can be practically impossible to distinguish between true contagion (occurrence dependence) and spurious contagion which habitually is the result of unobserved heterogeneity (Abbring, Chiappori, Heckman, & Pinquet, 2003; Cameron & Trivedi, 1998).

- Duration dependence: Counts depend on duration when non-occurrences of the event matters, so that the time elapsed since the last occurrence alters the probability of occurrence of future events. In a similar way of occurrence dependence, duration dependence can be positive or negative (Winkelmann, 2000).

- Non-Stationarity: The assumptions of the model may be violated because the probability of occurrence of an event changes due to exogenous variables while being unaffected by previous occurrences. This situation does not necessarily invalidate the Poisson distribution (Winkelmann, 2000).

Overdispersion

The normal distribution is determined by its two first moments, the mean μ and the variance σ^2. Therefore models with a normal response variable include a variance parameter which can be estimated from data, so that any degree of variability about the mean can be suited in the model.

Poisson distribution as well as other non-normal distributions, by contrast, does not have such an independent variance parameter. Instead, the variance is determined by the mean through some sort of fixed relationship (Krzanowski, 1998). One of the characteristic properties of the Poisson distribution is the identity relationship between mean and variance, which is referred to as *equidispersion*, i.e.:

$$\mathrm{Var}(y_i \mid x_i) = \mathrm{E}(y_i \mid x_i) = \lambda_i = \mu_i$$

Therefore, and by contrast to other multi-parameter distributions, a violation of this variance assumption entails a violation of the Poisson assumption (Winkelmann, 2000).

Departures from equidispersion can cause either:

- *Overdispersion* if $\mathrm{Var}(y_{i \mid} x_i) > \mathrm{E}(y_{i \mid} x_i)$, or
- *Underdispersion* if $\mathrm{Var}(y_{i \mid} x_i) < \mathrm{E}(y_{i \mid} x_i)$

As a result of the limitations imposed by the restrictive assumptions of the Poisson distribution equidispersion becomes quite uncommon in practice (Gardner, Mulvey, & Shaw, 1995; Krzanowski, 1998; McCullagh & Nelder, 1989; Winkelmann, 2000). Moreover, as McCullagh & Nelder (1989, p. 124) point out "overdispersion is the norm in practice and nominal dispersion the exception".

Overdispersion becomes important not only because of its common presence in applied research but, and specially, due to its consequences concerning inference. When the model used to analyze count data fails to take account of overdispersion standard errors can be underestimated and hence inferences about regression parameters mislead (Krzanowski, 1998).

Sources of overdispersion

There main reasons why count data can be overdispersed are (McCullagh & Nelder, 1989; Winkelmann, 2000):

- True positive contagion (occurrence dependence).
- Negative duration dependence.
- Unobserved heterogeneity, which has been succinctly described before, and is also referred to as *proneness* or *frailty* (Ezell et al., 2003; Navarro et al., 2001). Unobserved heterogeneity can be considered as a mean function misspecification as it is the result of the omission of some relevant explanatory variables without which

the model cannot account for heterogeneity in the proneness of the individuals to present a given event. This differential proneness or lack of interunit equiprobability of the occurrence of events does not influence future event occurrences, so the independence restriction remains unaltered. That is, subjects have constant but unequal probability of experiencing the event of interest.

- Excess zeros: A situation in which the zero counts exceed the predicted probability by the Poisson distribution is referred to as excess zeros. As Mullahy (1997) points out, although unobserved heterogeneity is typically considered apart from excess zeros, the latter may be considered a strict implication of the former.

Overdispersion diagnostics

The knowledge about the sources of overdispersion is necessary but it is not a way through which overdispersion can be detected or diagnosed. Overdispersion diagnostics can be performed through a wide variety of tests, which can be classified in three categories:

- Nested models tests: Overdispersion detection is based on testing a Poisson model against a more general parametric model, so that the former is said to be nested in the latter. Within this category three tests are used in the present study:

    - Likelihood ratio test (*LR*): This test is based on the difference between the restricted and the unrestricted log-likelihood values (Winkelmann, 2000):

$$LR = -2(\hat{\ell}_r - \hat{\ell}_{nr}) \sim \chi^2_{(df)}$$

    - Wald test: In contrast to the *LR* test, Wald test does not need the estimation of two models, the estimation of the unrestricted model is enough (Breslow, 1996; Winkelmann, 2000):

$$W = [\mathrm{R}\,\hat{\theta} - \mathrm{q}]' \, [\mathrm{R}\, \hat{Var}(\hat{\theta})\, \mathrm{R}']^{-1} [\mathrm{R}\,\hat{\theta} - \mathrm{q}]$$

The performance of *LR* and Wald tests have been found to be similar (see, e.g. Rothenberg, 1984) while in some other studies *LR* has been found to be better than (see, e.g. Tu & Zhou, 1999). In order to reduce complexity in the results we have excluded Wald test from our study.

- Lagrange multiplier test (*LM*): Instead of computing both the restricted and the unrestricted models, as in *LR* test, or computing the unrestricted model, as in Wald test, the *LM* test or score test, is based on the estimation of the restricted model (Gurmu & Trivedi, 1992; Long, 1997; Winkelmann, 2000)ï

$$LM = \sqrt{\left[\sum_{i=1}^{n} \frac{1}{2\hat{\mu}_i^2}\right]} \; \frac{1}{2}\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 - y_i \text{ [JVB1]}$$

- Regression based tests: Cameron and Trivedi (1990) introduce overdispersion tests based on the Poisson regression model residuals, that do not need to estimate an unrestricted model like the negative binomial regression. As these tests coincide with *LM* overdispersion test for Poisson against negative binomial (Cameron & Trivedi, 1998) are not included in this study.

- Goodness-of-fit tests: Due to its high frequency of use (Breslow, 1996), other some more generic overdispersion tests are also used in this study. These tests assess the relationship between χ^2 or deviance and:

  - χ^2: This test evaluates the relationship between χ^2 and degrees of freedom:

$$\chi^2 = \chi^2 / \mathrm{df}$$

  - *D*: This test evaluates the relationship between deviance and degrees of freedom:

$$D = \frac{\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right)}{df}$$

Purpose

The main two aims of this study are:

a) Compare the nominal error rate (NER) of *LR*, *LMR I*, *LMR II*, χ^2 and *D* overdispersion diagnostic tests under different equidispersion situations.

b) Compare the empirical estimation of the *LR*, *LMR I*, *LMR II*, χ^2 and *D* overdispersion tests statistical power under different overdispersion levels.

Experiment 1

The aim of this simulation experiment is the comparison of the NER of the *LR*, *LMR I*, *LMR II*, χ^2 and *D* overdispersion diagnostic tests.

Method

Apparatus

Simulation was performed under R (version 1.8.1) statistical-computing environment (Ihaka & Gentleman, 1996), an independent and open-source implementation of the S (Chambers, 1998) language (Fox, 2002) which has another well-known implementation: S-Plus (Insightful Corporation, 2001). We have chosen R instead of S-Plus because it is less resource-consuming and because of its greater modularity that makes it easier and faster to incorporate new developments.

Procedure

5000 random samples of sizes:

$n = 500$, $n = 100$, $n = 50$ and $n = 20$

were extracted under Poisson distributions with λ parameters:

$\lambda = 0.3$, $\lambda = 1$ and $\lambda = 5$

The overdispersion diagnostic tests compared were:

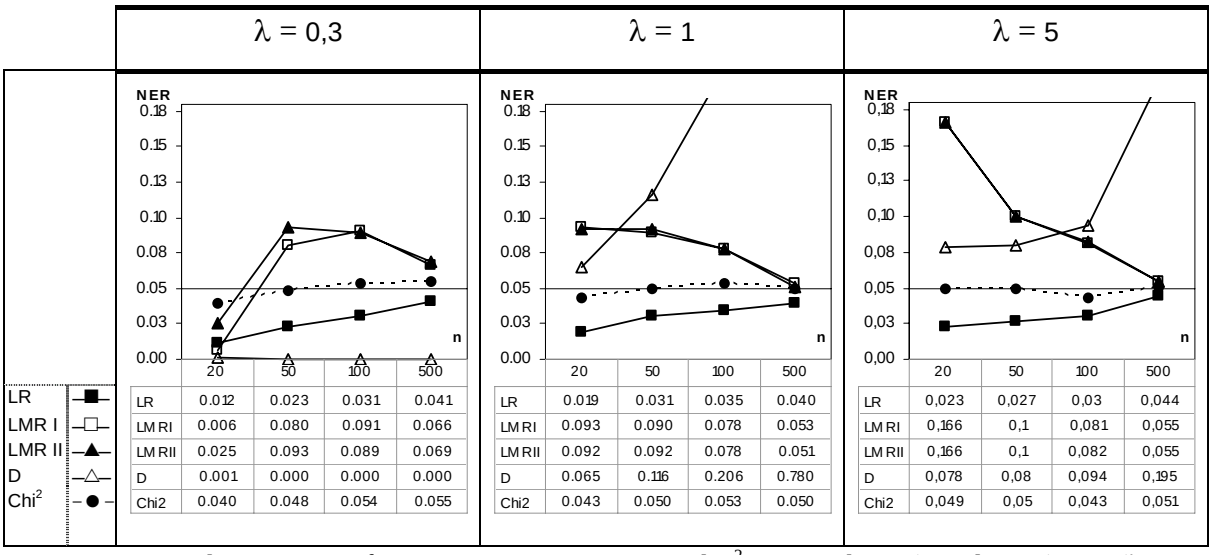
- $LR \sim \chi^2_{1,\,2\alpha}$
- $LMR\ I \sim t_{n-1,\,\alpha}$
- $LMR\ II \sim t_{n-1\,,\,\alpha}$
- $\chi^2 \sim \chi^2_{n-p\,,\,\alpha}$
- $D \sim \chi^2_{n-p,\,\alpha}$

Note that, as some authors point out (Cameron & Trivedi, 1998; Long, 1997) *LR* test is implemented as a one-tailed test due to the restriction that α cannot be negative.

We obtained the sample distribution of the above-mentioned tests under the equidispersion hypothesis so that the proportion of these tests based statistical signification decisions ($\alpha = 0.05$) is the empirical estimation of their NER.

Results

Figure 1 shows the NER of the overdispersion diagnostic tests under each experimental condition.



	λ = 0,3			
	20	50	100	500
LR	0.012	0.023	0.031	0.041
LMRI	0.006	0.080	0.091	0.066
LMRII	0.025	0.093	0.089	0.069
D	0.001	0.000	0.000	0.000
Chi2	0.040	0.048	0.054	0.055

	λ = 1			
	20	50	100	500
LR	0.019	0.031	0.035	0.040
LMRI	0.093	0.090	0.078	0.053
LMRII	0.092	0.092	0.078	0.051
D	0.065	0.116	0.206	0.780
Chi2	0.043	0.050	0.053	0.050

	λ = 5			
	20	50	100	500
LR	0,023	0,027	0,03	0,044
LMRI	0,166	0,1	0,081	0,055
LMRII	0,166	0,1	0,082	0,055
D	0,078	0,08	0,094	0,195
Chi2	0,049	0,05	0,043	0,051

Figure 1: Nominal error rates of LR, LMR I, LMR II, D and χ^2 tests under 4 (sample size) × 3 (λ parameter) conditions.

As can be seen at first glance λ does not influence NER in a systematic way. On the other side, NER of most of the tests depends on sample size in the sense that most tests appear to be consistent so that as sample size grows up so does NER. χ^2 has NER values remarkably stable and close to pre-fixed α in all experimental conditions so that χ^2 test is almost unaffected by sample size or λ. LR test performance was similar to χ^2 test, with NER values no as close to α as the latter except for n = 500.

Experiment 2

The aim of this second simulation experiment is the comparison of the statistical power empirical estimation of the same overdispersion diagnostic tests used in Experiment 1 (*LR, LMR I, LMR II,* χ^2 and *D*).

Method

Procedure

As in Experiment 1, the simulation program was implemented and executed using R version 1.8.1.

5000 random samples of sizes:

$n = 500$, $n = 100$, $n = 50$ and $n = 20$

were extracted under Poisson distributions with λ parameters:

$\lambda = 0.3$, $\lambda = 1$, $\lambda = 5$ and $\lambda = 10$

Three levels of overdispersion were simulated using the Negbin II variance function of the negative binomial distribution, with the parameter settings shown in Table 1.

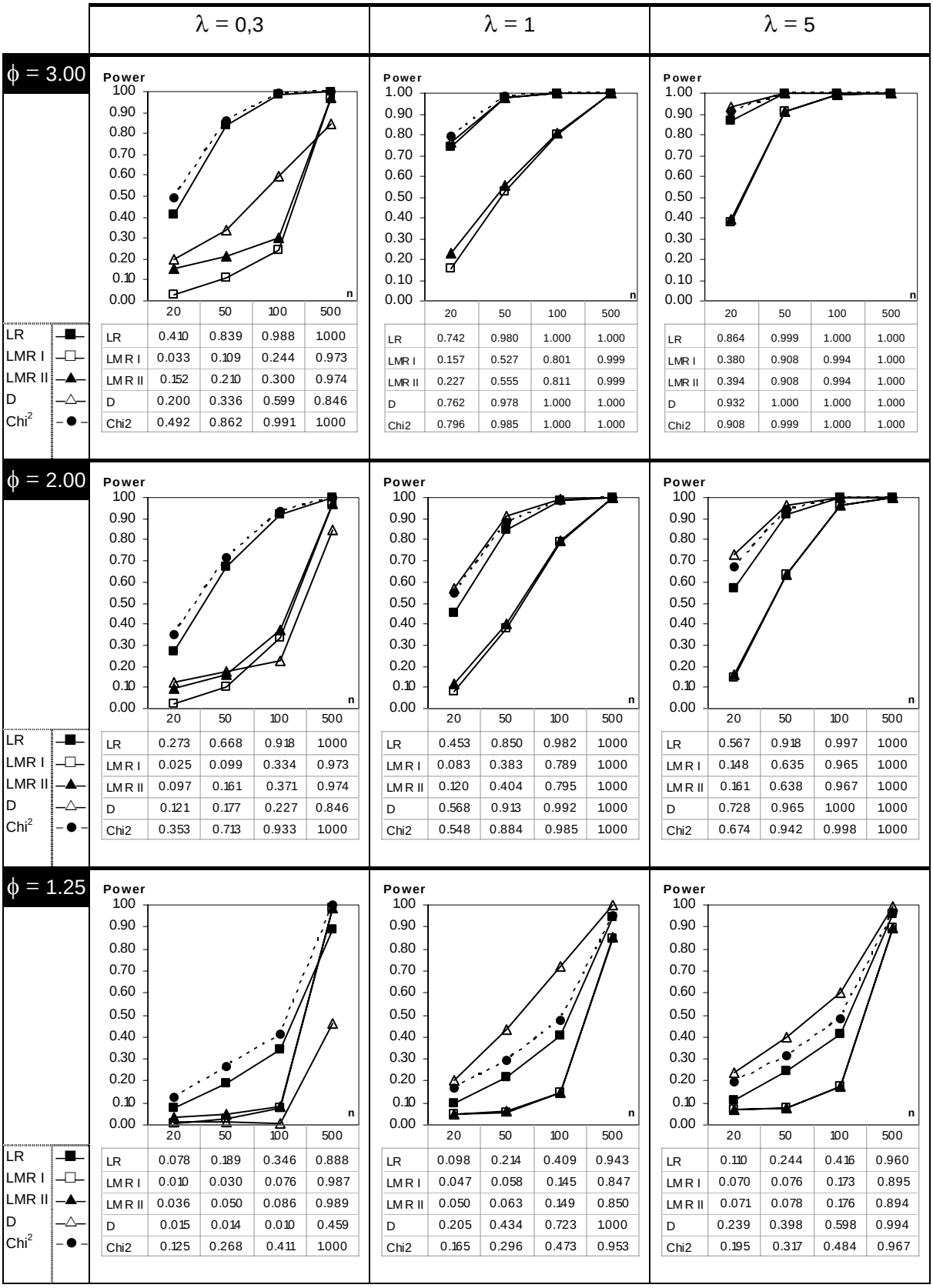
Table 1: Three scale parameter ϕ values (1.25, 2 and 3) were used for each λ to simulate different levels of overdispersion ranging from low overdispersion ($\phi = 1.25$) to high ($\phi = 3$). These ϕ values were obtained combining different τ within each λ.

λ	τ	Var = $\lambda + \lambda^2 / \tau$	Expected value of the scale parameter ϕ
0.3	1.2	$0.3 + 0.3^2 / 1.2 = 0.375$	$0.375 / 0.3 = 1.25$
	0.3	$0.3 + 0.3^2 / 0.3 = 0.6$	$0.6 / 0.3 = 2$
	0.15	$0.3 + 0.3^2 / 0.15 = 0.9$	$0.9 / 0.3 = 3$
1	4	$1 + 1^2 / 4 = 1.25$	$1.25 / 1 = 1.25$
	1	$1 + 1^2 / 1 = 2$	$2 / 1 = 2$
	0.5	$1 + 1^2 / 0.5 = 3$	$3 / 1 = 3$
5	20	$5 + 5^2 / 20 = 6.25$	$6.25 / 5 = 1.25$
	5	$5 + 5^2 / 5 = 10$	$10 / 5 = 2$
	2.5	$5 + 5^2 / 2.5 = 15$	$15 / 5 = 3$

Empirical estimations of statistical power were obtained for the same overdispersion diagnostic tests used in Experiment 1 ($LR \sim \chi^2_{1,\,2\alpha}$, $LMR\ I \sim t_{n-1,\,\alpha}$, $LMR\ II \sim t_{n-1\,,\,\alpha}$, $\chi^2 \sim \chi^2_{n-p\,,\,\alpha}$, $D \sim \chi^2_{n-p,\,\alpha}$).

Results

Figure 2: Statistical power of LR, LMR I, LMR II, D and χ^2 tests under 4 (sample size) × 3 (λ parameter) × 3 (ϕ parameter) conditions.



ϕ = 3.00

	λ = 0,3				λ = 1				λ = 5			
	20	50	100	500	20	50	100	500	20	50	100	500
LR	0.410	0.839	0.988	1.000	0.742	0.980	1.000	1.000	0.864	0.999	1.000	1.000
LMR I	0.033	0.109	0.244	0.973	0.157	0.527	0.801	0.999	0.380	0.908	0.994	1.000
LMR II	0.152	0.210	0.300	0.974	0.227	0.555	0.811	0.999	0.394	0.908	0.994	1.000
D	0.200	0.336	0.599	0.846	0.762	0.978	1.000	1.000	0.932	1.000	1.000	1.000
Chi2	0.492	0.862	0.991	1.000	0.796	0.985	1.000	1.000	0.908	0.999	1.000	1.000

ϕ = 2.00

	λ = 0,3				λ = 1				λ = 5			
	20	50	100	500	20	50	100	500	20	50	100	500
LR	0.273	0.668	0.918	1.000	0.453	0.850	0.982	1.000	0.567	0.918	0.997	1.000
LMR I	0.025	0.099	0.334	0.973	0.083	0.383	0.789	1.000	0.148	0.635	0.965	1.000
LMR II	0.097	0.161	0.371	0.974	0.120	0.404	0.795	1.000	0.161	0.638	0.967	1.000
D	0.121	0.177	0.227	0.846	0.568	0.913	0.992	1.000	0.728	0.965	1.000	1.000
Chi2	0.353	0.713	0.933	1.000	0.548	0.884	0.985	1.000	0.674	0.942	0.998	1.000

ϕ = 1.25

	λ = 0,3				λ = 1				λ = 5			
	20	50	100	500	20	50	100	500	20	50	100	500
LR	0.078	0.189	0.346	0.888	0.098	0.214	0.409	0.943	0.110	0.244	0.416	0.960
LMR I	0.010	0.030	0.076	0.987	0.047	0.058	0.145	0.847	0.070	0.076	0.173	0.895
LMR II	0.036	0.050	0.086	0.989	0.050	0.063	0.149	0.850	0.071	0.078	0.176	0.894
D	0.015	0.014	0.010	0.459	0.205	0.434	0.723	1.000	0.239	0.398	0.598	0.994
Chi2	0.125	0.268	0.411	1.000	0.165	0.296	0.473	0.953	0.195	0.317	0.484	0.967

Figure 2 shows the statistical power of the overdispersion diagnostic tests under each experimental condition.

The variable that more clearly affect the statistical power is the sample size in the sense that as sample size grows up so does the statistical power. But the effect of the sample size varies under the different overdispersion levels. In the low overdispersion level ($\phi = 1.25$) the statistical power of the overdispersion tests grows slower through the different sample sizes than it does in higher overdispersion levels ($\phi = 2$ and $\phi = 3$). Therefore all overdispersion diagnostic tests can be considered remarkably consistent.

Independently of sample size, the overdispersion level effect is also important especially when lower overdispersion level ($\phi = 1.25$) is compared with the other two levels ($\phi = 2$ and $\phi = 3$).

λ values have some influence in the statistical power of *D*, *LMR I* and *LMR II* tests while *LR* and χ^2 are not much sensible to λ.

χ^2 and *LR* provide the greater statistical power in almost all the experimental conditions.

Discussion

The decision about the appropriateness of certain statistical models should be taken after thoroughly considering the level of measurement of the variables involved along with their distributional characteristics. Count data is quite usual in Social and Health Sciences research (Gardner et al., 1995). The benchmark model to analyse count data is the Poisson regression model which usability becomes limited both due to its deterministic function character and to its distributional assumptions. One such distributional assumption is equidispersion which in practice becomes an exception more than a norm. Usually count data conditional variances exceed conditional means, i.e., overdispersion, so that inferences about can be seriously mislead. High presence of count data in applied research, count data usually overdispersed and its influence in the statistical decision process, justify the importance of overdispersion diagnostics tests as first step in the count data analysis process.

In this study we have compared the NER and statistical power of some of the most commonly used overdispersion tests.

The most remarkable result in NER simulation experiment (Experiment 1) is the performance of χ^2 test: it provided the closest stable approximation to the fixed α across all experimental conditions. Most of the rest of the overdispersion diagnostic tests showed consistent NER values in all lambda conditions. The notable exception was *D* test which, although widely used as an overdispersion diagnostic test, showed unexpected progressive departures from the expected α value as sample size grew up.

In Experiment 2 χ^2 and *LR* tests statistical power were better than the rest in almost all experimental conditions.

When researcher is faced with count data analysis, we recommend as a first step in the data analysis process, the testing of overdispersion through χ^2: a computational simple and efficient with a noteworthy performance in some relevant situations like the ones manipulated in our study.

On the other side, the also commonly used *D* test should be used carefully especially due to its proneness to NER (Type I) errors.

References

Abbring, J. H., Chiappori, P. A., Heckman, J. J., & Pinquet, J. (2003). Adverse selection and moral hazard in insurance: Can dynamic data help to distinguish? *Journal of the European Economic Association, 1,* 512-521.

Breslow, N. (1996). Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata, 8,* 23-41.

Cameron, A. C. & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46,* 347-364.

Cameron, A. C. & Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics, 1,* 29-53.

Cameron, A. C. & Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Chambers, J. M. (1998). *Programming with Data. A Guide to the S Language.* New York: Springer-Verlag.

Ezell, M. E., Land, K. C., & Cohen, L. E. (2003). Modeling Multiple Failure Time Data: A Survey of Variance-Corrected Proportional Hazards Models with Empirical Applications to Arrest Data. *Sociological Methodology, 33,* 111-167.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, CA: Sage.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*, 392-404.

Gurmu, S. & Trivedi, P. K. (1992). Overdispersion tests for truncated Poisson regression models. *Journal of Econometrics, 54*, 347-370.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*, 299-314.

Insightful Corporation (2001). *S-Plus for Windows user´s guide*. Seattle, WA: Author.

Krzanowski, W. J. (1998). *An Introduction to Statistical Modelling*. London: Arnold.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*. Oxford: Oxford University Press.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. New York: Springer-Verlag.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Lunneborg, C. E. (1994). *Modelling Experimental and Observational Data*. Belmont, CA: Duxbury Press.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. (2ª ed.) London: Chapman & Hall.

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics, 12*, 337-350.

Navarro, A., Utzet, F., Puig, P., Caminal, J., & Martín, M. (2001). La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes. *Gaceta Sanitaria, 15*, 447-452.

Rodríguez, G. (1998). *Lecture Notes on Generalized Linear Models*. Retrieved February 22, 2002, from http://data.princeton.edu/wws509/notes

Rothenberg, T. J. (1984). Hypothesis testing in linear models when the error covariance matrix is nonscalar. *Econometrica, 52*, 827-842.

Tu, W. & Zhou, X. (1999). A Wald test comparing medical costs based on log-normal distributions with zero valued costs. *Statistics in Medicine, 18*, 2749-2761.

Winkelmann, R. (2000). *Econometric Analysis of Count Data*. (3ª ed.) Berlin: Springer-Verlag.

Ajuste y estimación de los errores estándar de los parámetros del modelo de regresión de Poisson con sobredispersión.

Artículo publicado en AEMCCO 2004; supl., 329-335. Reproducción autorizada.

Metodología de las Ciencias del Comportamiento, Suplemento 2004, 329-335

Ajuste y estimación de los errores estándar de los parámetros del modelo de regresión de Poisson en presencia de sobredispersión

Noelia Llorens Aleixandre*[1], Alfonso Palmer Pol* y Josep María Losilla Vidal**
**Universidad de las Islas Baleares*
*** Universidad Autónoma de Barcelona*

Resumen

La crítica más notable al modelo de regresión de Poisson reside en la asunción del supuesto de equidispersión, si esta condición no se satisface, aparecen en general datos sobredispersos (overdispersed data). Para solucionar este problema el modelo puede ser modificado de dos formas, haciendo un ajuste post hoc de los errores estándar de los parámetros, mediante la utilización de índices directos o bien estimando nuevos errores estándar. El objetivo de este trabajo es comprobar mediante un estudio de simulación, la adecuación de los diferentes ajustes y estimaciones a realizar según las condiciones manipuladas en la muestra.

Abstract

The most remarkable criticism to the pattern of Poisson regression is the assumption of the equidispersion supposition. If this condition is not satisfied, overdispersed data is observed. To solve this problem the pattern can be modified in two ways: making a post hoc adjustment of the standard errors of the parameters (using direct indexes) or estimating new standard errors. The objective of this work is to check, by a simulation study, the fitting of the different adjustments and estimates according to the conditions of the sample.

Introducción

El modelo de regresión de Poisson (MRP) ha sido usado como modelo de referencia en el análisis de datos de recuento (Cameron & Trivedi, 1986, 1990; Gurmu,1991; Lee, 1986, Lindsey, 1998). Este modelo presenta una estructura simple y puede ser fácilmente estimado (Greene, 1993, 2000; Lee,1986), sin embargo, esta simplicidad es el resultado, como señala entre otros Sturman (1999), de algunas limitaciones en sus asunciones, el incumplimiento de las cuales tienen efectos sustanciales en la eficiencia de los coeficientes del modelo.

En la distribución de Poisson la relación entre media y variancia se caracteriza por la equidispersión, esto es:

$$Var\langle y_i \,|\, x_i \rangle = E\langle y_i \,|\, x_i \rangle$$

La equidispersión constituye un supuesto básico de diferentes modelos lineales generalizados, un incumplimiento de la asunción de la variancia es suficiente para incumplir el supuesto distribucional de Poisson (Winkelmann, 2000).

Tal como señalan McCullagh y Nelder (1989), en la práctica, la equidispersión (presencia de variancia nominal) es más la excepción que la norma. Así pues, la crítica más notable al modelo de regresión de Poisson es la presunción de que la media de y_i es igual a su variancia. Si esta condición no se satisface, aparecen en general datos sobredispersos (overdispersed data), aunque podrían ser también datos infradispersos. Ante esto Krzanowski (1998) y Winkelmann (2000) señalan que, en ausencia de equidispersión es mas frecuente una situación de sobredispersión que de infradispersión. De hecho, las pruebas para evaluar equidispersión son denominadas habitualmente pruebas de sobredispersión.

En esta situación, al ajustar el modelo de regresión de Poisson se obtiene una infraestimación de la matriz de covariancias de los parámetros de regresión produciendo una sobreestimación de los valores de la prueba de conformidad de los parámetros y por tanto de su significación (Liao, 1994).

El principal problema de la existencia de sobredispersión es que, estando bien especificado el modelo, las estimaciones de los parámetros son correctas pero no así sus errores estándar (EE). En este escenario, resulta imprescindible el poder disponer de pruebas para el diagnóstico de la sobredispersión, así como de procedimientos para poder abordarla.

[1] Dirección postal: Noelia Llorens, Facultad de Psicología, Universidad de las Islas Baleares. Ctra. Valldemosa, Km. 7.5, 07122 Palma de Mallorca . E-mail: noelia.llorens@uib.es

Soluciones a la sobredispersión

Para corregir la infraestimación de los errores estándar de los parámetros del predictor lineal producida por la sobredispersión, se pueden aplicar diferentes soluciones, entre las que cabe destacar por un lado índices directos que modifiquen los EE obtenidos en el MRP y por otro lado la estimación del EE a través de técnicas relacionadas con la simulación (Cameron y Trivedi, 1998) como son: Sandwich, Jackknife y Bootstrap.

Índices directos

-Wedderburn (1974), así como Cox and Snell (1989), sugiere utilizar un factor de inflación de la variancia definido por medio de $\sqrt{\chi^2 / gl}$. La estimación de este factor se debe obtener a partir del modelo ampliado, utilizando la prueba de bondad de ajuste de Pearson.

-Otro factor de inflación de la variancia, señalado en McCullagh y Nelder (1989), es $\sqrt{D / gl}$, la estimación de este factor se obtiene a partir de la discrepancia del modelo ampliado y de sus grados de libertad. La discrepancia se obtiene por medio de la siguiente expresión:

$$D = -2 \times \sum_{i=1}^{n} y_i \times \log\left(\frac{m_i}{y_i}\right) + \left(y_i - m_i\right)$$

Técnicas relacionadas con la simulación

Sandwich

El estimador Sandwich, que fue propuesto inicialmente por Huber (1967), Eicker (1967) y White (1980), también es conocido como estimador robusto de matrices de covariancias y tiene como objetivo la estimación consistente de la variancia. Esta aproximación da estimaciones consistentes de la matriz de covariancias tanto si trabajamos con covariancias no especificadas como si trabajamos bajo heterocedasticidad de errores. (Kauermann y Carroll, 2001). El estimador sandwich se ha utilizado considerablemente para corregir el problema de la heterocedasticidad en el Modelo Lineal General, no así en el Modelo Lineal Generalizado.

HC0 es la forma más común del sandwich, conocida también como "White estimator" y "Huber estimator". Como ha demostrado White (1980), entre otros, HC0 es un estimador consistente de la variancia en presencia de sobredispersión o heterocedasticidad, cuya expresión es:

$$\text{HC0} = (\text{X'X})^{-1}\text{X'}\,\Phi\text{X}\,(\text{X'X})^{-1} = (\text{X'X})^{-1}\text{X'diag}\,(\text{e}^{2}_{\text{i}})\text{X}\,(\text{X'X})^{-1}$$

El estimador de la variancia sandwich se obtiene mediante la expresión:

$$\text{Var}_{\text{s}}\,(\hat{\beta}) = I^{-1}\left[\sum_{i=1}^{n} w_i^2\left(\frac{y_i - \mu_1}{v(\mu_i)}\right)^2 x_i x_i^t\right] I^{-1}$$

donde I es la matriz de información de Fisher.

Posteriormente aparecieron diferentes versiones del sandwich, MacKninnon and White (1985) señalan tres estimadores alternativos diseñados para mejorar las características de HC0 en muestras pequeñas.

El ajuste más simple, sugerido por Hinkley (1977), obtiene una corrección utilizando los grados de libertad, concretamente ajusta aplicando el factor de escala N/(N-K). Esta versión es conocida como HC1:

$$\text{HC1} = \frac{N}{N-K}\,(\text{X'X})^{-1}\text{X'diag}\,(\text{e}^{2}_{\text{i}})\text{X}\,(\text{X'X})^{-1} = \frac{N}{N-K}\,\text{HC0}$$

Un segundo ajuste propuesto por Belsley, Kuh and Welsch (1980) y Wu(1986) introduce un factor de escala $1/(1-h_i)$ basándose en el análisis de valores influyentes y valores alejados. Esta versión es conocida como HC2:

$$\text{HC2} = (\text{X'X})^{-1}\text{X'diag}\left(\frac{e_i^2}{1-h_{ii}}\right)\text{X}\,(\text{X'X})^{-1}$$

Un tercer ajuste, sugerido también por Mackninnon and White (1985) aproxima el sandwich al estimador Jackknife de Efron (1982):

$$HC3 = (X'X)^{-1}X'\text{diag}\left(\frac{e_i^2}{(1-h_{ii})^2}\right)X\,(X'X)^{-1}$$

Cribari-Neto (2003) propone un cuarto ajuste que, según este autor, funcionará mejor que HC2 y HC3 cuando aparezcan valores alejados y valores influyentes,.

$$HC4 = (X'X)^{-1}X'\text{diag}\left(\frac{e_{in}^2}{(1-h_{in})^{\delta_{in}}}\right)X\,(X'X)^{-1}$$

Un quinto ajuste es el presentado por Salibian en su tesis doctoral (2000)

$$\text{HC Salibian: } ((X'X)^{-1}X'\text{diag}\,(e^{2}_{i})X\,(X'X)^{-1})\sqrt{\quad}\ [\text{disp (quasi poisson)}]$$

Jackknife

El Jackknife es una técnica ideada por Maurice Quenouille (1949,1956) y perfeccionada por John W. Tukey (1958). Este método, como señala Losilla(2002) consiste básicamente en, definida una muestra de observaciones de tamaño n, suprimir cada vez un conjunto de observaciones g y calcular sobre el conjunto (n-g) restante de datos el estadístico de interés. La aplicación más generalizada se basa en excluir cada vez una única observación, debido a que como indica Miller (1974) se evita así la arbitrariedad en la formación de subgrupos y parece haberse mostrado como la forma óptima de aplicación.

La estimación del error estándar viene dada por:

$$\hat{\sigma}_j = \sqrt{\frac{(n-1)}{n}\sum_{i=1}^{n}\left[\hat{\theta}_{-i} - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i}\right)\right]^2}$$

$$\text{Var}_{jackknife} = \frac{n-p}{n}\sum_{i=1}^{n}\left(\hat{\beta}^{*}_{(i)} - \hat{\beta}_{(.)}\right)\left(\hat{\beta}^{*}_{(i)} - \hat{\beta}_{(.)}\right)^{t}$$

$$\hat{\beta}^{*}_{(i)} = \hat{\beta} - \frac{\left(X^T\hat{V}X\right)^{-1}x_i r_i^R}{1-\hat{h}_i}$$

donde $\hat{V} = diag(V - (\hat{\mu}))$ es una matriz (n x n). X es una (n x p) matriz de covariadas, $\hat{h}_i$ es la i-ésima diagonal de la matriz hat. $\hat{\beta}$ es el vector de estimación de coeficientes usando todas las puntuaciones y $\hat{r}_i^R$ es la estimación de respuesta residual.

Bootstrap

El Bootstrap (Efron, 1979) se basa en remuestrear con reposición a partir de la muestra original para obtener nuevas muestras.

En los experimentos de simulación del presente trabajo se aplicarán las correcciones de los EE a través de los índices directos, así mismo se realizará la estimación de los EE utilizando las técnicas de Sandwich, Jackknife y Bootstrap, para verificar la adecuación real de estos procedimientos bajo diferentes promedios de recuentos, grados de sobredispersión y distintos tamaños muestrales, al tiempo que se compararan entre sí para establecer reglas de aplicación práctica.

Diseño del estudio de simulación

El estudio de simulación se ha llevado a cabo a través del paquete estadístico R, para cada muestra, la variable de respuesta Y será una variable de recuento distribuida según la ley o mecanismo de probabilidad correspondiente, y la única variable explicativa X no guardará ninguna relación con Y, de modo que es fácil conocer a priori los valores que deberán tomar los dos coeficientes b_0 y b_1 estimados mediante los diferentes modelos de regresión:

$$b_0 = \log(\lambda) \Rightarrow \exp(b_0) = \lambda$$
$$b_1 = 0 \Rightarrow \exp(b_1) = 1$$

La simulación de la sobredispersión se ha realizado a través de la distribución binomial negativa, concretamente con la función variancia correspondiente a Negbin II, con las configuraciones de los parámetros λ y τ que aparecen en la tabla 1, en la que se indica la relación entre los valores de los parámetros y de los valores esperados para el coeficiente b_0 en el modelo de regresión que se ajustará en cada muestra, así como el valor esperado del parámetro de escala que se estimará mediante quasi-Poisson:

Tabla 1. Configuración de parámetros λ y τ

λ	τ	**Var =** $\lambda + \lambda^2/\tau$	**Valor esperado del parámetro de escala** ϕ
0.3	**1.2**	$0.3+0.3^2/1.2=0.375$	$0.375/0.3=1.25$
	0.3	0.6	2
	0.15	0.9	3
1	**4**	1.25	1.25
	1	2	2
	0.5	3	3
5	**20**	6.25	1.25
	5	10	2
	2.5	15	3

Para llevar a cabo la simulación de este estudio, se procede a la extracción aleatoria de 3000 muestras, de tamaños: n= 50, n= 100, n= 250 a partir de distribuciones de Poisson con parámetros λ: $\lambda = 0.3$, $\lambda = 1$, $\lambda = 5$ cada una de ellas con un grado de dispersión bajo, moderado y alto.

Objetivos

El primer objetivo que nos propusimos fue comparar diferentes procedimientos de corrección del EE usados habitualmente en la literatura para corregir la presencia de sobredispersión. En concreto manejamos los siguientes índices:

- EE estimado por MRP y multiplicado por $\sqrt{\chi^2/gl}$
- EE estimado por MRP y multiplicado por $\sqrt{D/gl}$

El segundo objetivo del estudio de simulación pretendía estimar los EE utilizando diferentes métodos de estimación, a la vez que esto nos permitía comparar los diferentes métodos y comprobar en qué circunstancias uno de ellos puede ser más adecuado que otro. En concreto manejamos las siguientes estimaciones:

- Estimación sandwich del EE
- Estimación Bootstrap no paramétrica del EE
- Estimación Jackknife del EE

Finalmente, el tercer objetivo del estudio de simulación pretendía comparar los dos conjuntos anteriores para comprobar qué correcciones se ajustan mejor al verdadero EE. También consideramos interesante comprobar en qué circunstancias es mejor el ajuste o la estimación del EE.

Resultados

En la tabla 2 se resumen algunos de los resultados obtenidos en la simulación, en concreto los que hacen referencia al parámetro b1 del MRP utilizado. La primera columna "b1 Poisson" representa el valor teórico de referencia, es decir el valor del EE que deberíamos obtener si se tuviera en cuenta la sobredispersión de los datos. En la segunda columna aparece el valor del EE obtenido al aplicar directamente el modelo de regresión de Poisson sin hacer ningún ajuste.

En las siguientes columnas aparecen los índices o estimadores comparados, es decir se presentan los dos índices directos, dos estimadores sandwich HC0 y HC1, el Bootstrap y el Jackknife.

Vemos que aunque el tamaño de la muestra aumente, aplicar el MRP sin ningún ajuste va a provocar una infraestimación del EE, comprobándose así la necesidad de solucionar el problema de la sobredispersión. En cuanto a los diferentes métodos, señalar que:

En cuanto a los índices directos, presenta un mejor ajuste el índice $\sqrt{\chi^2/gl}$. Aunque podíamos pensar que ante un tamaño de muestra grande los dos índices directos funcionan igual, si la media es pequeña y el grado de dispersión grande, $\sqrt{D/gl}$ sigue infraestimando el EE, acercándose más al dado por Poisson que al teórico de referencia. Por ello podemos señalar que, en muestras pequeñas funciona mejor el $\sqrt{\chi^2/gl}$ independientemente de la media y del grado de dispersión; en muestras medias el EE es mejor ajustado en la mayoría de las ocasiones por el $\sqrt{\chi^2/gl}$; si fijamos N, en medias grandes y grados de dispersión pequeños y medios, funciona bien $\sqrt{D/gl}$, pero en el resto de combinaciones es más adecuado utilizar $\sqrt{\chi^2/gl}$.

De los dos sandwich presentados vemos como el HC1 presenta una mejor estimación del EE que el HC0. Las diferencias entre ellos son mínimas y se presentan principalmente en muestras pequeñas, lo que verifica que el ajuste que realiza HC1 sobre el HC0 mejora su comportamiento en muestras pequeñas.

En cuanto al Bootstrap y al Jackknife, cabe subrayar, que el funcionamiento de los dos estimadores es prácticamente el mismo, y podemos observar una aproximación casi perfecta al EE, dando estimaciones eficientes. Solo podrían señalarse pequeñas diferencias cuando se dan las condiciones más adversas, esto es, tamaño pequeño, media pequeña y grado de dispersión grande. En esta situación aunque los dos sobreestiman el EE, el Bootstrap se acerca más al valor real, dando valores más pequeños del EE.

En general el Bootstrap es el que presenta una mejor estimación del EE en casi todas las ocasiones, tanto en las condiciones adversas como en las más favorables. El HC1 funciona correctamente a partir de medias moderadas, independientemente del grado de dispersión y del tamaño de la muestra. Con tamaño de muestra grande todos los métodos comparados funcionan de forma correcta, presentando todos ellos un buen ajuste al EE real.

Conclusiones

En este trabajo se ha presentado el modelo que se considera de referencia en el análisis de datos de recuento: el modelo de regresión de Poisson. El hecho de que el modelo de regresión de Poisson reconozca la naturaleza de las variables de recuento hace de él un candidato idóneo para el análisis de este tipo de variables. Sin embargo, tal como se ha indicado, la propia restrictividad del modelo como consecuencia de las asunciones de las que parte, hacen que su aplicabilidad se vea restringida.

Ante las situaciones de sobredispersión, las estimaciones de los coeficientes en el MRP son insesgadas, aunque las estimaciones de los errores estándar sí presentan un sesgo hacia la infravaloración, bajo cualquier mecanismo generador de sobredispersión, influyendo no obstante el tamaño muestral y el grado de sobredispersión, que inciden directamente sobre el nivel de infraestimación de los EE de los parámetros del modelo de regresión.

Por último en cuanto a los procedimientos de corrección y estimación del error estándar de los coeficientes de regresión de Poisson, claramente los resultados indican la superioridad de las estimaciones no paramétricas Bootstrap y Jackknife, sobre la corrección directa del error estándar infraestimado mediante su producto por la raíz de alguna forma de estimación del parámetro de dispersión.

Bibliografía

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980), *Regresion Diagnostics: Identifying Influential Data ans Sources of Collinearity,* New York: Wiley.

Cameron, A. C. y Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests [Versión electrónica]. *Journal of Applied Econometrics, 1*, 29-53.

Cameron, A. C. y Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cameron, A. C. y Trivedi, P. K. (1998). Regression Analysis of Count Data. *Econometric Society Monographs*, *30*. Cambridge: Cambridge University Press.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Cribari-Neto, F. (2003). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis.*

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Statístics*,7,1-26.

Efron, B. (1982). *The jackknife, the bootstrap and another resampling plans.* CBMS Regional Conference Series in Applied Mathematics 38. Philadelphia: SIAM Publications.

Eicker, F. (1967). Limit Theorems for Regressions With Unequal and Dependent Errors, *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley , CA: University of California Press, pp. 59-82.

Greene, W. H. (1993) *Econometric Analysis.* New York: Macmillan.

Greene, W. H. (2000). *Econometric analysis.* (4ª ed.). New York: Prentice Hall.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business and Economic Statistics, 9*(2), 215-222.

Hinkley, D.V. (1977), Jackknifing in Unbalanced Situations. *Technometrics, 19,* 285-292.

Huber, P. J. (1967), The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions. *In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley: University of California Press, pp.221-233.

Kauermann, G. y Carroll, R.J. (2001). A note on the efficiency of Sandwich Covariance Matrix Estimation. *Journal of the American Statistical Association, 96*(456), 1387-

Krzanowski, W. J. (1998). *An Introduction to Statistical Modelling.* London: Arnold

Lee, L.F. (1986). Specification test for Poisson regression models. *International Economic Review, 27*, 689-706.

Liao, T.F. (1994). *Interpreting Probability Models. Logit, Probit and other Generalized Linear Models.* London: Sage

Lindsey, J. K. (1998). Counts and times to events. *Statistics in Medicine, 17*(15-16), 1745-1751.

Losilla, J.M. (2002). Computación intensiva para el Análisis de Datos en el siglo XXI. *Metodología de las ciencias del comportamiento*, 4(2), 201-221.

MacKninnon, J. G. and White, H. (1985). Some Heteroskedasticity-consistent covariance matrix estimators with improvement finite samples properties. *Journal of Econom.*29, 305-325.

McCullagh, P. y Nelder, J. A. (1989). *Generalized linear models.* (2ª ed.). London: Chapman & Hall.

Miller, R.G. (1974). The Jackknife – a review. *Biometrika,* 61(1), 1-15.

Quenouille,M. H. (1949). Aproximate test of correlation in time series. *Journal of the Royal Statistical Society – Series B*, 11, 68-84.

Quenouille,M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

Tukey, J.W. (1958). Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics, 29,* 614.

Salibian, M. (2000). Contributions to the theory of robust inference. University of British Columbia. Tesis doctoral no publicada.

Sturman, M. C. (1999). Multiple approaches to analyzing count data in studies of individual differences: The propensity for Type I errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement, 59*(3), 414-430.

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* (3ª ed.). Berlin: Springer-Verlag. ediction. *Educational and Psychological Measurement, 59*(3), 414-430.

White, H. (1980). A Heteroskedastic-Consistent Covariance Matrix Estimator and Direct Test of Heteroskedasticity", *Econometrica,* 48, 817-838.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss-Newton methods. *Biometrika*,61(3), 439-447.

Wu, C.F. J. (1986). Jackknife, bootstrap and another resampling methods in regression analysis. *Ann. Statistics,* 14, 1261-1295.

Tabla 2. Estimaciones realizadas por los diferentes métodos en las diferentes condiciones

						b1 Poisson	EE b1 Poisson	EE b1 Discrep/gl	EE b1 X2/gl	EE hc0 b1	EE hc1 b1	Bootstrap b1	Jackknife b1
						Desviación típ.	Media	Media	Media	Media	Media	Media	Media
Tamaño	50,0	Media de Y	,3	Grado Disp.	,15	,264841	,168549	,173112	,244782	,210721	,219501	,253227	,295907
					,30	,221093	,160336	,159964	,207936	,187460	,195271	,222654	,240653
					1,20	,182621	,156390	,145266	,169363	,158979	,165603	,183941	,187337
			1	Grado Disp.	,50	,140883	,083181	,122070	,136176	,126684	,131962	,137688	,144689
					1,00	,115363	,082070	,108534	,112623	,106852	,111304	,114082	,118301
					4,00	,094544	,081485	,093468	,089665	,086104	,089692	,091263	,092798
			5	Grado Disp.	2,50	,060116	,036128	,061381	,061556	,058940	,061396	,060533	,062908
					5,00	,050717	,035964	,050845	,050191	,048281	,050293	,049848	,051157
					20	,041033	,035928	,040885	,040046	,038671	,040282	,039574	,040777
	100	Media de Y	,3	Grado Disp.	,15	,191501	,111494	,120825	,176677	,163319	,166652	,185253	,192229
					,30	,159140	,109189	,111830	,147116	,140034	,142892	,152345	,157384
					1,20	,122496	,107488	,101759	,118271	,114363	,116697	,122029	,123036
			1	Grado Disp.	,50	,100526	,058351	,086548	,098268	,094848	,096784	,097281	,100994
					1,00	,081473	,057918	,076965	,080649	,078407	,080007	,080252	,082237
					4,00	,066699	,057561	,066198	,063840	,062527	,063803	,063689	,065416
			5	Grado Disp.	2,50	,043913	,025616	,043803	,044058	,043054	,043932	,043775	,044714
					5,00	,036567	,025615	,036407	,036055	,035343	,036064	,035962	,036283
					20	,028941	,025599	,029158	,028581	,028074	,028647	,028347	,028737
	250	Media de Y	,3	Grado Disp.	,15	,119217	,068383	,075660	,113492	,109029	,109908	,113841	,117022
					,30	,096432	,067369	,070255	,093750	,091584	,092322	,094201	,095889
					1,20	,075512	,067060	,064045	,074515	,073565	,074159	,075126	,075512
			1	Grado Disp.	,50	,062568	,036735	,054728	,062634	,061627	,062124	,062496	,063502
					1,00	,053930	,036512	,048708	,051309	,050762	,051171	,052525	,051779
					4,00	,040019	,036369	,041957	,040578	,040329	,040654	,039821	,040939
			5	Grado Disp.	2,50	,028073	,016285	,027837	,028061	,027838	,028062	,028381	,028267
					5,00	,022973	,016273	,023156	,022975	,022839	,023023	,022793	,023018
					20	,018105	,016280	,018519	,018153	,018042	,018188	,018103	,018213

Overdispersion in the Poisson regression model: A comparative simulation study.

Overdispersion in the Poisson regression model: A comparative simulation study

Llorens, N., Palmer, A., Losilla, J.M. & Vives, J.

Abstract

This simulation study compares different strategies to solve the problem of underestimating Standard Errors in the Poisson regression model when overdispersion is present. The study analyses the possible importance of sample size, Poisson distribution mean and dispersion parameter when choosing the best index. The results show that the indices obtained by resampling (non-parametric bootstrap and jackknife) are the least biased, followed by the direct index based on the chi-square and the so-called robust indices, in third place. Nevertheless, the resampling indices' inefficiency is also evident, especially in small samples.

Keywords: Overdispersion, non-parametric bootstrap, jackknife, robust, direct indices.

1. Introduction

The Poisson regression model (PRM) is being used as a reference in analysing count data (Cameron & Trivedi, 1986, 1990; Gurmu, 1991; Lee, 1986, Lindsey, 1998). Although this model has a simple structure and can be easily estimated (Greene, 1993, 2000; Lee, 1986), Sturman (1999), among others, indicates that this simplicity is the result of several limitations in its assumptions, which have substantial effects on the efficiency of the model's coefficients when violated.

The relationship between mean and variance in the Poisson distribution is characterised by equidispersion, i.e.:

$$Var\ (y_i \mid x_i) = E(y_i \mid x_i) \quad (1)$$

If this condition is not satisfied, then the data may overdisperse or underdisperse. Equidispersion is a basic assumption in different Generalised Linear Models and the violation of the assumption of variance is sufficient to violate the Poisson's distributional assumption (Winkelmann 2000). In practice, as McCullagh and Nelder (1989) indicated, equidispersion is the exception rather than the rule and Krzanowski (1998) and Winkelmann

(2000), among others, further indicated that overdispersion is more common than underdispersion in the absence of equidispersion.

An indication of the magnitude of overdispersion can be obtained simply by comparing the sample mean and variance of the dependent count variable.

In a Generalised Linear Model, interest focuses on both the estimate of the parameters of the linear predictor and its standard errors (SE). As Cox and Snell (1989) indicate, standard errors do not adjust to overdispersion when it is present and are lower than what they should be, thereby producing faulty inferences. When the Poisson regression model is adjusted in this case, it underestimates the regression parameter's covariances matrix, producing an overestimate of the parameters' conformity test values and their significance (Liao, 1994), as well as narrower confidence intervals.

If the exact mechanism that produces the over or underdispersion is known, specific mechanisms can be applied to model the data (see e.g. McCullagh and Nelder (1989)); however, correcting the SE is appropriate in the absence of this knowledge, as Breslow (1996), Heinzl and Mittlböck (2003) suggest.

Gardner, Mulvey and Shaw (1995) indicate two alternatives which can be used when overdispersion is present. In the first place, researchers can correct the Poisson regression model's statistical inferences by estimating a dispersion parameter and using it to correct them. On the other hand, but along the same lines, it may be preferable to obtain the model's SE estimates through different methods, rather than correcting them. There are various ways to obtain this estimate, such as the robust estimator or resampling. This first alternative may be the method of choice when the researcher is primarily interested in testing hypotheses on the Poisson regression model's coefficients. Nevertheless, overdispersed Poisson models do not specify the probability distribution of the data, which is why changing distribution may be recommended in this case.

The most popular alternative to the Poisson model with overdispersion is the Negative Binomial Regression Model (NBRM), which considers unobserved heterogeneity the source of overdispersion. As Lindsey (1999) indicates, the NBRM may be the best model for estimating the probability distribution of a specific count. Nonetheless, prudence is called for

when using models because the paper of Guo & Li (2002) shows that the measurement errors in covariates in general lead to the overdispersion on the observed data. Monte Carlo simulations show that the estimates for the parameters in the mean function of a Poisson regression could be severely biased if the overdispersion caused by measurement errors is falsely modeled as arising from the unobserved heterogeneity.

The objective of this article is to verify the conditions appropriate for the different options currently being used to correct and estimate standard errors. Section 2 describes the different corrections and ways to estimate standard errors in the presence of overdispersion; Section 3 describes the simulation study; and Section 4 presents and discusses the results from the simulation study. The article concludes with a critical discussion and recommendations.

2. Handling SE when overdispersion is present

This section presents several methods that may be used to handle SE when overdispersion is present.

Different solutions may be applied to correct the underestimation of standard errors produced by overdispersed linear predictor parameters. Two of the most widely used are on one hand, direct indices, which modify the SE obtained by the PRM, and on the other, SE estimates through resampling techniques, such as the jackknife and non-parametric bootstrap procedures or through the so-called robust or sandwich estimator (Cameron and Trivedi, 1998).

2.1.- Scaling Standard Errors

If the precise mechanism that produces the overdispersion is known, specific methods may be applied to model the data (McCullagh and Nelder, 1989). In the absence of such knowledge, it is convenient to assume approximatively that $Var(Y)=\Phi\mu$ for some $\Phi>1$ constant. This is a rather robust approach to tackle the problem, since even quite substantial deviations in the assumed simple linear functional form, $Var(Y)=\Phi\mu$, generally have a merely minor effect on the conclusions related to standard errors, confidence intervals an p-values

(McCullagh and Nelder, 1989). Now, fitting this model is a three-step procedure. Firstly, regression coefficients and their standard errors are estimated by maximizing the likelihood of a standard Poisson regression model. Secondly, the dispersion parameter Φ is estimated separately. And thirdly, the standard errors are adjusted for the estimated dispersion parameter so that proper confidence intervals and test statistics can be obtained. Note that the resulting likelihood-ratio tests are based on quasi-likelihood theory (McCullagh and Nelder, 1989).

Wedderburn (1974), like Cox and Snell (1989), suggests several procedures to model the absence of equidispersion and obtain improved estimates of the variance-covariance matrix. This factor is estimated from the global model, using Pearson's goodness of fit test and its degrees of freedom.

Then, a standard error inflation factor is calculated simply as the square root of the coefficient between Pearson's chi-square and its degrees of freedom, $\sqrt{\chi^2 / df}$. Scaled standard errors are obtained by multiplying the original standard error by this factor.

Another standard error inflation factor, indicated by McCullagh and Nelder (1989), is based on the deviance between the global model and its degrees of freedom, expressed as $\sqrt{D / df}$. This factor is estimated by calculating the global model deviance expressed in the Poisson regression model as:

$$D = -2 \times \sum_{i=1}^{n} y_i \times \log\left(\frac{m_i}{y_i}\right) + \left(y_i - m_i\right) \qquad (2)$$

where D represents Deviance, y_i represents the observed count and m_i is the expected count.

One only looks for overdispersion if the deviance is at least twice the number of degrees of freedom. Such a criterion has a theoretical justification: it is equivalent to comparing the fitted model with the saturated model by using AIC (Lindsey, 1999).

2.2. Estimating Standard Errors

There are two possible ways to estimate SE in model parameters: obtaining the so-called robust estimate and using resampling techniques, such as the jackknife or bootstrap procedures.

2.2.1. Robust Estimator: Sandwich

The sandwich estimator, initially proposed by Huber (1967), Eicker (1967) and White (1980), is also known as the robust covariance matrix estimator and is designed to consistently estimate the variance.

This approach yields consistent estimates of the covariance matrix even under misspecified working covariances as well as under heteroskedastic errors (Kauermann y Carroll, 2001).

The sandwich estimator has been widely used to correct heteroskedasticity in the General Linear Model, but not in the Generalised Linear Model.

HC0, also known as the *White estimator* and *Huber estimator*, is the most common form of the sandwich estimator and is expressed as:

$$HC0 = \left(X'X\right)^{-1} X'\Phi X \left(X'X\right)^{-1} = \left(X'X\right)^{-1} X' diag\left(e_i^2\right) X \left(X'X\right)^{-1} \quad (3)$$

where $\hat{\Phi}$ = diag { e_1^2,.....e_n^2}. That is, $\hat{\Phi}$ is a diagonal matrix formed out the vector of squared least-squares residuals. This estimator is consistent under both homoskedasticity and heteroskedasticity of unknown form; see White (1980). However, it can be considerably biased in finite samples; see, e.g., Cribari-Neto (2004); Cribari-Neto and Zarkos (1999, 2001) and MacKinnon and White (1985).

Different versions of the sandwich estimator were subsequently proposed; MacKinnon and White (1985) indicate three alternative estimators designed to improve the features of the HC0 method in small samples.

The most simple adjustment, suggested by Hinkley (1977), achieves correction by using the degrees of freedom; it specifically adjusts the values by applying scale factor *n/(n-p)*. This version is known as HC1.

$$HC1 = \left(n/\left(n-p\right)\right)\left(X'X\right)^{-1}X'diag(e_i^2)X\left(X'X\right)^{-1} = \left(n/\left(n-p\right)\right)HC0 \quad (4)$$

where *n* is the observations on a dependent variable and *p* is the number of predictors.

A second adjustment, proposed by Belsley, Kuh and Welsch (1980) and Wu (1986), introduces a scale factor of $1/(1-h_i)$ based on the analysis of influential and extreme values. This version is known as HC2.

$$HC2 = \left(X'X\right)^{-1}X'diag\left(\frac{e_i^2}{1-h_i}\right)X\left(X'X\right)^{-1} \quad (5)$$

where h_i is the ith diagonal element of the "hat matrix" $H=\boldsymbol{X(X'X)^{-1}X'}$, $i=1,\ldots\ldots n$.

An alternative estimator with superior finite-sample behavior also suggested by Mackinnon and White (1985) approaches the sandwich to the Jackknife estimator of Efron (1982). It can be devised by modifying the HC0 estimator. The idea is to use $\widehat{\Phi}=diag\{e_1^2/(1-h_1)^2,\ldots\ldots\ldots,e_n^2/(1-h_n)^2\}$:

$$HC3 = \left(X'X\right)^{-1}X'diag\left(\frac{e_i^2}{\left(1-h_i\right)^2}\right)X\left(X'X\right)^{-1} \quad (6)$$

Cribari-Neto (2004) proposes a fourth adjustment which, according to the author, functions better than HC2 and HC3, when extreme and influential values are present. The idea is to use $\widehat{\Phi}=diag\{e_1^2/(1-h_1)^{\delta_1},\ldots\ldots\ldots,e_n^2/(1-h_n)^{\delta_n}\}$, where $\delta_i = \min\left\{4,\frac{nh_i}{p}\right\}$, with n observations and p parameters:

$$HC4 = \left(X'X\right)^{-1}X'diag\left(\frac{e_i^2}{\left(1-h_i\right)^{\delta_i}}\right)X\left(X'X\right)^{-1} \quad (7)$$

2.2.2. Estimating by resampling

a. Jackknife: Jackknifing is a technique conceived by Maurice Quenouille (1949,1956) and perfected by John W. Tukey (1958). Basically, it consists of defining an *n*-size sample of observations, suppressing a set of observations (*g*) each time and calculating the statistics of interest from the remaining set (*n-g*) of data. The most generalised and apparently optimum form of application is based on excluding one sole observation each time, since this prevents arbitrarily formed subgroups, as Miller (1974) indicates. Once the *n* estimates of the statistic for each one of the jackknife samples have been obtained, an estimate of the bias associated with the original sample can be calculated, as can a non-parametric estimate of the statistic's standard error.

The jackknife estimate of standard error of an estimator $\hat{\theta} = s(X)$ is defined by:

$$SE_{jack} = \left[\frac{n-1}{n} \sum (\hat{\theta}_i - \hat{\theta}_\cdot)^2 \right]^{1/2} \tag{8}$$

Where *n* is the size samples, $\hat{\theta}_\cdot = \sum_{i=1}^{n} \hat{\theta}_i / n$; $\hat{\theta}_i = s(X_i)$ where $X_i=(x_1, x_2, \ldots\ldots x_n)$ except x_i, for $i= 1,2,3\ldots n$, called jackknife samples..

b. Non-parametric Bootstrap: The bootstrap procedure was originally proposed by Efron in 1979 as an intensive computational method which allows measures of accuracy for statistical estimates to be obtained.

The application of the non-parametric bootstrap estimate algorithm obtains a high number of repeated random samples based on the original data sample and the size of *n* itself. The non-parametric bootstrap estimate of standard error of an estimator $\hat{\theta} = s(x)$ is defined by:

$$SE_{boot} = \left[\sum_{b=1}^{B} \left[\hat{\theta}_b - \hat{\theta}_\cdot \right]^2 / (B-1) \right]^{1/2} \tag{9}$$

where *B* is the number of independent samples, $\hat{\theta}_\cdot = \sum_{b=1}^{B} \hat{\theta}_b / B$; $\hat{\theta}_b = s(X_b)$, for *b=1,2,3,….B,* called bootstrap samples.

In this paper, the SE were corrected through direct indices in the simulation experiments and estimated using sandwich, jackknife and bootstrap techniques to verify the real adaptation of these procedures under different count averages, degrees of overdispersion and sample sizes and they were compared to each other to establish practical rules of application.

3. The simulation study design

This section presents the simulation study, which compares the different procedures for post hoc adjustment and SE estimation habitually used in literature to correct overdispersion.

The following estimators and post-hoc adjustments were considered: the HC0, HC1, HC2, HC3, HC4, non-parametric bootstrap, jackknife, $\sqrt{\chi^2/df}$ and $\sqrt{D/df}$.

The simulation study was conducted through statistical package R (v.1.8.1). In each sample, the Y_i response variable is a count variable distributed according to the corresponding law or mechanism of probability and the only explanatory variable X_i does not have any relation with Y_i, so is it easy to ascertain a priori the values for the two coefficients b_0 and b_1 estimated by different regression models:

$$b_0 = \log(\lambda) \Rightarrow \exp(b_0) = \lambda$$
$$b_1 = 0 \Rightarrow \exp(b_1) = 1$$

Where b_0 is the constant of the regression model and b_1 is the parameter of the variable X_i and λ is the Poisson mean.

Overdispersion is simulated through negative binomial distribution, specifically with the variance function corresponding to Negbin II: $var\left(y_i|x_i\right) = \mu_i + \alpha\mu_i^2$, and the configurations of parameters λ and ζ (dispersion parameters) stated in Table 1.

Likewise, table 1 indicates the relationship between parameter values and the expected values of the b_0 coefficient in the regression model that is adjusted in each sample, as well as the expected value of scale parameter Φ which is estimated through quasi-Poisson.

Table 1: Parameter relationship

λ	τ	**Var =** $\lambda + \lambda^2 / \tau$	**Expected value of scale parameter** ϕ
0.3	**1.2**	$0.3+0.3^2/1.2=0.375$	0.375/0.3=1.25
	0.3	0.6	2
	0.15	0.9	3
1	**4**	1.25	1.25
	1	2	2
	0.5	3	3
5	**20**	6.25	1.25
	5	10	2
	2.5	15	3

To conduct this simulation study, 3000 random samples of sizes *n*=50, *n*=100, *n*=250 were extracted based on Poisson distributions with parameters of $\lambda = 0.3,\ \lambda = 1,\ \lambda = 5$, where 0.3 showed a Poisson distribution with a much lower frequency, 1 showed a Poisson distribution with low frequency but with at least the occurrence of one incident and 5 showed a Poisson distribution with high frequency, approaching normal distribution.

4. Results

Since the average of the sample distribution of a statistic coincides with the parameter of the sample's origin population, it can be assumed that standard deviation from the statistic's sample distribution is the real SE. Thus, this SE was used as a reference value with which to compare the different SE obtained. Box-plots were used to illustrate the adjustments of each one of the tests compared.

Because of the residuals' asymmetry, which occurred mainly in small samples and under the most adverse conditions, Huber's M-Estimator was calculated instead of the mean, since it is more robust and takes outliers and asymmetry into consideration.

Table 2: Huber's M-estimator (k=1.339)

N	λ	ϕ	Sd. real	D/df	χ^2/df	HC0	HC1	HC2	HC3	HC4	Boot	Jack
50	0.3	3	0.3080	0.1707	0.2380	0.2017	0.2102	0.2074	0.2134	0.2114	0.2994	0.2669
		2	0.2323	0.1582	0.2020	0.1795	0.1870	0.1846	0.1898	0.1878	0.2292	0.2189
		1.25	0.1824	0.1442	0.1654	0.1530	0.1593	0.1573	0.1617	0.1598	0.1774	0.1739
	1	3	0.1378	0.1214	0.1335	0.1235	0.1286	0.1270	0.1306	0.1290	0.1329	0.1374
		2	0.1167	0.1083	0.1114	0.1041	0.1084	0.1070	0.1101	0.1087	0.1112	0.1143
		1.25	0.0937	0.0933	0.0888	0.0842	0.0877	0.0866	0.0891	0.0879	0.0897	0.0910
	5	3	0.0604	0.0614	0.0612	0.0582	0.0606	0.0598	0.0616	0.0607	0.0594	0.0612
		2	0.0507	0.0507	0.0500	0.0477	0.0497	0.0491	0.0505	0.0498	0.0494	0.0508
		1.25	0.0402	0.0407	0.0399	0.0383	0.0399	0.0395	0.0406	0.0400	0.0392	0.0402
100	0.3	3	0.1912	0.1199	0.1731	0.1581	0.1613	0.1601	0.1625	0.1616	0.1778	0.1828
		2	0.1586	0.1112	0.1446	0.1356	0.1384	0.1375	0.1393	0.1384	0.1475	0.1503
		1.25	0.1224	0.1014	0.1167	0.1120	0.1143	0.1136	0.1152	0.1144	0.1184	0.1192
	1	3	0.1013	0.0864	0.0972	0.0926	0.0945	0.0939	0.0953	0.0946	0.0956	0.0985
		2	0.0821	0.0768	0.0800	0.0772	0.0787	0.0782	0.0793	0.0788	0.0793	0.0810
		1.25	0.0641	0.0661	0.0636	0.0619	0.0632	0.0628	0.0637	0.0633	0.0635	0.0644
	5	3	0.0449	0.0438	0.0440	0.0427	0.0436	0.0434	0.0440	0.0437	0.0435	0.0442
		2	0.0360	0.0364	0.0360	0.0351	0.0358	0.0356	0.0361	0.0359	0.0355	0.0361
		1.25	0.0291	0.0291	0.0285	0.0279	0.0285	0.0283	0.0287	0.0285	0.0282	0.0285
250	0.3	3	0.1159	0.0755	0.1122	0.1077	0.1085	0.1082	0.1088	0.1086	0.1108	0.1141
		2	0.0949	0.0702	0.0930	0.0906	0.0913	0.0911	0.0916	0.0914	0.0922	0.0939
		1.25	0.0765	0.0640	0.0741	0.0729	0.0735	0.0733	0.0737	0.0735	0.0744	0.0751
	1	3	0.0628	0.0547	0.0623	0.0609	0.0614	0.0613	0.0616	0.0614	0.0614	0.0623
		2	0.0523	0.0487	0.0511	0.0503	0.0507	0.0505	0.0508	0.0507	0.0507	0.0514
		1.25	0.0409	0.0420	0.0405	0.0401	0.0404	0.0403	0.0406	0.0405	0.0404	0.0408
	5	3	0.0283	0.0278	0.0280	0.0277	0.0279	0.0279	0.0280	0.0279	0.0279	0.0281
		2	0.0232	0.0232	0.0230	0.0228	0.0229	0.0229	0.0230	0.0229	0.0229	0.0231
		1.25	0.0184	0.0185	0.0182	0.0180	0.0181	0.0181	0.0182	0.0181	0.0181	0.0182

Intervals of 95% were obtained based on the 2.5 and 97.5 percentiles to see the adjustment of each one of the SE estimates and the root mean square of error (RMSE) was also calculated, using the real SE value as a reference, instead of the predicted value.

Figures 1, 2 and 3 show box-plots illustrating the different indices and adjustments which were compared, grouped according to the sample size simulated.

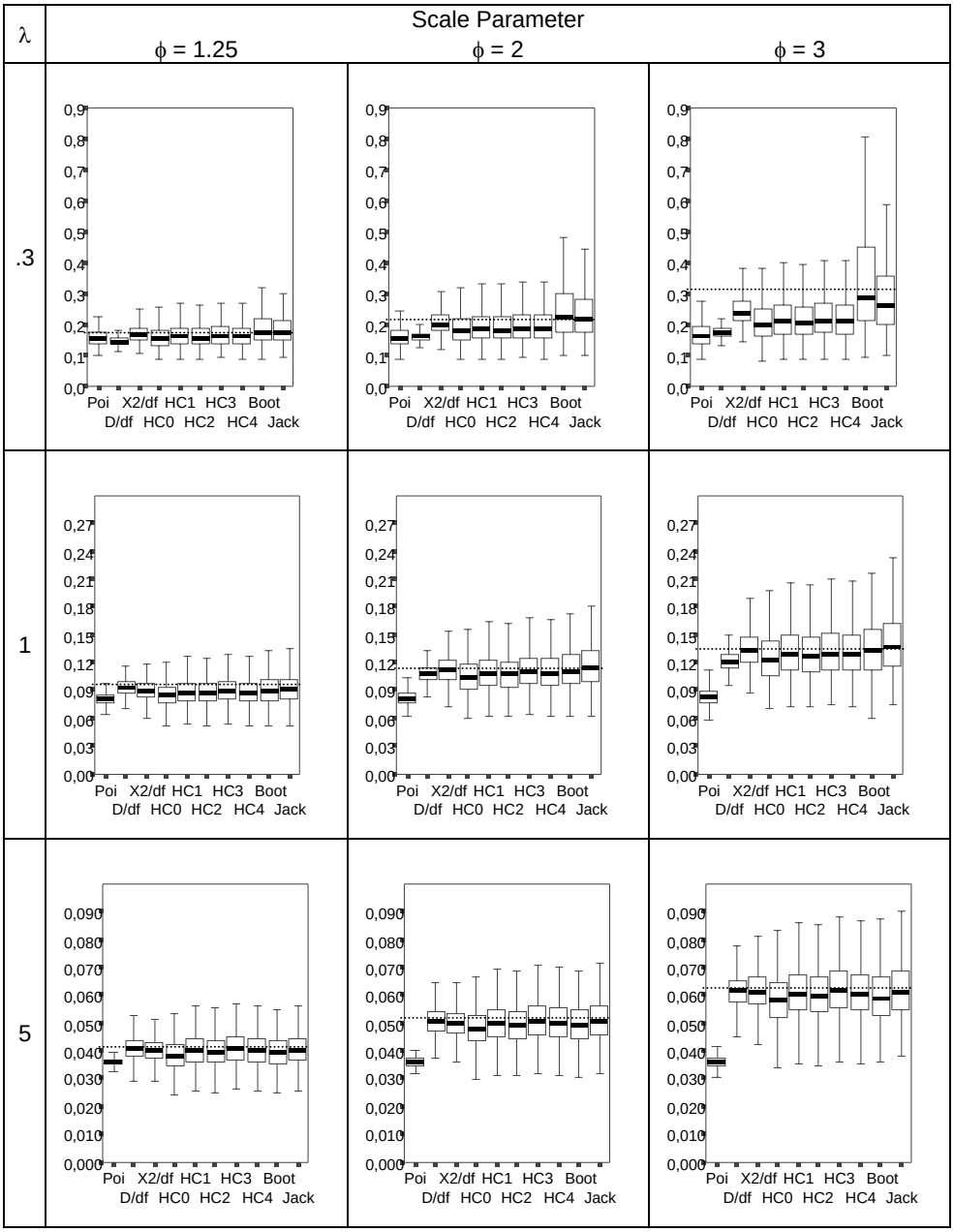


Fig. 1. Boxplots of the distributions of 3000 simulated samples for standard errors of Poisson regression parameter for n = 50, scale parameter ϕ = 1.25, 2 and 3 (in columns), and mean Poisson distribution λ = 0.3, 1 and 5 (in rows). Horizontal dot lines are drawn for the Monte Carlo standard error estimates (reference value).

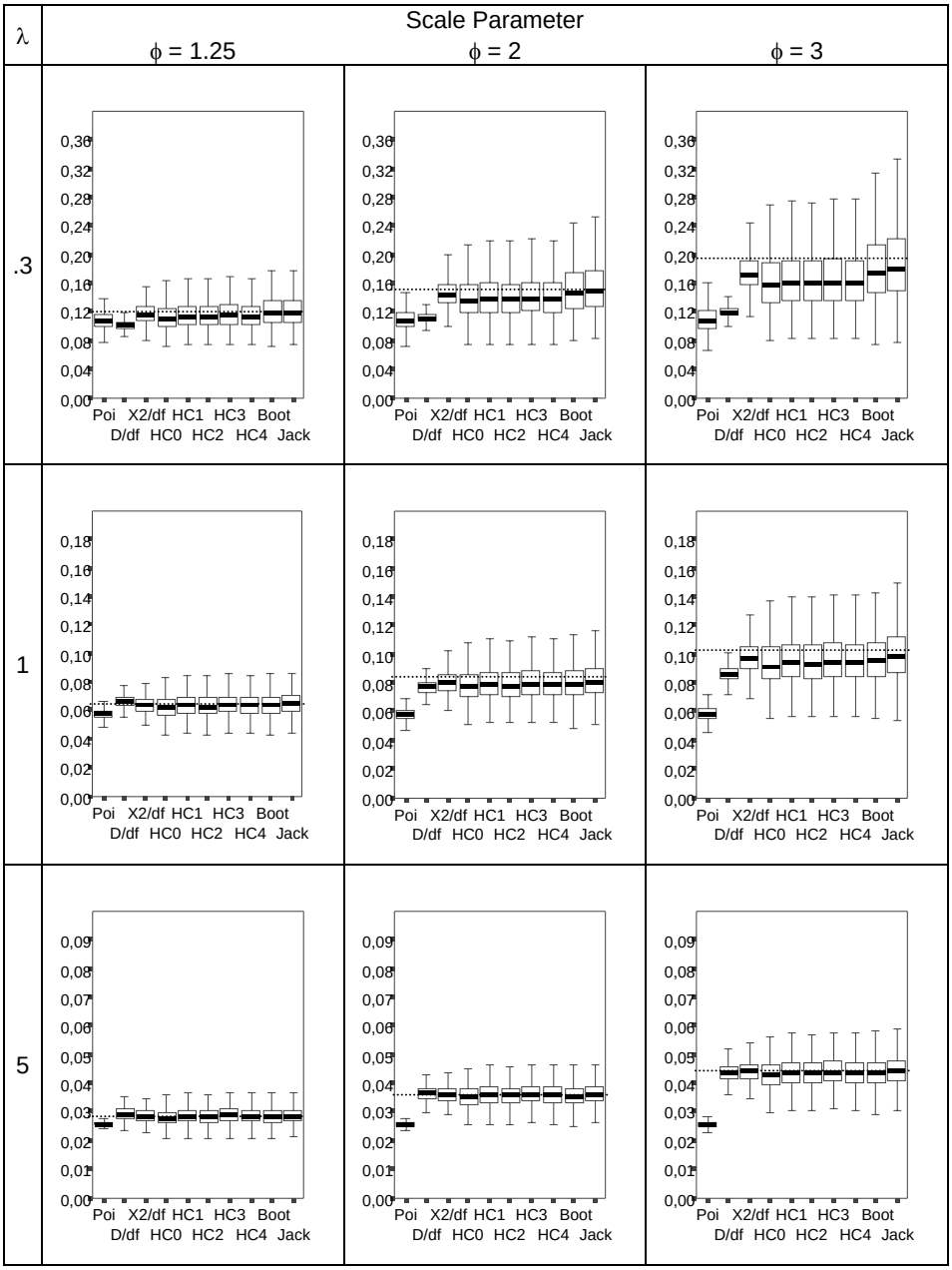


Fig. 2. Boxplots of the distributions of 3000 simulated samples for standard errors of Poisson regression parameter for n = 100, scale parameter ϕ = 1.25, 2 and 3 (in columns), and mean Poisson distribution λ = 0.3, 1 and 5 (in rows). Horizontal dot lines are drawn for the Monte Carlo standard error estimates (reference value).

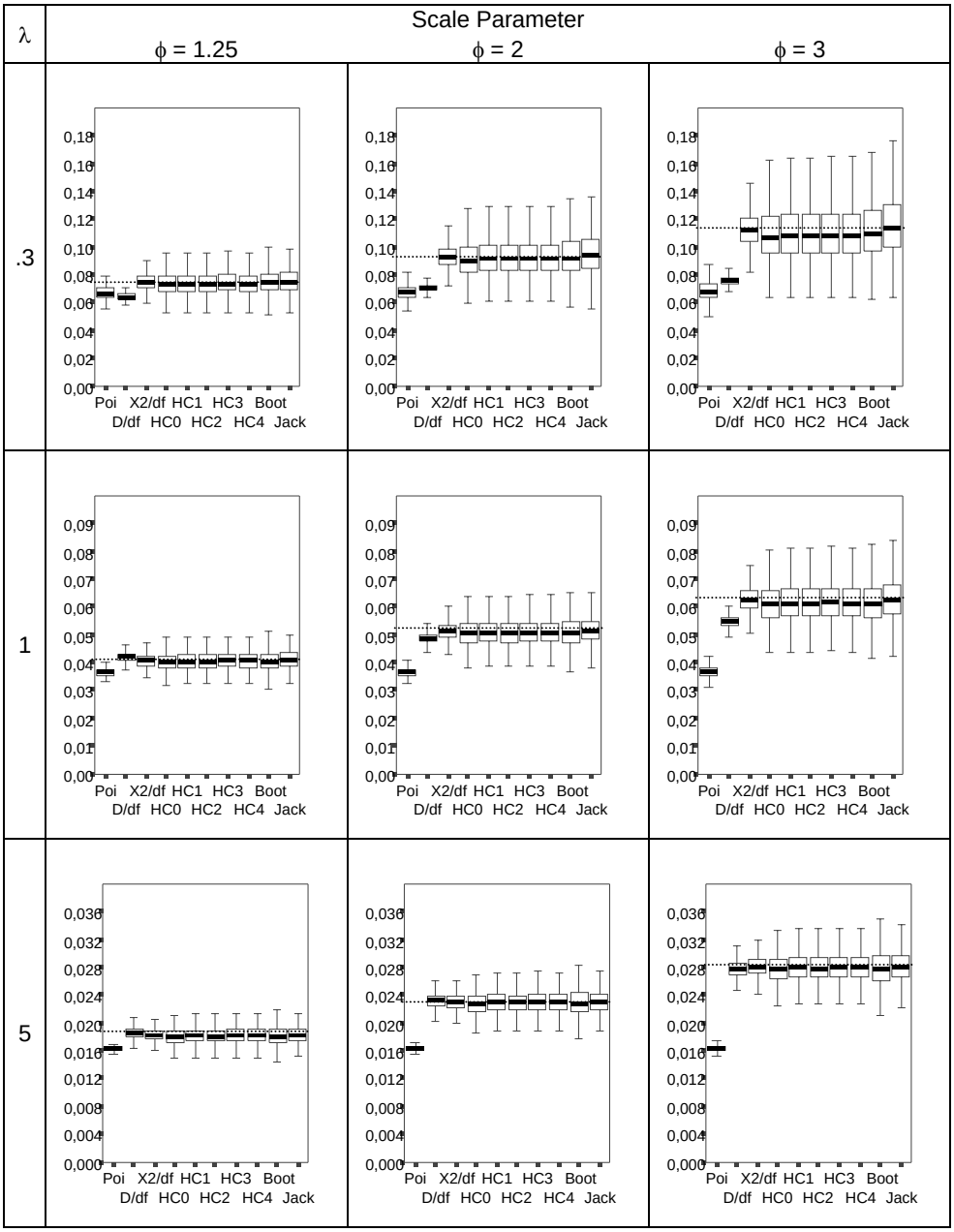


Fig. 3. Boxplots of the distributions of 3000 simulated samples for standard errors of Poisson regression parameter for n = 250, scale parameter ϕ = 1.25, 2 and 3 (in columns), and mean Poisson distribution λ = 0.3, 1 and 5 (in rows). Horizontal dot lines are drawn for the Monte Carlo standard error estimates (reference value).

In the first place, the inefficiency of the SE given by the Poisson model without correcting overdispersion was verified. The figure shows how the Poisson box-plot, given the size of a sample and a value of lambda, is fixed in one position regardless of the dispersion parameter, as it is independent from the model. Likewise, it can be seen that the more overdispersion there is, the farther away the SE obtained by the Poisson model are from the real values, i.e., they are more biased. In comparison, all the other indices vary their values according to the degree of dispersion.

Although the direct index $\sqrt{D/df}$ is generally the most precise, it does not estimate SE very accurately, as in most situations it underestimates the real value, especially and most clearly in situations where λ = 0.3 and in the most adverse situations where λ =1. The direct index is more precise than the other indices analysed and similar to them in accuracy in the remaining cases.

The direct index $\sqrt{\chi^2/df}$ performs very similarly to sandwich estimators, although it is more precise than they are in all simulation situations. This index' Huber M-estimator shows values closer to the real SE than those obtained by sandwich estimators, especially when *n*=50 and $\lambda = 0.3$. Graphs do not reflect the different behaviours of the sandwich estimators presented, nevertheless a numerical analysis shows that the HC3 and HC4 estimate SE better than HC0, HC1 and HC2. HC0's poor performance in small samples was verified, as was the better performance of its correction, HC1, in adjusting these conditions better. In general, all the sandwich estimators underestimate the SE to a certain degree; the differences between them are minimum and mainly present in small samples. The box-plots (fig. 1, 2 and 3) reveal that sandwich estimators are more precise than non-parametric bootstrapping and jackknifing, the two resampling estimators, especially in samples where *n*= 50 and *n*=100, λ =0.3 and λ =1.

These two estimators, bootstrapping and jackknifing, perform in practically the same manner. Table 2 illustrates an almost perfect approximation of the SE by Huber's M-estimators. Small differences can be seen in the most adverse conditions, in which both estimators present slight underestimations. Non-parametric bootstrapping presents a better

estimate of the SE in small samples with a low mean, although it must be pointed out that these estimates are marked by their low precision (fig. 1), which leads to a greater oscillation in estimated values, providing values far removed from real values in some situations. Jackknifing makes a more exact adjustment as from a mean of 1 in small samples on all other occasions.

As the results did not discriminate clearly between the different indices and estimators, the RMSE goodness of fit index was calculated which, in our case, reported the average error between each value given by the different indices and estimators and the real value.

Table 3: RMSE values

N	λ	ϕ	D/df	χ^2/df	HC0	HC1	HC2	HC3	HC4	Boot	Jack
50	0.3	3	55.8174	25.7501	48.6193	45.7776	48.4581	49.8022	65.3416	96134.0	43634.4
		2	17.0303	8.8998	16.4181	15.4411	16.2929	16.7321	21.3165	32697.8	20843.6
		1.25	4.7946	3.8957	48.6193	6.2853	6.4376	6.3664	6.5594	7006.36	12.3644
	1	3	1.0901	1.3679	2.7986	2.7462	2.7942	2.8889	2.9276	3.6234	5.2768
		2	.4831	.8108	1.6770	1.5809	1.6226	1.6328	1.6598	1.7338	2.2036
		1.25	.2281	.4345	.8136	.7345	.7597	.7451	.7598	.8284	.9122
	5	3	.1133	.1647	.2951	.3180	.3118	.3487	.3345	.3483	.4291
		2	.0813	.0922	.1746	.1704	.1714	.1812	.1777	.1788	.2025
		1.25	.0609	.0565	.1039	.1053	.1046	.1136	.1099	.1059	.1097
100	0.3	3	15.1223	3.3001	8.4197	8.1731	8.3336	8.2917	8.5471	2144.60	12.7508
		2	6.7418	1.8623	4.1420	3.9741	4.0511	3.9869	4.0693	146.749	5.4759
		1.25	1.4038	.7308	1.4031	1.3567	1.3767	1.3671	1.3854	1.7217	1.7836
	1	3	.7368	.4214	1.0216	.9920	1.0071	1.0045	1.0163	1.0752	1.3332
		2	.1489	.2172	.4780	.4700	.4750	.4801	.4827	.4922	.5600
		1.25	.0707	.0994	.1904	.1912	.1912	.1971	.1952	.2213	.2287
	5	3	.0299	.0415	.0880	.0842	.0855	.0855	.0861	.1023	.0973
		2	.0205	.0236	.0457	.0462	.0461	.0481	.0473	.0520	.0491
		1.25	.0149	.0152	.0287	.0273	.0277	.0277	.0278	.0313	.0272
250	0.3	3	4.8977	.5288	1.4578	1.4502	1.4574	1.4595	1.4656	1.5138	1.8306
		2	1.8448	.2453	.6451	.6440	.6467	.6501	.6524	.7243	.7918
		1.25	.4833	.1168	.2390	.2331	.2354	.2328	.2348	.2751	.2518
	1	3	.2078	.0673	.1809	.1810	.1814	.1828	.1829	.1987	.2051
		2	.0498	.0381	.0881	.0859	.0867	.0858	.0865	.0996	.0865
		1.25	.0120	.0163	.0330	.0327	.0328	.0330	.0330	.0459	.0346
	5	3	.0048	.0067	.0145	.0142	.0143	.0143	.0144	.0217	.0165
		2	.0033	.0041	.0086	.0084	.0085	.0085	.0085	.0122	.0080
		1.25	.0023	.0024	.0044	.0042	.0043	.0042	.0043	.0069	.0043

In general, direct indices had the lowest RMSE values, followed by the HC3 and HC4 sandwich procedures. This would seem to support that direct indices provide the most precise estimate, as can be seen in the box-plots, but that does not include accuracy, since the index box-plots do not include the parameter's real SE value in some cases.

By size, the indices analysed produced the most notable differences in small samples (n=50). The index based on the deviance should not be used when n=50 and λ=.3, for any value of Φ, as it is as biased as the Poisson model. In this case, bootstrap and jackknife estimator averages are closest to the real value, although since they provide very high RMSE values, their low precision must be taken into account. Compared to all the sandwich estimators, the direct index based on the chi-square offers the mean closest to the real mean and is also the most precise of all, although on average it underestimates the real value more than resampling indices do.

The index based on the deviance presents the M-estimator value closest to the real value as well as the lowest RMSE value when n=50, λ=1 and Φ=1.25, and thus appears to be the best possible adjustment in this condition.

When n=50, λ=1 and Φ=2, the jackknife resampling index provides values closest to real values, followed by the chi-square and bootstrap indices; however jackknifing presents a RMSE three times higher than the chi-square index's, while the bootstrapping index is 22% lower than the jackknife's.

For the other situation, when n=50, λ=1 and Φ=3, the jackknife estimator presents the mean closest to the real one, followed by the chi-square and bootstrap indices. As in the previous cases, the index based on chi-square is the most precise and the jackknife's RMSE value is four times higher than the chi-square's.

When n=100, λ=.3 and λ=1 for any value of Φ, the index based on the deviance should not be used, as it is always more biased than any other index. The jackknife index is preferable in these situations, followed by the chi-square and bootstrapping indices, although the resampling indices' RMSE is at least double the chi-square's and its average values are very close to both each other's and the real value.

This is also true when n=250, λ=.3 and λ=1 for any value of Φ, in which the jackknifing, chi-square and bootstrap indices are the most exact. In these situations, the most biased index is based on the deviance while the most stable index is based on chi-square, whose RMSE is at least half that of any other index.

When λ=5 for any size and any value of Φ, the best estimates are given by the resampling and direct indices based on the deviance or on the chi-square although, as with other indices in accuracy, these are the most precise.

5. Discussion

This article has suggested two alternative procedures for correcting overdispersion: post-hoc correction of SE and estimating SE.

When the Poisson regression model is overdispersed, it is known that the model's parameter estimates are unbiased, but the parameters' SE are underestimated, which seriously affects the inferential process.

In the overdispersed Poisson approach, the analyst can correct the inferential statistics from the Poisson regression by estimating an overdispersion parameter and using it to correct the inferential statistics. This is likely to be the method of choice when the researcher is interested primarily in the hypothesis tests about the regression coefficients of the log-linear model (Gardner, Mulvey, and Shaw, 1995).

This study has attempted to ascertain the importance of both sample size and degree of overdispersion in data in estimating the original SE as well as in correcting them.

It has been proved how applying an unadjusted PRM will cause the SE to be underestimated even though the size of the sample increases, as Sturman (1999) indicates, this causes an excess of false positives, thus proving the need to resolve the overdispersion problem.

At the general level, when the value of the scale parameter is fixed, the overall behaviour of the set of indices remains the same, regardless of the value of lambda or size. When the value of lambda is fixed, the higher the scale parameter increases and the bias of the indices

is steeper and less precise. On the other hand, as expected, increasing sample size decreases error estimates and increases precision.

As in Dean, Eaves and Martinez (1995) and Breslow (1996), it was found that the sandwich estimators often appear to underestimate real variability with small and moderate-sized samples. Literature on the subject points to the HC3 (Davidson and Mackinnon, 1993; Long and Erving, 1998, 2000) and the HC4 (Cribari-Neto, 2004) as the most suitable alternatives to HC0 of all the sandwich estimators compared and in this research, these sandwich estimators performed better in the situations analysed.

Nevertheless, in all cases, sandwich estimators were less accurate than resampling estimators and also lower than the chi-square-based index.

As for correcting and estimating the standard error of Poisson regression coefficients, the results clearly indicate that resampling estimators, basically jackknifing, are less biased than indices based on the direct correction of standard error, basically chi-square, or robust estimators; however, it is also clear that these estimators have the highest RMSE values and are therefore less precise in almost all situations, especially in small samples. Thus, our results follow the lines suggested by Cameron and Trivedi (1998), who recommend resampling estimators for small samples.

Nevertheless, we believe that various bootstraps should be run with small samples to analyse their stability and thus be able to decrease the effect its low precision may have on the estimated value; in this sense, continued research on procedures to make bootstrapping more precise is needed.

In large samples, n=250, there are practically no differences between the values provided by the different indices and estimators, except the index based on deviance, which is why the preferable index is the most precise of all of them, which in our study was the chi-square index, whose RMSE was 65% lower than the resampling's with this size sample.

In general, direct indices are more precise, followed by sandwich estimators and resampling estimators in third place; however, in general all the indices maintained a bias compared to the real value, especially in small and medium-sized samples. In view of the results, future research may be oriented towards finding correction factors for these

estimators and indices which would correct this bias and enable them to produce results closer to real values.

On of the most pointed results of this study is that in general, the index based on chi-square appears to be a good corrector of SE, because its average values are close to real value and because it is one of the most precise indices of all those compared.

To conclude, a good choice when having to modifying the SE would be using both types of indices, chi-square and resampling (bootstrap and jackknife), and evaluating where they coincidence. When they coincide, either of them could be chosen, with a preference for resampling and especially for jackknifing. Running the study of the indices' sample distribution by resampling is recommended when the chi-square and resampling indices do not coincide and the most suitable mean value should be chosen on that basis.

6. Referencias

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Breslow, N.E., 1996. Generalized linear models: checking assumptions and strengthening conclusions. Statistica Applicata 8, 23-41.

Cameron, A. C., Trivedi, P. K., 1986. Econometric models based on count data: comparisons and applications of some estimators and tests [Electronic version]. J. Appl. Econom. 1, 29-53.

Cameron, A. C., Trivedi, P. K., 1990. Regression-based tests for overdispersion in the Poisson model. J. Econometrics 46, 347-364.

Cameron, A. C., Trivedi, P. K. 1998. Regression Analysis of Count Data. Econometric Society Monographs, 30. Cambridge: Cambridge University Press.

Cox, D. R., Snell, E. J., 1989. Analysis of Binary Data. London: Chapman and Hall.

Cribari-Neto, F., 2004. Asymptotic inference under heteroskedasticity of unknown form. Comput. Stat. Data. An. 45, 215-233.

Cribari-Neto, F., Zarkos, S.G., 1999. Bootstrap Methods for Heteroskedastic Regression Models: Evidence on Estimation and Testing. Econometric Reviews 18, 211-228.

Cribari-Neto, F., Zarkos, S.G., 2001. Heteroskedasticity-Consistent Covariance Matrix Estimation. J. Statis. Comput. Sim. 68, 391-412.

Davidson, R., Mackinnon, J. G., 1993. Estimation and Inference in Econometrics. Oxford University Press.

Dean, C.B., Eaves, D.M., Martinez, C.J., 1995. A comment on the use of empirical covariances matrices in the analysis of count data. J. Stat. Plan. Infer. 48, 197-205.

Efron, B., 1979. Bootstrap methods: another look at the Jackknife. Ann. Statist. 7,1-26.

Efron, B., 1982. The jackknife, the bootstrap and another resampling plans. CBMS Regional Conference Series in Applied Mathematics 38. Philadelphia: SIAM Publications.

Eicker, F., 1967. Limit Theorems for Regressions With Unequal and Dependent Errors, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley , CA: University of California Press, 59-82.

Gardner, W., Mulvey, E., Shaw, E., 1995. Regression analyses of counts and rates: Poisson, overdispersed Poisson and negative binomial models. Psychol. Bull. 118, 392-404.

Greene, W. H., 1993. Econometric Analysis. New York: Macmillan.

Greene, W. H., 2000. Econometric analysis (4th Ed). New York: Prentice Hall.

Guo, J.Q., Li, T., 2002. Poisson regression models with errors-in-variables: implication and treatment. J. Stat. Plan. Infer., 104, 391-401.

Gurmu, S., 1991. Tests for detecting overdispersion in the positive Poisson regression model. J. Bus. Econ. Stat. 9, 215-222.

Heinzl, H., Mittlböck, M., 2003. Pseudo R-squared measures for Poisson regression models with over- or under-dispersion. Comput. Stat. Data. An. 44, 253-271.

Hinkley, D.V., 1977. Jackknifing in Unbalanced Situations. Technometrics 19, 285-292.

Huber, P. J., 1967. The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions. In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, pp.221-233.

Kauermann, G., Carroll, R.J., 2001. A note on the efficiency of Sandwich Covariance Matrix Estimation. J. Am. Stat. Assoc. 96, 1387-1396.

Krzanowski, W. J., 1998. An Introduction to Statistical Modelling. London: Arnold

Lee, L.F., 1986. Specification test for Poisson regression models. Int. Econ. Rev. 27, 689-706.

Liao, T.F., 1994. Interpreting Probability Models. Logit, Probit and other Generalized Linear Models. London: Sage

Lindsey, J. K., 1998. Counts and times to events. Stat. Med. 17, 1745-1751.

Lindsey, J. K., 1999. On the use of corrections for overdispersion. Appl. Stat.-J. Roy. St. C 48(4), 553-561.

Long, J.S,. Ervin, L.H, 1998. Correcting for Heteroscedasticity with Heteroscedasticity Consistent Standard Errors in the Linear Regression Model: Small Sample Considerations. Working Paper.

Long, J.S, Ervin, L.H., 2000. Using Heteroskedasticity Consistent Standard Errors in the Linear Regression Model. Amer. Statist. 54, 217-224.

Mackinnon, J.G., White, H., 1985. Some Heteroskedasticity-consistent covariance matrix estimators with improvement finite samples properties. J. Econom. 29, 305-325.

McCullagh, P., Nelder, J. A. 1989. Generalized linear models, 2nd Edition. London: Chapman & Hall.

Miller, R.G., 1974. The Jackknife – a review. Biometrika 61, 1-15. New York: Oxford University Press.

Quenouille, M. H., 1949. Aproximate test of correlation in time series. J. Roy. Stat. Soc. B 11, 68-84.

Quenouille,M. H., 1956. Notes on bias in estimation. Biometrika 43, 353-360.

Sturman, M.C., 1999. Multiple approaches to analyzing count data in studies of individual differences: The propensity for Type I errors, illustrated with the case of absenteeism prediction. Educ. Psychol. Meas. 59, 414-430.

Tukey, J.W., 1958. Bias and confidence in not quite large samples (Abstract). Ann. Math. Stat. 29, 614.

Wedderburn, R. W. M., 1974. Quasi-likelihood functions, generalized linear models, and Gauss-Newton methods. Biometrika 61, 439-447.

White, H., 1980. A Heteroskedastic-Consistent Covariance Matrix Estimator and Direct Test of Heteroskedasticity. Econometrica 48, 817-838.

Winkelmann, R., 2000. Econometric Analysis of Count Data, 3rd Edition. Berlin: Springer-Verlag.

Wu, C.F.J., 1986. Jackknife, bootstrap and another resampling methods in regression analysis. Ann. Statist. 14, 1261-1295.

Modelado del número de días de consumo de cannabis.

Artículo publicado en Psicothema, 2005. Vol. 17. Reproducción autorizada.

Modelado del número de días de consumo de cannabis

Palmer, A., Llorens, N. y Perelló, M.J.

Resumen

Este artículo pretende mostrar, desde un punto de vista práctico, la manera adecuada de modelar una variable de respuesta de tipo recuento. Los resultados muestran que el modelo de regresión de Poisson es un modelo más adecuado que el de regresión lineal, ya que tiene en cuenta las características propias de los datos de recuento, aunque la presencia de sobredispersión indicará la corrección de los errores estándar de los parámetros del modelo de Poisson, y puede provocar el modelado mediante el modelo binomial negativa. El modelo de ceros aumentados será usado cuando haya un número excesivo de ceros. Para ello se modela el número de días de consumo de cannabis en función del consumo por parte del grupo de iguales y del consumo de cannabis del padre, de la madre y de los hermanos.
Palabras clave: recuento, regresión de Poisson, Binomial Negativa, ceros aumentados, cannabis.

Abstract

This article aims to show the appropriate way to model a count type response variable from a practical viewpoint. The results show that the Poisson regression model is more appropriate than the linear regression model, as it takes count data characteristics into consideration, although the presence of overdispersion indicates the correction of standard errors in the Poisson model parameters and may cause modelling through the negative binomial model. The Zero Inflated model shall be used whenever there is an excessive number of zeros. To do so, the number of days of cannabis consumption in accordance with peer group consumption and cannabis consumption by parents and siblings is modelled.
Keywords: count, Poisson regression, Negative Binomial, Zero inflated, cannabis.

Innumerables teorías y modelos intentan explicar la conducta de consumo y el abuso de sustancias adictivas. En una revisión realizada por Lettieri, Seyers y Pearson (1980) se encontraron más de cuarenta perspectivas teóricas que intentan explicar los problemas y las conductas de consumo de sustancias adictivas. El paso de las drogas legales a las ilegales también ha sido muy estudiado, el modelo de Danise Kandel defiende la hipótesis de la escalada donde se postulan la existencia de estadios secuenciales en el consumo de drogas (Kandel,1975). Según esto, el consumo de cannabis es el paso intermedio entre el no consumo de drogas ilegales y el consumo de drogas potencialmente de riesgo como la cocaína y la heroína. Este modelo propone la existencia de una progresión secuencial en la conducta de consumo de drogas y distingue dos grandes tipos de variables predictoras en la conducta de consumo: la historia del consumo de drogas y las variables psicosociales del sujeto. Dentro de las variables psicosociales, la Teoría del aprendizaje social de Bandura (1977) explica la conducta humana como un fenómeno de aprendizaje basándose en las leyes del condicionamiento clásico, operante y vicario, haciendo hincapié en el poder predictivo de las variables antecedentes familiares (Muñoz-Rivas y Graña, 2001; Martínez, Fuertes, Ramos y Hernández, 2003; Pérez y Delgado, 2003; Delgado y Pérez, 2004) e influencia del grupo de iguales (Gonzalez, García–Señoran y Gonzalez, 1996) en el consumo de drogas. Por su parte, Kandel y Davies (1992) encontraron que la influencia ejercida por los compañeros en el consumo de cannabis es muy importante, pero no tanto para el consumo de alcohol u otras drogas. Sin embargo, la influencia de los padres tenía un gran valor predictivo para el consumo de drogas ilegales diferentes al cannabis.

Toda conducta de consumo sigue un proceso, hay una serie de fases por las que se pasa del uso al abuso y a la posterior dependencia. El proceso es lento y complejo pero predecible. Es importante detectar en qué punto de ese proceso se encuentra el sujeto, y también puede resultar útil saber qué variables de las señaladas por las diferentes teorías están afectando más, en la conducta de consumo de sustancias, en el punto concreto en el que se encuentra el sujeto. Para ello se ha analizado el número de días que han consumido cannabis en el último mes, tomada como variable de recuento.

Hay pocos estudios que realmente utilizan las técnicas estadísticas adecuadas para este tipo de variables. Concretamente, el problema que interesa resolver es como una o más variables pueden explicar o predecir el número de ocurrencias (recuento) que se dará de un hecho determinado. Hasta ahora se ha utilizado con frecuencia el modelo de regresión lineal, pero como señala Long (1997, p.217) una variable de recuento tiene unas características que hacen que la utilización del modelo de regresión lineal para su modelado producirá estimaciones ineficientes, inconsistentes y sesgadas, mientras que el modelo de regresión de Poisson (MRP) es el modelo específico para analizar este tipo de datos.

El siguiente estudio pretende analizar, en una muestra de jóvenes adultos, el valor predictivo de las variables psicosociales en el consumo de sustancias. Dentro de las variables de historia de consumo se ha tenido en cuenta el consumo actual de cannabis, porque como señalan Martinez y Robles (2001), el cannabis es la droga ilegal más consumida en la actualidad considerándose también la droga puente entre el consumo de drogas legales e ilegales. Dentro de las variables psicosociales hemos tenido en cuenta la influencia de la familia y la influencia del grupo de iguales.

Método

Participantes

Se utilizó un muestreo incidental llevado a cabo en varios edificios del Campus Universitario correspondientes a diferentes estudios, así como en zonas de marcha de la ciudad. La muestra estaba formada por 314 sujetos de edades comprendidas entre los 18 y los 30 años. La edad media de los sujetos fue de 22 años. Por sexos, el 53.9% eran hombres y el 46.1% eran mujeres.

Procedimiento

Se les administró un cuestionario anónimo desarrollado para medir la frecuencia de uso de diferentes sustancias adictivas, entre ellas el cannabis, así como el consumo del grupo de iguales y antecedentes familiares en el consumo. Los cuestionarios fueron administrados por un entrevistador, que preguntaba personalmente a los sujetos de forma individual y recogía las respuestas dadas por estos.

El modelado utilizado para analizar los datos de recuento, número de días de consumo de cannabis durante el último mes, se enmarca dentro del Modelo Lineal Generalizado. El análisis se realizó con el programa Stata 8.0.

Variables

La variable *consumo de cannabis* por parte del grupo de iguales era una variable categórica, representada en el modelo por las variables codificadas c1 (la mayoría), c2 (la mitad) y c3 (pocos) con categoría de referencia “ningún amigo consume”. La variable *antecedentes familiares* se registró separando el consumo de cannabis del padre, de la madre y de los hermanos. Las tres variables antecedentes familiares se introdujeron en el modelo mediante las variables codificadas, pc1 (Si), pc2 (No

procede) para el padre, mc1 (Si), mc2 (No procede) para la madre y hec1 (Si), hec2 (No procede) para los hermanos, todas respecto a la categoría No consumo.

Resultados

Puesto que la variable número de días de consumo es una variable cuantitativa, tiende a ser modelada mediante el modelo de regresión lineal, lo que llevaría a los resultados mostrados en la parte izquierda de la tabla 1.

Tabla 1

Modelo de regresión lineal y con transformación logarítmica

	Modelo Regresión Lineal			MRL con transformación Logarítmica		
	Coef.	EE	P	Coef.	EE	P
c1	9.1151	1.1769	<0.001	2.1888	.2048	<0.001
c2	1.3819	1.1368	0.224	.6215	.1978	0.002
c3	.2515	1.2189	0.837	.2183	.2121	0.303
pc1	4.4806	2.9662	0.131	.7338	.5162	0.155
pc2	-1.7955	1.1432	0.116	-.2927	.1989	0.141
mc1	8.5836	2.2195	<0.001	.8591	.3862	0.026
mc2	-.5697	2.5887	0.826	-.0116	.4505	0.979
hec1	1.7565	.8764	0.045	.2108	.1525	0.167
hec2	-.2238	.8884	0.801	.0426	.1546	0.783
Const	.1655	1.0017	0.869	-.6776	.1743	<0.001

A partir de esta tabla puede verse que el modelo de regresión lineal nos dice que las variables c1, mc1 y hec1 son las variables explicativas significativas. Así pues, que todos los amigos consuman, que la madre consuma y que los hermanos consuman aumenta la probabilidad de que el sujeto consuma un mayor número de días. Sin embargo, si atendemos a las características de la variable respuesta utilizada, nos damos cuenta de que es una variable cuantitativa pero no continua sino discreta, y que tan solo puede tomar valores enteros y no negativos. Y además, cada dato indica el número de

veces (días) que el suceso “tomar cannabis” ha sido repetido en cada sujeto, lo que define dicha variable respuesta como una variable de recuento.

Un hecho paradigmático de que el modelo de regresión lineal no es adecuado para datos de recuento es que hace predicciones negativas para una variable que, claramente, no admite valores negativos. En concreto, en este modelo se obtienen 32 valores predichos inferiores a cero.

En general, en estas situaciones se acostumbra a realizar una transformación logarítmica en la variable respuesta (logdias), ya que esto proporciona una distribución cercana a la Normal, lo que posibilita su manejo mediante MCO (Mínimos Cuadrados Ordinarios). En la parte derecha de la Tabla 1 pueden verse los resultados del modelo de regresión lineal con la variable respuesta transformada en escala logarítmica.

El modelo de regresión con transformación logarítmica de la variable respuesta ajusta mejor que el modelo de regresión lineal ya que el índice AIC (Akaike Information Criterion) pasa de 6.519 a 3.022. Sin embargo, con la transformación logarítmica, existen por un lado problemas de estimación ya que un valor y=0, frecuente en una variable de recuento, necesita ser transformado para poder ser utilizado, en general sumándole una pequeña cantidad, y por otra parte existen problemas de interpretación ya que, aunque se cumpla que exp[log(y)] sea igual a *y*, el valor predicho por la ecuación viene dado por exp[E(log(y))] el cual es diferente al valor de E(y).

Así pues, será necesario elegir el modelo adecuado a este tipo de variable. En la Tabla 2 se presenta el modelado de la variable numero de días de consumo de cannabis por medio del modelo de regresión de Poisson (MRP). Los resultados señalan que el hecho de que los amigos consuman cannabis aumenta el número de días de consumo en un sujeto. El valor tan alto de los coeficientes será discutido posteriormente a partir de los

datos de la Tabla 4. En cuanto a la influencia de los antecedentes familiares, vemos que el hecho de no tener padre, respecto a tenerlo y que éste no consuma, disminuye un 40% el número de días de consumo de cannabis. Por lo que respecta a la madre, si ésta consume el hijo consumirá cannabis el doble de días que si la madre no consume. El consumo de los hermanos aumenta en un 42% el número de días de consumo.

Tabla 2

Modelos de Poisson, Poisson corregido y Binomial Negativa

	Modelo de Poisson			Modelo Poisson Corregido		Modelo Binomial Negativa		
Días	Coef.	EE	P	EE	P	Coef.	EE	P
C1	5.9263	1.0008	<0.001	2.3517	0.012	5.9419	1.0509	<0.001
C2	4.3359	1.0027	<0.001	2.3561	0.066	4.3517	1.0496	<0.001
C3	3.3054	1.0124	0.001	2.3789	0.165	3.3410	1.0680	0.002
Pc1	.0836	.1218	0.493	.2864	0.770	.1100	.8155	0.893
Pc2	-.4943	.1034	<0.001	.2430	0.042	-.4728	.3829	0.217
Mc1	.7548	.0954	<0.001	.2243	0.001	.5609	.6179	0.364
Mc2	-.2758	.1990	0.166	.4678	0.555	.8990	.9118	0.324
Hec1	.3556	.0650	<0.001	.1529	0.020	.4497	.2635	0.088
Hec2	-.0603	.0880	0.493	.2070	0.771	-.0233	.2979	0.938
Const	-3.7725	1.0005	<0.001	2.3510	0.109	-3.8515	1.0378	<0.001

Los resultados dados hasta el momento parecen apoyar las teorías planteadas al inicio del artículo, no obstante cuando trabajamos con el MRP es importante comprobar el supuesto básico de equidispersión y en su defecto la aparición de sobredispersión que, como señalan McCullagh y Nelder (1989), es la norma en datos de recuento. El principal problema de la sobredispersión es que estando bien especificado el modelo, las estimaciones de los parámetros son correctas pero no sus errores estándar, lo que comporta una sobreestimación del valor de la prueba de conformidad del parámetro así como de la amplitud de su intervalo de confianza. Un primer indicio de la existencia de

sobredispersión se tiene a partir de los resultados del MRP en los que se comprueba que el valor del cociente entre la Discrepancia y sus grados de libertad, de valor 5.52, está alejado del valor 1 que indicaría equidispersión. Se ha comprobado la sobredispersión de los datos a través de la prueba basada en la regresión (Cameron y Trivedi, 1990) cuyo resultado t=5.85 ($p<0.01$) indica que nos encontramos ante datos sobredispersos.

Ante la sobredispersión existen dos opciones, como señalan entre otros Hardin y Hilbe (2001): se puede realizar un ajuste *post hoc* de los errores estándar, utilizando para ello los diferentes índices que existen para tal fin, o por otro lado modelar con un modelo que sea más tolerante con la falta de equidispersión, como el modelo de regresión de la Binomial Negativa (MRBN) (Lindsey, 1995).

Realizamos el ajuste de los errores estándar por medio de la raíz cuadrada del parámetro de dispersión, porque se ha señalado, entre otros, como un ajuste adecuado. Esto significa multiplicar, en este caso, los errores estándar obtenidos en el MRP por un factor de corrección de valor 2.35.

Los resultados de esta corrección se muestran en la parte central de la tabla 2. En ella se observa el aumento de los errores estándar respecto a los del MRP original, lo que produce una disminución del valor de la prueba z de conformidad de los coeficientes del modelo, así como de sus intervalos de confianza, y esto conlleva un menor número de variables consideradas predictoras en el MRP. Concretamente podemos comprobar que ahora no son significativas las variables c2 y c3.

El modelo más ampliamente utilizado en situaciones de sobredispersión es el MRBN ya que es capaz de recoger la sobredispersión causada por heterogeneidad no observada. En este modelo la variancia viene dada por $V(y) = \mu + \alpha\mu^2$, por lo que se necesita una

estimación de la constante alfa que, por máxima verosimilitud, en nuestro caso, vale 2.706.

Utilizando el MRBN se observa (parte derecha de la tabla 2) como ninguna de las variables indicadoras relacionadas con la familia aparecen como significativas, resultando significativa solo la relación en función de los amigos. Se comprueba así, cómo varían los resultados en función del procedimiento estadístico aplicado.

MRP versus MRBN

La no-adecuación del MRP puede verse a través de los índices BIC y AIC ya que en este modelo sus valores son de –68.12 y 6.95, siendo –1493.21 y 3.79 en el MRBN, lo que indica un mejor ajuste del MRBN.

Una manera de comparar la eficacia de ambos modelos es por medio de los residuales de discrepancia producidos en cada modelo. Las observaciones ajustadas correctamente por un modelo tendrán unos residuales que se moverán en el intervalo –2 a +2. En la Figura 1 puede verse cómo prácticamente todos los residuales del MRBN se encuentran en el intervalo adecuado, mientras que muchos de los residuales del MRP caen fuera del intervalo adecuado. El grafico ha sido dividido en 9 rectángulos para poder visualizar mejor donde se producen las observaciones mal ajustadas. Así, se observa que todas las observaciones ajustadas por Poisson también son bien ajustadas por la Binomial Negativa, pero observaciones mal ajustadas por MRP son ajustadas correctamente por el MRBN.

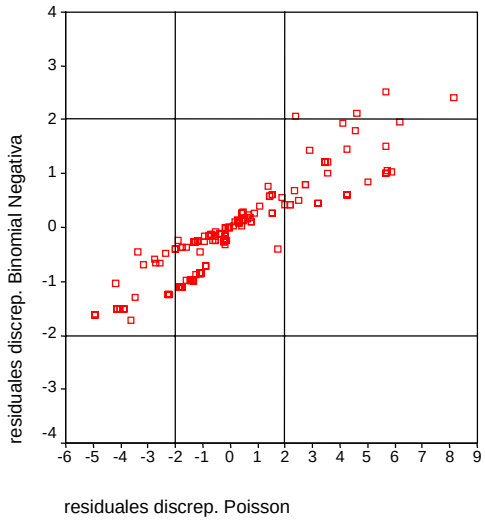


Figura 1. Residuales de discrepancia en los modelos de Poisson y Binomial Negativa

Así pues, mientras en el MRP hasta 91 observaciones, un 28.86%, están mal ajustadas, solo 4 observaciones, un 1.27%, están mal ajustadas en el MRBN, lo que significa que ésta consigue estimar correctamente el 91.6% de las observaciones mal ajustadas por el MRP. Todo ello implica que el modelo MRBN se adecua muy bien a los datos, ya que ajusta correctamente el 98.73% de las observaciones.

Otra manera de evaluar la diferencia entre MRP y MRBN es por medio de las probabilidades predichas por cada modelo respecto a los valores observados. Ambos modelos proporcionan una tasa predicha de 1.39 lo que indica que las estimaciones del modelo de Poisson son consistentes, aun en presencia de sobredispersión. Sin embargo las probabilidades predichas por cada modelo indican que el MRBN ajusta mucho mejor los ceros observados y su distribución es mucho más dispersa que la del MRP, ya que realiza predicciones en recuentos superiores a 8 cuando Poisson prácticamente los descarta. Por otra parte la Binomial Negativa tampoco ajusta correctamente los

recuentos inferiores a 5 aunque su ajuste, en todos los casos, es mejor que el realizado por Poisson.

Modelo de ceros aumentados

La existencia de un número excesivo de ceros, es decir un número superior al predicho por el modelo, nos puede indicar la existencia de una mezcla de distribuciones. En nuestro caso es habitual encontrar esta situación, ya que en la muestra pueden haber sujetos que no sean consumidores de cannabis por lo que su consumo en el último mes será cero, aunque sean sujetos expuestos a la posibilidad de consumir. El modelo de ceros aumentados (se utilizará la notación ZIP (Zero-Inflated Poisson), por ser la notación estándar para este tipo de modelo) permite distinguir entre dos procesos conducentes a que un sujeto tenga valor cero: por un lado los ceros estructurales (sujetos no consumidores) y por otro los ceros aleatorios (consumidores que no han consumido). Si se aplica el modelo ZIP se obtienen los resultados expuestos en la Tabla 3.

Tabla 3: Modelo de ceros aumentados (ZIP)

	Coef.	EE	P
C1	6.1995	1.0018	<0.001
C2	5.4679	1.0037	<0.001
C3	5.1132	1.0146	<0.001
Pc1	-.0104	.1189	0.930
pc2	-.2095	.1041	0.044
mc1	.5368	.0936	<0.001
mc2	-.2730	.1996	0.171
hec1	.2580	.0652	<0.001
hec2	-.1943	.0887	0.028
Const	-3.7381	1.0014	<0.001

El test de Vuong (Vuong, 1989) permite comparar el modelo ZIP frente al modelo de Poisson, es decir modelos no anidados, proporcionando un valor z=6.97 (P<0.0001) según el cual el modelo ZIP proporciona un mejor ajuste que el modelo de Poisson.

En la Figura 2 se comparan las diferencias entre lo observado y lo predicho en cada recuento, para cada uno de los modelos: MRP, MRBN y ZIP.

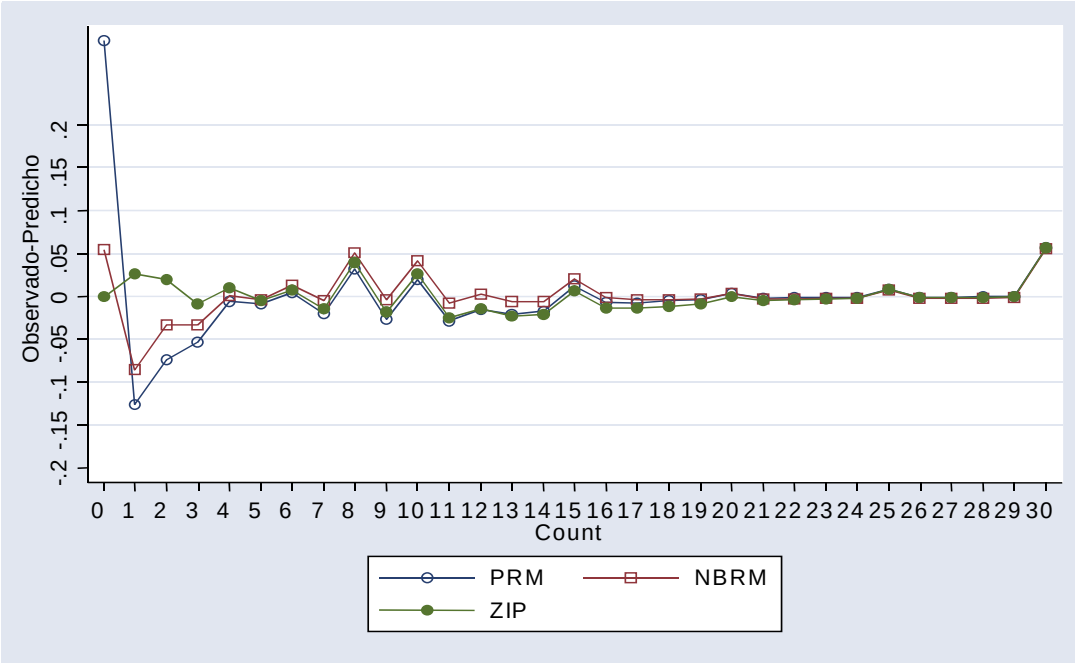


Figura 2. Comparación Observado menos Predicho entre Poisson (MRP), Binomial Negativa (NBRM) y ZIP

En la figura 2 se comprueba que el número de ceros es predicho de forma perfecta por el ZIP. Asimismo podemos ver que, en general, el ajuste del ZIP es mejor que el realizado por el MRBN, sobre todo en recuentos bajos.

En el modelado mediante el ZIP se obtiene que el grupo de amigos es fundamental a la hora de explicar el número de días de consumo de cannabis, observándose que los coeficientes asociados a estas variables son bastante altos. Para explicar esta relación

podemos analizar los datos de la Tabla 4 en la que se especifican únicamente las frecuencias de los sujetos consumidores.

Tabla 4

Días de consumo y número de amigos consumidores

Días	Mayoría	Mitad	Pocos	Ninguno
1-7	24	20	8	1
8–14	29	10	2	
15–21	8	3		
22–30	21			

Podemos observar que prácticamente no hay ningún consumidor de cannabis que tenga un grupo de amigos en el que no haya ningún consumidor ya que esto tan solo ocurre en un 0.8% de los casos. Cuando se tienen pocos amigos consumidores, lo que se da en un 8% de los casos, los sujetos mayoritariamente consumen de 1 a 7 días al mes, lo que ocurre en un 80% de estos casos, mientras que el 20% restante lo hacen de 8 a 14 días. Cuando la mitad de los amigos son consumidores, lo que ocurre en un 26% de los casos, un 9% de los sujetos ya consume entre 15 y 21 días al mes, un 30% lo hace entre 8 y 14 días, mientras que un 61% lo hace de 1 a 7 días. Cuando la mayoría del grupo de amigos son consumidores, lo que se da en un 65% de los casos, un 26% consume más allá de 22 días al mes incluyendo a los que consumen diariamente, casi un 10% fuma entre 14 y 21 días al mes, un 35% fuma entre 7 y 14 días y el 29% restante consume de 1 a 7 días. Así pues, a partir de la estructura triangular de la tabla de contingencia se deduce que a medida que aumenta el número de amigos que consumen aumenta también el número

de días de consumo, tal como refleja la magnitud de los coeficientes de las variables c1, c2 y c3.

Asimismo se obtiene que, no tener padre respecto a tenerlo y que éste no consuma disminuye un 19% el número de días de consumo de cannabis. Por lo que respecta a la madre, si ésta consume aumenta un 71% el número de días de consumo del hijo, respecto a que la madre no consuma. Los hermanos también influyen, de manera que si éstos consumen aumenta un 29% el número de días de consumo, y el ser hijo único disminuye un 18% el número de días de consumo respecto a tener hermanos no consumidores.

Discusión

El consumo de cannabis ha ido aumentando e instaurándose en la sociedad con una gran aceptación en las últimas décadas. El papel de los amigos es fundamental en el número de días de consumo que se realizará del cannabis, así como los antecedentes familiares, tanto del padre, de la madre como de los hermanos.

En un primer análisis a través del MRP corregido, debido al incumplimiento de la equidispersión, y posteriormente con el modelo ZIP, hemos obtenido que el papel de la familia es fundamental en la frecuencia de consumo de cannabis. Estos mismos resultados los obtuvieron previamente Duncan, Duncan y Hops (1990) que mostraron que el papel de la familia era determinante en el inicio y mantenimiento de la conducta de consumo de diferentes sustancias de los menores. Sin embargo, realizando los análisis con el MRBN, en nuestro estudio no se confirma que la familia modele la conducta de consumo que hacen los jóvenes adultos del cannabis.

Elliot, Huizinga y Ageton (1982) señalaron que el consumo por parte del grupo de iguales predecía la presencia y cantidad del consumo de sustancias en adolescentes. En

nuestro estudio se comprueba, a través del modelo de ceros aumentados ZIP, cómo el número de días que un sujeto va a consumir cannabis está parcialmente explicado por la cantidad de personas de su grupo que consumen la sustancia en cuestión y asimismo por los antecedentes familiares.

El presente estudio pretendía mostrar, utilizando las técnicas estadísticas adecuadas al tipo de datos a analizar, la capacidad explicativa de variables incluidas en el aprendizaje vicario. Parece demostrarse que un análisis estadístico no adecuado a los objetivos de las investigaciones, o un análisis no adecuado al tipo de datos a analizar, puede desembocar en resultados que no se ajusten tanto a la realidad. En Sturman (1999) se recogen algunos de los problemas que podemos encontrar al utilizar un procedimiento no adecuado a los datos con los que trabajamos.

La utilización de las técnicas estadísticas apropiadas en la investigación empírica puede considerarse un indicador de la madurez científica conseguida en ese ámbito, ya que estas técnicas son las que nos permitirán llegar a las mejores conclusiones basadas en nuestros datos.

Referencias

Bandura, A. (1977). *Social learning* theory. Englewood Cliff, NJ: Prentice-Hall.

Bobes, J., Bascarán, M.T., González, M.P. y Saiz, P.A. (2000) Epidemiología del uso/abuso de cannabis. *Adicciones,*12(2), 31-40.

Cameron, A.C. y Trivedi, P.K.(1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cameron, A.C. y Trivedi, P.K. (1998). *Regression analysis of count data*. New York, NY: Cambridge University Press.

Duncan, S.C., Duncan, T.E. y Hops, H. (1998). Progressions of alcohol, cigarette and marijuana use in adolescence. *Journal of Behavioral Medicine*, 21(4), 375-388.

Elliot, D.S., Huizinga, D., y Ageton, S.S. (1982). *Explaining delinquency and drug use.*, Boulder, CO: Behavioral Research Institute.

Golub, A. y Johnson, B.D. (1994). The shifting importance of alcohol and marijuana as gateway substances among serious drug abusers. *Journal of Studies on Alcohol*, 55(5), 607-614.

Gonzalez, F., Garcia-Señorán, M.M. y Gonzalez, S.G. (1996). Consumo de drogas en la adolescencia. *Psicothema*, 8(2), 257-267.

Hardin, J. y Hilbe, J. (2001). *Generalized Linear Models and Extensions.* College Station, TX: Stata Press.

Hops, H. (1990). Parent-adolescent problem solving interactions and drug use. *A Journal Drug Alcohol Abuse*, 16(3-4), 151-164.

Kandel, D. (1975). Stages in adolescent involvement in drug use. *Science*, 190(4217), 912-914.

Kandel, D.B. y Davies, M. (1992). Progression to regular marijuana involvement: Phenomenology and risk factors for near-daily use. En M. Glantz y R. Pickens (Eds.), *Vulnerability to drug abuse* (pp. 211-253). Washington, DC: American Psychological Association.

Lettieri, D.J., Sayers, M. y Pearson, H.W. (1980). *Theories on Drug Abuse: Selected contemporary perspectives.* NIDA Research Monograph 30. Rockville, MD: National Institute on Drug Abuse.

Lindsey, J.K. (1995). *Modelling frequency and count data.* Oxford: Clarendon Press.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Long, J.S. y Freese, J. (2003). *Regression models for categorical dependent variables using Stata.* College Station, Texas: Stata Press.

Martínez, J.L., Fuertes, A., Ramos, M. y Hernández, A. (2003). Consumo de drogas en la adolescencia: importancia del afecto y la supervisión parental. *Psicothema*, 15(2), 161-166.

Martínez, J.M. y Robles, L. (2001). Variables de protección ante el consumo de alcohol y tabaco en adolescentes. *Psicothema*, 13(2), 222-228.

McCullagh, P. y Nelder, J.A. (1989). *Generalized linear models.* (2ª ed.). London: Chapman & Hall.

Muñoz-Rivas, M.J. y Graña, J.L. (2001). Factores familiares de riesgo y de protección para el consumo de drogas en adolescentes. *Psicothema*, 13(1), 87-94.

Pérez, A. y Delgado, D. (2003). La codependencia en familias de consumidores y no consumidores de drogas: estado del arte y construcción de un instrumento. *Psicothema*, 15(3), 381-387.

Pérez, A. y Delgado, D. (2004). La codependencia en familias de consumidores y no consumidores de sustancias psicoactivas. *Psicothema*, 16(4), 632-638.

Sturman, M.C. (1999). Multiple approaches to analyzing count data in studies of individual differences: The propensity for Type I errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement, 59*(3), 414-430.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57(2)*, 307-334.

Las estrategias de afrontamiento: factores de protección en el consumo de alcohol, tabaco y cannabis.

Artículo publicado en Adiciones 2004; vol. 16(4):261-266. Reproducción autorizada.

Las estrategias de afrontamiento: factores de protección en el consumo de alcohol, tabaco y cannabis.

Noelia Llorens Aleixandre*, Miguel Perelló del Río, Alfonso Palmer Pol*****

* Profesor ayudante del Área de Metodología, Universidad de las Islas Baleares.
** Psicólogo práctica privada. Centro de Aplicaciones Psicológicas. Valencia.
*** Profesor Titular del Área de Metodología, Universidad de las Islas Baleares.

Enviar correspondencia a:
Noelia Llorens Aleixandre, e-mail: noelia.llorens@uib.es. Carretera Valldemossa, km. 7,5. Universitat de les Illes Balears. Edificio Guillem Cifre de Colonya. Palma de Mallorca, CP. 07122. Tfno. 971 172 495

Recibido: ¿? de ¿?¿?¿?¿?¿ de 200?.
Aceptado: ¿? de ¿?¿?¿?¿ de 200?.

RESUMEN

Objetivo: El objetivo del presente estudio es analizar el valor explicativo de las estrategias de afrontamiento, habilidades sociales y habilidades propias, en el consumo de diferentes sustancias adictivas legales e ilegales.

Método: Se administró un cuestionario anónimo, a una muestra de 314 jóvenes adultos, entre 18 y 30 años, desarrollado para medir la cantidad de sustancia consumida en una semana, las estrategias de afrontamiento, habilidades sociales y habilidades propias. Las variables de respuesta analizadas fueron el número de bebidas destiladas (whisky, vodka, etc..), número de bebidas fermentadas (cerveza, vino, etc...) y número de cigarrillos de cannabis y de tabaco.

Resultados: Déficits en estrategias de afrontamiento como "Pensar en las consecuencias negativas" disminuye el consumo de bebidas fermentadas y de cannabis en un 24% y en un 40% respectivamente. La habilidad propia "Ser disciplinado" disminuye el consumo de bebidas destiladas, fermentadas y cannabis, en un 38%, 31% y 33% respectivamente. Déficits en habilidades sociales, como la dificultad para relacionarse con personas del sexo opuesto, influyen disminuyendo el consumo de bebidas destiladas (45%), de cannabis (70%) y aumentando el consumo de tabaco por cuatro.

Conclusiones: Únicamente déficits en habilidades sociales influyen en el consumo de tabaco. Déficits en habilidades sociales y propias influyen en el consumo que los sujetos hacen de las bebidas destiladas y de cannabis y déficits en estrategias de afrontamiento son las que llevarán a un sujeto a consumir bebidas fermentadas. La modificación de todos éstos déficits puede permitir hacer programas de prevención más efectivos.

Palabras clave: *Estrategias Afrontamiento, Poisson, Sobredispersión, Cannabis, Tabaco, Alcohol.*

ABSTRACT

Purpose: The aim of this study is to analyse the explanatory value of coping skills, own skills and social skills, over use of different legal and illegal addictive substances in youth adults' population.

Methods: An anonymous questionnaire was administered to a sample 314 youths adults, between 18 and 30 years. The instrument measures the quantity of substance consumed in one week, coping skills, social skills and own skills. The analysed answer variables were the number of consumed distilled drinks (whisky, vodka, etc.), fermented drinks (beer, wine, etc.) and cannabis cigarettes and tobacco.

Results: Lacks in coping skills like to "Think of the negative consequences" diminishes the use of fermented drinks and cannabis a 24% and 40% respectively. The own skill "To be disciplined" diminishes the use of distilled drinks, fermented drinks and cannabis a 38%, 31% and 33% respectively. Lacks in social skills, as the difficulty to be related with people of the opposite sex, influences diminishing the use of distilled drinks (45%) and cannabis (70%), and increasing the use of tobacco fourth time.

Conclusions: Lacks in social skills are the only significant influence in the use of tobacco. Lacks in social and own abilities have an effect to the use of distilled drinks and cannabis, and lacks in coping skills facilitate a subject to use fermented drinks. The modification of all these lacks could guide to develop effective prevention programs.

Key words: *Coping Skills, Poisson, Overdispersion, Cannabis, Tobacco, Alcohol.*

INTRODUCCIÓN

En el pasado uno de los primeros acercamientos teóricos a las conductas adictivas, fue el modelo de enfermedad, que enfatizaba la importancia de los parámetros biológicos de las adicciones haciendo hincapié en los efectos farmacológicos de las sustancias. Esto implicaba que la persona adicta no tenía control voluntario sobre su conducta adictiva, ya que dicha conducta estaba determinada por factores fisiológicos internos como compulsiones, impulsos o deseos irresistibles a consumir (Marlatt, 1985). Una alternativa al modelo de enfermedad ha sido la Teoría del Aprendizaje Social de Bandura (1977) que explica la conducta humana como un fenómeno de aprendizaje basándose en las leyes del condicionamiento clásico, operante y vicario. Este modelo y los derivados del mismo, enfatizan la capacidad de autocontrol y/o autorregulación que el sujeto tiene sobre su conducta, derivándose tratamientos que implican el entrenamiento en estrategias de afrontamiento.

Dentro de las teorías cognitivo-conductuales sobre los procesos adictivos, el Modelo de Afrontamiento al Estrés (Wills y Shiffman, 1985; Wills y Hirky, 1996) asume que el consumo de sustancias podría ser una respuesta del sujeto a los estresores vitales a los que se enfrenta a lo largo de su vida, el consumo de sustancias reduciría los efectos negativos del estrés o aumentaría la capacidad del sujeto para hacer frente a los mismos. Siguiendo este modelo se puede deducir que si el sujeto tiene habilidades adecuadas y adaptativas para afrontar el estrés tendrá menos tendencia a desarrollar trastornos adictivos que si carece de estas habilidades.

Wagner, Myers y Ininch, (1999) en un estudio para validar empíricamente el modelo de afrontamiento al estrés, comprobaron que los adolescentes que utilizan estrategias centradas en el afrontamiento a las situaciones de estrés consumían menos que los que utilizaban estrategias de evitación, similares resultados se han encontrado en población adulta (Finney y Moos, 1995). Estos estudios apoyan la idea de que las estrategias de afrontamiento al estrés predicen el uso de drogas en adolescentes y adultos. Por otro lado, el modelo de Prevención de Recaídas de Marlatt y Gordon (1985) hace énfasis en la importancia de las estrategias de afrontamiento en el uso de drogas. Este modelo señala que después de un periodo de abstinencia se puede producir la vuelta al consumo cuando el sujeto tiene que enfrentarse a situaciones de alto riesgo y no utiliza las estrategias de afrontamiento adecuadas (Marlatt y Gordon, 1980). Este modelo subraya la importancia del entrenamiento en estrategias de afrontamiento tanto conductuales como cognitivas (Chaney, O`Leary y Marlatt, 1978). Estas estrategias de afrontamiento consisten básicamente en prever las posibles situaciones de alto riesgo futuras, el entrenamiento en resolución de problemas (D´Zurilla y Goldfried, 1971), en dar respuestas asertivas en dichas situaciones (Flowers, 1975) y en estrategias de manejo del estrés (Lazarus y Folkman, 1984). Por su parte el Modelo Transteórico del Cambio (Prochaska y DiClemente, 1982; Prochaska, DiClemente y Norcross, 1992; Prochaska y Prochaska, 1993) describe y delimita las variables y elementos que los sujetos ponen en práctica de forma intencional en el proceso de cambio de una conducta adictiva. Propone que en el proceso de cambio de una conducta adictiva se dan tres dimensiones: estadios, procesos y niveles de cambio. Los estadios hacen referencia a diferentes etapas motivacionales, los procesos son conductas observables o no que los sujetos realizan encaminadas a modificar su comportamiento adictivo y los niveles son cambios necesarios que se tienen que producir para abandonar una conducta adictiva. Estos autores cuando hablan de procesos se están refiriendo a las diferentes estrategias de afrontamiento que tienen que poner en práctica los sujetos en los diferentes estadios de cambio. Hay diferentes estudios que demuestran la influencia de las habilidades sociales en el consumo de drogas. Se ha detectado que los adolescentes que utilizan más habilidades asertivas (decir "no") y habilidades referidas a toma de decisiones consumen menos drogas ilegales (Barkin, Smith y DuRant, 2002). Hay estudios que apoyan el entrenamiento de habilidades sociales para el tratamiento de diferentes adicciones (Chaney y otros, 1978, Chaney, Roszell y Cummings, 1982; Marlatt y Gordon, 1980; Monti, Abrams, Binkoff y Zwick, 1986), por lo que se parte de que las personas que abusan de sustancias tienen déficit de habilidades sociales que pueden ser de dos tipos: Secundarios, que hacen referencia a falta de asertividad en situaciones relacionadas con el consumo, o primarios, problemas de habilidades sociales generales en diversas situaciones (Trower, Yardley, Briant y Shaw, 1978). Todas estas teorías y estudios muestran la importancia de las estrategias de afrontamiento, tanto conductuales como cognitivas, en el consumo de sustancias adictivas.

El objetivo del presente estudio es analizar, en una muestra de jóvenes adultos, el valor explicativo de las estrategias de afrontamiento cognitivo y conductuales, entre ellas las habilidades sociales especificas a la situación de consumo y las generales a situaciones interpersonales. Concretamente se pretende averiguar el peso de cada una de las variables en la explicación de la conducta de consumo de diferentes drogas, como son el consumo de bebidas destiladas, fermentadas, tabaco y cannabis. La variable a explicar es la variable de respuesta número de bebidas, número de cigarrillos y número de cigarrillos de THC consumidos en una semana. Como las variables de respuesta son variables de recuento se aplicará para su análisis, en

el contexto del Modelo Lineal Generalizado, el modelo de regresión de Poisson.

MÉTODO

La muestra estaba formada por 314 sujetos de edades comprendidas entre los 18 y los 30 años, con una edad media de 22.7 años. Por sexos, el 53.9% eran hombres y el 46.1% eran mujeres. La muestra estaba formada por 291 universitarios y 23 no universitarios. Se utilizó un muestreo incidental llevado a cabo en varios edificios del Campus Universitario correspondientes a diferentes estudios, así como en zonas de marcha de la ciudad.

En cuanto al consumo que presentaban, 22.1% eran No consumidores (Si no habían probado nunca ninguna de las sustancias), el 12.1% eran experimentadores o ex-consumidores (habían probado en una o dos ocasiones o habían consumido alguna droga en otro tiempo, pero actualmente no consumían ninguna droga), el 69.4% eran consumidores ocasionales de alguna de las sustancias y el 55.4% eran consumidores habituales (fin de semana, entre semana y a diario) de alguna de las sustancias analizadas. Se les administró un cuestionario anónimo desarrollado para medir la frecuencia de uso de diferentes sustancias adictivas, así como las habilidades propias, habilidades personales y estrategias de afrontamiento.

Los cuestionarios fueron administrados por un entrevistador, que preguntaba personalmente a los sujetos de forma individual y recogía las respuestas dadas por estos.

Las variables explicativas introducidas en el modelo fueron: a) Estrategias de afrontamiento, consideradas como habilidades específicas puestas en práctica por el sujeto para evitar el consumo. b) Habilidades propias, recogidas como percepción que tiene el sujeto de sus habilidades generales frente a cualquier situación problemática. c) Habilidades sociales, percepción que tiene el sujeto de sus habilidades de interrelación social (ver tabla1).

Todas estas variables eran variables categóricas binarias para evitar las respuestas de tendencia intermedia y obligar a contestar en algún sentido.

Las variables de respuesta modeladas fueron el número de copas de bebidas destiladas (ginebra, ron, whisky, combinados, etc.) consumidas en una semana, el número de copas de bebidas fermentadas (vino, cerveza, cava, etc.) consumidas en una semana, el número de cigarrillos consumidos en una semana y el número de cigarrillos de THC consumidos en una semana.

Se planteó un diseño de carácter explicativo, el objetivo del cual es comprobar los factores que per-

TABLA 1: Items del cuestionario.

Estrategias de Afrontamiento	Habilidades Propias	Habilidades Sociales
1-Decir que no quiero cuando me ofrecen.	1-Tiene facilidad para encontrar soluciones cuando tiene un problema	1-Tienes dificultades para conocer gente nueva
2-Evitar lugares y personas relacionados con drogas.	2-Se considera una persona disciplinada.	2-Tienes dificultades para expresar tus sentimientos.
3-Pensar en las consecuencias negativas que acarrea el consumo.	3-Sabe distraerse cuando quiere dejar de pensar en algo.	3-Te cuesta iniciar, mantener o cerrar una conversación con personas que no conoces.
4-Simplemente no llevar dinero.	4-Sabe relajarse sin utilizar drogas.	4-Tienes problemas para relacionarse con personas del sexo opuesto.
		5-Tienes dificultades para decir "no" a otras personas.
		6-Te resulta muy difícil negarte cuando te ofrecen implicarte en conductas de consumo de sustancias.

miten predecir las variables de respuesta número de copas consumidas de bebidas fermentadas y destiladas, número de cigarrillos y número de cigarrillos de THC consumidos en una semana. Las bebidas se han convertido a UBES (Unidades de bebidas estándar) para realizar los análisis.

El modelo estadístico utilizado para analizar los datos se enmarca dentro del Modelo Lineal Generalizado, concretamente se aplicará el modelo de regresión de Poisson como modelo adecuado para datos de recuento. El análisis se realizó con el Stata 8.0.

RESULTADOS

Se ha utilizado la prueba de hipótesis basada en la regresión (Cameron y Trivedi, 1998) para evaluar el supuesto de equidispersión del modelo de regresión de Poisson, obteniéndose que en bebidas destiladas (t=2,82 ; P=0.006), en tabaco (t=7,96 ; P<0.001) y en cannabis (t=4,49 ; P<0.001) se incumple el supuesto, mientras que éste se cumple en bebidas fermentadas (t=1,46 ; P=0.148), por lo que en los tres primeros casos se ha utilizado el error estándar robusto.

Aplicamos el modelo a las diferentes muestras de consumidores y las variables que aparecen como explicativas para cada una de las variables de respuesta son las que se presentan a continuación.

En bebidas fermentadas encontramos que los sujetos que dicen utilizar la estrategia de afrontamiento "Evito lugares y personas relacionados con drogas" consumen un 46% menos respecto a los que no utilizan esta estrategia (b=-0.617, p=0.003), siendo el número esperado de Ubes para los que utilizan esta estrategia de 2.8 Ubes menos a la semana. Por otro lado, los que dicen que usan la estrategia de afrontamiento "No llevar dinero", multiplican por 6 el consumo de estas sustancias (b=1.81, p=0.001), presentando un valor esperado de 29 Ubes a la semana más que los que no utilizan esta estrategia de afrontamiento.

Aquellos que piensan en las consecuencias negativas que acarrea el consumo disminuyen un 24% su consumo respecto a los que no utilizan esta estrategia de afrontamiento (b=-0.271, p=0.016), lo que conlleva disminuir 1.5 Ubes a la semana su valor esperado en Ubes consumidas.

Las personas que dicen ser disciplinadas consumen un 31% menos que las que señalan que no son disciplinadas (b=-0.365, p=0.003), esperándose un consumo de 2.5 Ubes menos a la semana. Si hablamos de los que señalan dificultades para negarse cuando les ofrecen implicarse en conductas de consumo de sustancias, vemos como consumen casi el triple, esperando un consumo de 10.2 Ubes más a la semana (b=1.031, p<0.001).

Los que señalan facilidad para solucionar problemas consumen un 31% más de bebidas fermentadas (b=0.270, p=0.005), 1.46 Ubes más a la semana, que los que dicen que no les resulta fácil solucionar problemas.

En bebidas destiladas los sujetos que se consideran disciplinados consumen un 38% menos que los que señalan no ser disciplinados, concretamente el cambio esperado en el número de Ubes es de 5 unidades menos en los sujetos disciplinados respecto a los no disciplinados (b=-0.474, p=0.004). Los que señalan tener dificultades para conocer gente nueva consumen un 69% más, es decir el valor esperado es 5.7 Ubes más a la semana respecto a los que señalan no tener dificultades para conocer gente nueva (b=-0.527, p=0.022). Los sujetos que dicen tener problemas para relacionarse con las personas del sexo opuesto consumen un 45% menos, 4 Ubes menos a la semana, que los que dicen no tener problemas de relación con personas del sexo opuesto (b=-0.592, p=0.006).

Hay una tendencia a la significación (p=0.059) en aquellos que manifiestan tener problemas para decir "no", que origina una disminución del 20% en el consumo con una bajada esperada de 1.8 Ubes en el consumo semanal.

En tabaco, los sujetos que señalan tener dificultad para expresar sus sentimientos consumen un 42% más, concretamente presentan un valor esperado de 32.1 cigarrillos más a la semana que los que no presentan dificultades de expresión de sentimientos (b=0.351, p=0.003). Los sujetos que dicen no tener habilidades conversacionales consumen un 56% menos (47.7 cigarrillos menos a la semana) que los que presentan esta habilidad (b=-0.820, p=0.001) . Por otro lado los sujetos que señalan problemas para relacionarse con el sexo opuesto consumen 4 veces más que los que no señalan dicho problema (b=1.398, p<0.001).

Hay una tendencia a la significación (p=0.07) en la estrategia de no llevar dinero, los cuales multiplican por dos la prevalencia de consumo de tabaco.

En cannabis, los sujetos que señalan saber relajarse sin drogas consumen un 78% menos (18.2 cigarrillos de THC menos a la semana) que los que señalan necesitar las drogas para relajarse (p=0.057). Los que dicen tener problemas para relacionarse con las personas del sexo opuesto consumen un 70% menos respecto a los que no presentan problemas de relación con el sexo opuesto (b=-1.194, p=0.006). Se espera una disminución de 18.2 cigarrillos de THC en los sujetos que señalan saber relajarse sin drogas y una disminución de 4 cigarrillos de THC en aquellos que señalan tener

problemas de relación son el sexo opuesto. También en esta sustancia aparece una variable con tendencia a la significación (p=0.057), que muestra que el uso de la estrategia de afrontamiento "pensar en las consecuencias negativas del consumo" disminuye el consumo en un 40%, traduciéndose esto en un valor esperado de 2.5 cigarrillos de THC menos en los que piensan en las consecuencias negativas del consumo. Una segunda variable con tendencia a la significación (p=0.061) corresponde a la habilidad propia de sentirse una persona disciplinada, lo que disminuye el consumo en un 33%, con un número esperado de 2.7 cigarrillos de THC menos que los que no poseen dicha habilidad.

Cada uno de los resultados expuestos se interpretan siempre suponiendo que se mantienen constante las otras variables del modelo.

DISCUSIÓN

En nuestra muestra hemos encontrado, en la misma dirección que Barkin y otros (2002), que en bebidas fermentadas las estrategias efectivas son las referentes a habilidades asertivas en situaciones concretas de consumo. La estrategia de afrontamiento de control de estímulos "no llevar dinero" no es efectiva, posiblemente por el bajo coste de estas bebidas, mientras que si parece funcionar en nuestra muestra, tal como señalan Prochaska y DiClemente (1982, 1984), la estrategia de control de estímulos "Evitar personas y lugares relacionados con las drogas". Por otro lado la estrategia cognitiva "Pensar en las consecuencias negativas" disminuye el consumo de alcohol justificándose así el uso que se hace de esta estrategia dentro de la terapia cognitiva de las drogodependencias (Beck, Wright, Newman y Liese, 1993).

Como también encontraron Todd, Kashdan, Charlene, Vetter y Collins (en prensa), entre otros, la autodisciplina está relacionada con un menor uso de alcohol.

En la Teoría de la Activación, Eysenck (1973, 1980 y 1981) habla del uso del tabaco como reductor de sentimientos de ansiedad e ira en determinadas situaciones tensas para el sujeto. En la muestra analizada se ha obtenido que la expresión de sentimientos, y los problemas de relación con el sexo opuesto influyen positivamente en el uso de tabaco, aumentando el consumo. Esto podría apoyar la teoría de Eysenck, ya que estos déficits en habilidades sociales podrían ser causantes de estrés y por tanto la utilización de tabaco podría servir para eliminar los sentimientos de ansiedad y/o hacer frente a estas situaciones sociales.

Por otro lado, en la muestra se ha encontrado que los sujetos con déficits en habilidades conversacionales consumen menos tabaco. Esto podría estar relacionado con los resultados del estudio de Eysenck (1980) que hallaron que los fumadores frente a los no fumadores eran más extravertidos y uno de los aspectos que recogían en el constructo de extraversión era la sociabilidad. Un amplio estudio sobre las variables de personalidad en diferentes usuarios de tabaco y derivados puede verse en Spielberger, Reheiser, Foreyt, Poston y Volding (2004).

Podemos concluir, no obstante, que déficits en la expresión de sentimientos y en las relaciones con el sexo opuesto, pueden ser variables de mantenimiento de un mayor uso de tabaco y déficits en habilidades sociales conversacionales parece disminuir el consumo de tabaco, según los resultados obtenidos.

El consumo de cannabis aumenta en los sujetos que utilizan esta sustancia para relajarse, esto puede estar relacionado con el efecto relajante que se produce tras el consumo de cannabis (Leza y Lorenzo, 2000). El tener problemas para relacionarse con el sexo opuesto, no obstante, parece disminuir el consumo de esta sustancia.

Según estos resultados los consumidores de THC tienen menos problemas de interrelación con el sexo opuesto, por lo que puede ser posible que el consumo de THC se utilice para facilitar las relaciones con el sexo opuesto.

Como conclusión podemos señalar que las estrategias de afrontamiento generales están mayoritariamente relacionadas con el consumo de tabaco, cannabis y bebidas destiladas y únicamente en bebidas fermentadas parecen influir las estrategias de afrontamiento especificas al consumo. Así, características de los sujetos no directamente relacionadas con el consumo de drogas parecen estar influyendo en el consumo que los sujetos hacen de las bebidas destiladas, tabaco y cannabis, mientras que déficits en estrategias de afrontamiento especificas a la situación de consumo son los que llevarán a un sujeto a consumir bebidas fermentadas.

Otros estudios han encontrado diferencias en función del sexo, estatus, etc. Estos elementos, claramente, son inmutables y poco se puede hacer para modificarlos. No obstante, nuestra investigación si que indica cómo elementos alterables, como las habilidades sociales o determinadas estrategias de afrontamiento, parecen marcar la diferencia en determinar el uso que se hace de las diferentes sustancias.

Por ello, los programas de tratamiento y prevención tienen que ir encaminados a incorporar el aprendizaje de estrategias de afrontamiento generales y específicas, no solo teniendo en cuenta las variables individuales y contextuales de los sujetos, sino también el tipo de sustancia.

REFERENCIAS

Bandura, A. (1977): *Social learning* theory. Englewood Cliff, N.J: Prentice-Hall.

Barkin, S.L, Smith, K.S y DuRant, R.H. (2002). Social skills and attitudes associated with substance use behaviors among young adolescents. *Journal of Adolescent Health, 30.* 448–454

Beck, A., Wright, F., Newman, C. y Liese, B. (1993). *Cognitive Therapy of substance abuse.* Guildford Press. New York.

Cameron, A. C. y Trivedi, P. K. (1998). *Regression analysis of count data.* Cambridge University Press.

Chaney, E.F., O`Leary, M.R. y Marlatt, G.A. (1978). Skill training with alcoholics. *Journal of Consulting and Clinical Psychology, 46,* 1092-1104.

Chaney, E.F., Roszell, D.K. y Cummings, C. (1982). Relapsed in opiate addicts: A behavioral analysis. *Addictive Behaviors, 7,* 291-297.

D´Zurilla, T.J. y Goldfried, M.R. (1971). Problem solving and behavior modification. *Journal of Abnormal Psychology,78,* 107-126.

Eysenck, H.J. (1973). Personality and the maintenance of the smoking habit. In W. L. Bunn (Ed.), *Smoking behavior: motives and incentives*. Washington, DC: Winston/Wiley.

Eysenck, H.J. (1980). *The causes and efects of smoking.* London: Temple Smith.

Eysenck, H.J. (1981). *A model for personality*. New York: Springer.

Finney, J.W. y Moos, R.H. (1995). Entering treatment for alcohol abuse: A stress and coping model. *Addiction, 90*, 1223–1240.

Flowers, J.V. (1975). Simulation and role playing methods. In F. H. Kanfer y A. P. Goldstein (Eds.), *Helping people change : A textbook of methods.* New York: Pergamon.

Lazarus, R.S. y Folkman, S. (1984). Coping and adaptation. In W. D. Gentry (Ed.). *Handbook of behavioral medicine.* New York: Guilford.

Leza, J.C. y Lorenzo, P.(2000). Efectos farmacológicos de los Cannabinoides. *Adicciones, vol.12, sup.2,* 109-133.

Marlatt, G.A. (1985). Relapse prevention: Theorical rationale and overview of the model. En G.A. Marlatt y J.R. Gordon (eds), *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors.* New York: Guilford

Marlatt, G.A. y Gordon, J.R. (1980). Determinants of relapse: Implications for the maintenance of behavior change. En P.O. Davidson y S.M. Davidson (eds.), *Behavioral medicine: Changing health life styles,* Nueva York, Brunner/Mazel.

Marlatt, G.A. y Gordon, J.R. (1985). *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors.* New York: Guilford

Monti, P.M., Abrams, D.B., Binkoff, J.A. y Zwick, W.R. (1986). Social skills training and substance abuse. En C.R. Hollin y P. Tower (eds.), *Handbook of socials skills training,* Nueva York, Pergamon Press.

Prochaska, J. y Prochaska, J.M. (1993). Modelo transteorético de cambio para conductas adictivas. En Casas y M. Gossop (eds.), *Recaída y prevención de recaídas. Tratamientos Psicológicos en Drogodependencias,* Barcelona, Ed. Neurociencias, Citrán.FISP.

Prochaska, J.O. y DiClemente, C.C. (1982). Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy: Theory, Research and Practice, 19*, 276-288.

Prochaska, J. O. y DiClemente, C. C. (1984). Stages and processes of self-change of smoking: Toward and integrative model of change. *Journal of Consulting and Clinical Psychology, 51,* 390-395.

Prochaska, J.O., DiClemente, C.C. y Norcross, J.C. (1992). In search of how people change. Applications to Addictive Behaviors. *American Psychologist, 47*, 1102-1114

Spielberger, C.D., Reheiser, E.C., Foreyt, J.P., Poston, W.S.C. y Volding, D.C. (2004). Personality determinants of the use of tobacco products. *Personality and Individual Differences, 36*, 1073-1082.

Todd, B., Kashdan, Charlene, J., Vetter, R. y Collins, L. (2004). Substance use in young adults:Associations with personality and gender . *Addictive Behaviors (*en prensa*).*

Trower, P., Yardley, K., Briant, G.M. y Shaw, P. (1978). The treatment of social failure. *Behavior Modification, 2,* 41-60.

Wagner, E.F., Myers, M.G. y Ininch, J.L. (1999). Stress-coping and temptation-coping as predictors of adolescent substance use. *Addictive Behaviors, 24* (6), 769–779,

Wills, T.A. y Hirky, A.E. (1996). Coping and substance abuse: A theoretical model and review of the evidence. In M. Zeichnec & N. S. Eudler (Eds.), *Handbook of coping: Theory Research and Applications,* 279–302. New York: Wiley.

Wills, T.A. y Shiffman, S. (1985). Coping and substance use: A conceptual framework. In S. Shiffman & T. A. Wills (Eds.), *Coping and Substance Use ,* 3-24. San Diego, CA: Academic Press.

Activity levels and drug use in a sample of Spanish adolescent.

Artículo publicado en Addictive Behaviors, 2005. Reproducción autorizada.

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Addictive Behaviors xx (2005) xxx–xxx

ADDICTIVE BEHAVIORS

Short communication

Activity levels and drug use in a sample of Spanish adolescents

Noelia L. Aleixandre*, Miguel J. Perello del Río, Alfonso L. Palmer Pol

Balearic Islands University, Department of Psychology, Ctra. Valldemossa, km 7.5, 07122 Palma de Mallorca, Spain

Abstract

The importance of adolescents taking part in various activities as a protective factor in substance consumption has been demonstrated over time. The objective of this paper is to analyze a sample of 1378 adolescents and the explanatory capacity of participating in different activities in relation to present-day alcohol, cannabis and tobacco consumption. Count variables were used as response variables, thus the Poisson Regression Model was applied in the analysis, within the context of a Generalized Linear Model. The results of the research demonstrate the explanatory value and differences in explanatory roles of each activity in connection with each the substances studied.

Keywords: Substance use; Substance-free behaviour; Adolescent; Alcohol; Tobacco; Cannabis

1. Introduction

According to the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV-TR, 2002), two of the criteria for substance dependency are related to dependent subjects' activity levels. It appears clear that substance dependent and abusive subjects decrease their educational, work-related, domestic or recreational activities, relegating them to second place in favour of activities related to drug consumption, as the latter end up acquiring a greater reinforcing power. From the functional analysis of behaviour, this decline in non-drug consumption-related activities explains why substances positively or negatively reinforce behaviours associated with their consumption and the external or internal stimulus situations associated with them.

* Corresponding author. Methodology of Behavior and Social Sciences Area, Department of Psychology, Guillem Cifre de Colonya Building (Room A-215), Balearic Islands University, Ctra. de Valldemossa, km. 7'5, 07122 Palma de Mallorca (SPAIN). Tel.: +0034 971 172495; fax: +0034 971 173190.

E-mail address: noelia.llorens@uib.es (N.L. Aleixandre).

0306-4603/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.addbeh.2005.02.009

One study conducted with 206 subjects with an average age of 18.89 who had consumed alcohol or other drugs in the preceding 30 days showed results that demonstrated that the reinforcement that drug-free activities provide subjects has a predictive value on drug consumption behaviours. It found a negative relationship between the reinforcing level of drug-free activities and drug consumption, to such a degree that frequency of consumption increases if the amount of reinforcement of the drug-free activity perceived by the subject decreases (Correia, Benson, & Carey, 2005; Correia, Simons, Carey, & Borsari, 1998).

However, on the other hand, when dealing with adolescents, and drug consumption behaviour is seen within a wide range of health risk behaviours, it was found that activity level is not as decisive. It seems that adolescents involved in multiple health risk behaviours were also involved in many other positive behaviours (Lindberg, Boggess, & Williams, 1999). Behavioural theories assume that drug or alcohol consumption should be understood within the context in which that behaviour occurs.

The objective of this study was to verify the relation between participating in pleasurable or reinforcing activities and the consumption of different addictive substances. It specifically aimed to prove the explanatory capacity of participating in different activities in present day substance consumption.

2. Method

The sample was made up of 1378 subjects of which 43.7% were men and 56.3% were women. Ages oscillated between 13 and 19 years of age and the average age of the subjects was 15.67 years of age. Of the 1378, 50.2% consumed fermented beverages, 52.3% consumed distilled beverages, 44.6% consumed cannabis cigarettes and 48.4% consumed cigarettes. In all the cases we are with non-problematic sample. The adolescents were asked for their informed consent verbally and were informed that their responses were confidential. A differentiation was made in the alcohol variable between fermented beverages (wine, beer...) and distilled beverages (whisky, vodka...) and their different associated graduations were taken into account. The measures were all subsequently converted to standard beverage units (SBU, when one SBU equals 10 g of ethanol) in order to carry out the analysis.

The Generalized Linear Model was used as a reference framework to analyze data whose response variable is a count and consequently the model to apply is the Poisson Regression. This model requires compliance with equidispersion. The analysis was conducted with the Stata 8.0 and SPSS 11.0.

3. Results

3.1. Descriptive

Subjects who have tried drugs participate in a low total number of activities (0–2), in a proportion statistical significantly higher than those who have not tried them ($\chi^2 = 10.69$; $p < .001$). Subjects' responses showed that those who have not tried drugs visit libraries and sports centers more often, while subjects who have tried drugs presented statistically higher values in visiting pubs and discotheques ($\chi^2 = 33.057$; $p < .001$). In our sample, there were no statistically significant differences in subjects' gender and having tried the different substances or not.

3.2. Modelling

One way to verify equidispersion is through the regression test proposed by Cameron and Trivedi (1990). After verifying the non-compliance of equidispersion and executing various possible fits, the model, which presents the best fit in our data, is the Negative Binomial Regression Model.

As the study works with count data, interpretation cannot be conducted directly on the model's coefficients, due to the lack of linearity; thus the transformation to IRR (Incidence Rate Ratio) through the exponentiation of the coefficients, exp(b), was used as the interpretation of this transformation is similar to the odds ratios, when the remaining variables are always maintained constant.

Table 1, shows the variables for all the substances, which appear significant. Social activities and the age of first use appear as explanatory variables in the consumption of fermented beverage. Subjects who claimed to participate customarily in social activities consume 54% more beverages than those who claimed not to participate in these activities. Furthermore, consumption decreases 5% for each year the age of first year increases.

Social activities, families and trips appear significant in distilled beverages. Subjects who indicated participating in social activities consume 34% more than those would do not participate in this type of activity, producing an expected change of 1.8 more drinks per week compared to those who do not customarily participate in these activities. Subjects who indicated participating in family activities consume 35% less than those who did not participate in these activities, producing an expected change in the number of drinks of 2.5 fewer standard drinks per week. Those who indicated taking trips increased their consumption of distilled beverages by 36% compared to those who do not participate in these

Table 1
Variables for all the substances, which appear significant

Number of:	Coef.	SE	*z*	P>*z*	[95% conf. interval]	
Fermented Beverages						
Age first use	−.0516	.0253	−2.03	.042	−.1014	−.0019
Social activities	.4327	.1283	3.37	.001	.1811	.6842
Distilled Beverages						
Age first use	−.0894	.0260	−3.43	.001	−.1405	.0384
Cultural activities	−.2156	.1177	−1.83	.067	−.4464	.0151
Social activities	.2965	.0968	3.06	.002	.1067	.4863
Family activities	−.4300	.0920	−4.67	<.001	−.6104	.2496
Trips	.3100	.0983	3.15	.002	.1171	.5028
Cannabis						
Age first use	−.3114	.0577	−5.39	<.001	−.4246	−.1981
Cultural activities	−.6358	.2372	−2.68	.007	−1.100	−.1708
Trips	.5053	.1961	2.58	.010	.1210	.8897
Tobacco						
Sports	−.8708	.2011	−4.33	<.001	−1.265	−.4766
Family activities	−.3415	.2007	−1.70	.089	−.7350	.0520

The response variables utilised in the research were the amounts consumed in the last week.

activities. The age of first use also appears to influence consumption of these substances, specifically a decrease of 9% for each year first use increases. Furthermore, a trend in the significance of cultural activities can also be indicated. Subjects who participated in cultural activities present an expected change in the number of distilled drinks of 1.18 fewer standard drinks than subjects who do not participate in this type of activity.

Cultural activities, trips and the age of first use appear as explanatory variables in cannabis consumption. Subjects who say that they customarily participate in cultural activities consume 48% less than subjects who do not participate in these activities (2.24 fewer cannabis cigarettes per week). Participating in activities related to trips increases consumption by 65% compared to those who do not participate in this type of activity (an expected change in consumption of 2.36 more cannabis cigarettes per week). As for the age of first use, consumption decreases by 27% for each year first use increases.

Only sports activities appear significant in the consumption of tobacco. Subjects who indicated participating in sports activities consume 59% less tobacco than subjects who do not participate in sports activities, which translates into an expected value of 20 fewer cigarettes per week. The remaining activities in our sample do not appear as explanatory in tobacco consumption.

4. Discussion

The results of this study are consistent with prior research indicating the need for alternative activities to substance consumption (Carroll, 1996; Vuchinich & Tucker, 1988). In general, subjects who consumed one of the different substances participate in fewer drug-free activities (Van Etten, Higgins, Budney, & Badger, 1998).

The age of first use influences current consumption of different substances. In all cases it appears that later first use of substance consumption leads to lower consumption and as Carroll (1996), among others, indicates, subjects who participate in more non-drug related activities begin consumption later. Activities not related to drugs are thus converted into a protective factor, because later first use of a drug means a delay associated with the consumption of other, potentially more dangerous drugs (Kandel & Yamaguchi, 1985; Yu & Williford, 1992). Nevertheless, it can be verified how the influence of this variable on each one of the substances varies differentially, as later first use means a greater decrease in the consumption of distilled beverages than in fermented beverages, but cannabis is where the greatest influence as regards a greater decrease in consumption can be observed.

Higher levels of activities are not always related to a lower consumption of addictive substances; participating in social activities appears related to greater substance consumption, specifically alcohol (distilled and fermented) and the weight of this variable in fermented beverages is much greater than in distilled beverages. On the other hand, taking trips is also predictive of greater cannabis and distilled beverage consumption and the influence of this variable is greater on cannabis.

Although the variables introduced in the model, which explain the consumption of each one of the substances, are different, all of them appear in the same sense. As previously indicated, behavioural theories assume that drug consumption must be understood within the context it takes place in. Within the context, the availability of alternative reinforces incompatible with consumption and an environmental restriction to obtain substances is determinant (Vuchinich & Tucker, 1996). Thus, the increase in proven protective activities, as well as the proposal of non-drug use related activities must be taken into account in preventing and treating problems related to addictive substances. Providing young people

with easy access to alternatives activities to consuming drugs, activities they can participate in frequently, allows the age of first use in consumption to be delayed as well as consumption to be reduced. This research was partially supported by the National Plan on Drugs, Spain.

References

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, *46*, 347–364.

Carroll, M. E. (1996). Reducing drug abuse by enriching the environment with alternative non-drug reinforcers. In L. Green, & J. Kagel (Eds.), *Advances in Behavioral Economics*, *vol. 3* (pp. 37–68). Norwood, NJ: Ablex Press.

Correia, C. J., Benson, T. A., & Carey, K. B. (2005). Decreased substance use following increases in alternative behaviors: A preliminary investigation. *Addictive Behaviors*, *30*, 19–27.

Correia, C. J., Simons, J., Carey, K. B., & Borsari, B. E. (1998). Predicting drug use: Application of behavioral theories of choice. *Addictive Behaviors*, *23*, 705–709.

DSM-IV-TR. (2002). *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*. Rockville, MD: American Psychiatric Association.

Kandel, D. B., & Yamaguchi, K. (1985). Developmental patterns of the use of legal, illegal, and prescribed drugs. In C. L. Jones, & R. J. Battjes (Eds.), *Etiology of Drug Abuse: Implications for Prevention, 1985*, *vol. 56* (pp. 193–235). Rockville, MD: National Institute on Drug Abuse.

Lindberg, L. D., Boggess, S., & Williams, S. (1999). Multiple threats: The co-occurrence of teen health risk-behaviors. *Trends in the Wellbeing of America's Children and Youth* (pp. 489–504). Washington, DC: U.S. Department of Health and Human Services.

Van Etten, M. L., Higgins, S. T., Budney, A. J., & Badger, G. J. (1998). Comparison of the frequency and enjoy ability of pleasant events in cocaine abusers vs non-abusers using a standardized behavioral inventory. *Addiction*, *93*, 1669–1680.

Vuchinich, R. E., & Tucker, J. A. (1988). Contributions from behavioral theories of choice to an analysis of alcohol abuse. *Journal of Abnormal Psychology*, *97*, 181–195.

Vuchinich, R. E., & Tucker, J. A. (1996). The molar context of alcohol abuse. In L. Green, & J. Kagel (Eds.), *Advances in Behavioral Economics*, *vol. 3* (pp. 133–162). Norwood, NJ: Ablex Press.

Yu, J., & Williford, W. (1992). The analysis of drug use progression of young adults in New York State. *The International Journal of the Addictions*, *27*, 1313–1323.

Características de personalidad en adolescentes como predictores de la conducta de consumo de sustancias psicoactivas.

Artículo publicado en Trastornos Adictivos 2005; 7(2):90-6. Reproducción autorizada por Ediciones Doyma S.L.

ÁREA CLÍNICA

Características de personalidad en adolescentes como predictores de la conducta de consumo de sustancias psicoactivas

Characteristic of personality in adolescents as predictors of psicoactives drugs use

LLORENS ALEIXANDRE, N.*, PALMER POL, A.** Y PERELLÓ DEL RÍO, M.***

**Profesor Ayudante. Área de Metodología. Universidad de las Islas Baleares. Palma de Mallorca.*
***Profesor Titular. Área de Metodología. Universidad de las Islas Baleares. Palma de Mallorca. España.*
****Psicólogo práctica privada. Centro de Aplicaciones Psicológicas. Valencia. España.*

RESUMEN: ***Objetivo:*** **Examinar el papel explicativo que tienen diferentes indicadores de personalidad y el autoconcepto en el consumo de diferentes sustancias adictivas (alcohol, cannabis y tabaco), comprobando también el carácter predictivo que presentan estas variables en el consumo futuro de cada una de las sustancias analizadas.**

Sujetos y métodos: **1.378 sujetos con edades comprendidas entre los 13 y los 19 años contestaron un cuestionario anónimo desarrollado para medir la cantidad de sustancia consumida en una semana, indicadores de personalidad, de autoconcepto, así como variables personales. Las variables de respuesta analizadas fueron las cantidades consumidas de cada sustancia en una semana. El análisis se realizó a través del modelo de regresión de Poisson con ceros aumentados.**

Resultados: **A mayor edad, mayor consumo o mayor probabilidad de pasar a ser consumidor. Las mujeres consumen más tabaco que los varones. La impulsividad predice un mayor consumo de tabaco en sujetos consumidores. En consumidores de bebidas destiladas y de cannabis, un buen autoconcepto protege del consumo. La conducta antisocial predice un paso al consumo en sujetos no consumidores de todas las sustancias y un aumento de consumo en consumidores de todas las sustancias, excepto de bebidas destiladas. La búsqueda de sensaciones predice un consumo mayor en sujetos consumidores de bebidas destiladas, fermentadas y cannabis.**

Conclusiones: **No aparece una personalidad adictiva, pero hay indicadores de personalidad que parecen explicar mejor o predecir el consumo futuro de los adolescentes, sin embargo estos indicadores influyen de forma diferente en función de la sustancia.**

PALABRAS CLAVE: Abuso de sustancias. Adolescentes. Personalidad. Autoconcepto. Distribución de Poisson.

ABSTRACT: ***Objective:*** **This study aims to analyze the explanatory value of different indicators of personality and the self-concept over the use of different addictive substances (alcohol, cannabis and tobacco), also checking the predictive nature these variables show in the future consumption of every analyzed substance.**

Subjects and methods: **1378 subjects, between 13 and 19 years, answered an anonymous questionnaire developed to measure the quantity of substance consumed in one week, indicators of personality, self-concept, as well as personal variables. The answer variables analyzed were the consumed**

Este trabajo se ha realizado, en parte, gracias a la ayuda del Plan Nacional Sobre Drogas.

Correspondencia:
NOELIA LLORENS ALEIXANDRE
Edificio Guillem Cifre de Colonya
Universitat de les Illes Balears
Ctra. Valldemossa, km 7,5
E-mail: noelia.llorens@uib.es

quantities of each substance in one week. The analysis was carried out through Poisson's model of regression with zeros inflated.

Results: **The older, the greater use or bigger probability of becoming consumer. Women consume more tobacco than men. The impulsiveness predicts a greater use of tobacco in subject consumers. In subject consumers of distilled drinks and of cannabis a good self-concept protects from consumption. The antisocial behavior predicts a step to the consumption in subjects' non-consumers of all the substances and a consumption increase in consumers of all the substances, except for distilled drinks. The search of sensations predicts a greater consumption in subject consumers of distilled, fermented drinks and cannabis.**

Conclusions: **It doesn't appear an addictive personality, but there are personality indicators that seems to explain better or to predict adolescents future use, however these indicators influence in different ways in function of the substance.**

KEY WORDS: Abuse of substances. Adolescent. Personality. Self-concept. Poisson's distribution.

Introducción

Sin duda alguna, el fenómeno de las drogodependencias responde a una gran diversidad de variables que requieren un análisis detallado y preciso, para conocer de forma objetiva este problema y poder desarrollar procedimientos de prevención y tratamiento eficaces. Sin olvidar la existencia de factores biológicos, psicológicos (diferentes a la personalidad), ambientales y socioculturales, el presente estudio se ha centrado en los indicadores de personalidad que pueden estar influyendo en la predisposición para que los jóvenes consuman diferentes sustancias adictivas o pasen a consumir en el futuro.

El estudio de las características de personalidad ha sido muy controvertido, y se ha estudiado en algunas sustancias, pero no aparecen diferencias claras entre consumidores y no consumidores. Entre los diferentes factores de riesgo en el consumo de sustancias adictivas en adolescentes, podemos encontrar la impulsividad[1,2], la conducta antisocial[3,4], la búsqueda de sensaciones[5] y el autoconcepto[6].

Algunos estudios han encontrado que sujetos que en la niñez son hiperactivos, con falta de atención, problemas de conducta, impulsividad y fracaso escolar, con una historia familiar de alcoholismo, cuando son adultos se relacionan con el abuso de alcohol y otras drogas[7-9].

Otras investigaciones han hallado que los sujetos drogodependientes realizan un gran número de conductas antisociales y tienen un pobre concepto de sí mismos o baja autoestima[6]. Aunque estas variables se pueden entender como consecuencias del consumo de drogas, en un estudio longitudinal[4] se demostró que jóvenes que no habían probado ninguna sustancia pero que mostraban más signos de comportamientos antisociales, tenían más probabilidad de consumir diferentes sustancias cuando tenían 17 años. Estrechamente relacionada con la conducta antisocial, se encuentra la búsqueda de sensaciones, que hace referencia a la necesidad de obtener experiencias variadas y nuevas y a aceptar riesgos físicos y sociales[5].

Las investigaciones de sujetos que con el tiempo han desarrollado problemas con el alcohol, no presentaban perfiles de personalidad desajustados antes de desarrollar el problema. Aunque sí que tenían unas características comunes entre ellos y que les diferenciaban de los sujetos que no desarrollaban problemas de drogodependencias, estas características consistían en ser personas gregarias, inconformistas e impulsivas[10,11].

El presente estudio ha tenido en cuenta los indicadores de personalidad más relevantes que se han hallado en las investigaciones dentro del área de personalidad asociadas a las conductas adictivas, tales como indicadores de conducta antisocial, de impulsividad, de búsqueda de sensaciones y autoconcepto, así como variables personales (edad y sexo). Para medir los indicadores de personalidad se realizó un cuestionario basado y adaptado en instrumentos específicos para medir autoconcepto, indicadores de impulsividad, de búsqueda de sensaciones y de conducta antisocial en adolescentes.

Sujetos y métodos

Muestra

La muestra estaba formada por 1.378 estudiantes de escuelas públicas, obtenidas aleatoriamente, de la ciudad de Palma de Mallorca, con edades comprendidas entre los 13 y los 19 años. A los adolescentes, se les pidió consentimiento informado de forma verbal. El cuestionario fue autoadministrado en sus respectivos centros escolares, durante los meses de abril a junio de 2004.

Instrumento

Se les administró un cuestionario anónimo desarrollado para medir la cantidad consumida de diferentes sustancias adictivas: alcohol, tabaco y cannabis, así como variables personales, autoconcepto e indicadores de personalidad. El cuestionario se adaptó para recoger la cantidad de consumo, que se midió preguntando la cantidad que habían consumido de cada una de las sustancias en la última semana.

Por otro lado, los sujetos debían señalar de 19 afirmaciones sobre su forma de pensar, sentir y actuar, cuáles de ellas se adaptaban mejor a ellos mismos. A continuación se enuncian las afirmaciones utilizadas, con sus indicadores, en los sujetos utilizados en el estudio.

A. Autoconcepto:
- *1.* Hago muchas cosas bien.
- *2.* Estoy satisfecho conmigo mismo.
- *3.* Me gusta como soy físicamente.
- *4.* Creo que tengo buenas cualidades.
- *5.* Soy un fracaso como persona.

B. Impulsividad:
- *6.* Hago cosas impulsivamente.
- *7.* Me resulta difícil estarme quieto.
- *8.* Digo las cosas sin pensarlas.
- *9.* Me resulta difícil esperarme cuando quiero conseguir algo.
- *10.* Soy una persona que normalmente se precipita.

C. Búsqueda de sensaciones:
- *11.* Me gusta practicar deportes y actividades de riesgo.
- *12.* Me gusta tener experiencias nuevas y excitantes.
- *13.* Me gusta hacer cosas que impliquen peligro.
- *14.* Me gusta el desenfreno y la desinhibición.

D. Conducta antisocial:
- *15.* Soy una persona que alborota y monta jaleo.
- *16.* Suelo hacer cosas prohibidas e ilegales.
- *17.* Rompo, quemo o deterioro cosas de otros.
- *18.* Suelo pelearme o insultar a otros.
- *19.* Contesto mal a las personas mayores (padres, profesores, etc.).

Todos se evaluaron de forma dicotómica, los sujetos debían señalar si se identificaban con estas afirmaciones o no.

Se planteó un diseño de carácter explicativo y predictivo, el objetivo del cual era comprobar los factores que permiten explicar la variable de respuesta cantidad consumida de diferentes sustancias en una semana, así como predecir el consumo futuro en función de las variables analizadas. El consumo de bebidas alcohólicas se ha expresado en UBE (unidades de bebidas estándar) para realizar los análisis.

Análisis estadísticos

Debido a las características de la variable respuesta, que es una variable de recuento, el análisis adecuado a aplicar es el modelo de regresión de Poisson, este modelo se enmarca en el modelo lineal generalizado[12]. El modelo de regresión de Poisson exige el cumplimiento del supuesto de equidispersión. La ausencia de equidispersión provoca la aparición de sobredispersión, lo que lleva a que, estando bien especificado el modelo, las estimaciones de los parámetros son correctas pero no así sus errores estándar (SE). Otro problema de la distribución de Poisson es el exceso de ceros que, en este caso, es debido a la existencia de 2 posibles tipos de ceros. En primer lugar, los debidos a personas que nunca consumen y, por otra parte, los que pertenecen a los sujetos que aunque sí consumen la sustancia en cuestión, no la han consumido en la última semana. Para solucionar este problema, utilizamos el análisis bajo el modelo de Poisson con ceros aumentados (ZIP).

También se realizó un estudio descriptivo de los datos. El análisis se efectuó con el Stata 8.0 y el SPSS 11.0.

La adecuación de los indicadores utilizados en este estudio se comprobó a través de un análisis factorial confirmatorio, en el que se obtuvo que el conjunto de los 19 indicadores se ajustaban a 4 factores: autoconcepto, búsqueda de sensaciones, impulsividad y conducta antisocial.

Resultados

Estadísticos descriptivos

El 43,7% son varones y el 56,3% mujeres. Edades comprendidas entre los 13 y los 19 años. Con una mayor presencia de sujetos que acuden a 3.º de ESO. El 37,7% de los sujetos entrevistados habían consumido al menos una copa de bebidas destiladas (whisky, vodka, etc.) en la última semana, el 26% lo habían hecho de bebidas fermentadas (vino, cerveza, etc.), el 17,6% había consumido cigarrillos de cannabis y el 24,5% cigarrillos de tabaco. Se les preguntó por el consumo de otras drogas, tales como éxtasis, LSD y anfetaminas pero al aparecer una frecuencia de uso muy baja, alrededor del 1%, se omitió su análisis.

Modelado

Una forma de comprobar la equidispersión es a través de la discrepancia del modelo. Para evaluar la existencia de sobredispersión se utiliza la prueba de la regresión propuesta por Cameron y Trivedi (1990)[13] cuyos valores se dan entre paréntesis. En las bebidas fermentadas el valor de este cociente es 2,47 ($z = 3,51$; $p < 0,001$), en las bebidas destiladas el grado de dispersión es de 6,69 ($z = 5,30$; $p < 0,001$), en el cannabis el grado de dispersión es de 4,36 ($z = 3,07$; $p < 0,05$) y en el tabaco es de 28,59 ($z = 3,07$; $p < 0,002$). Estos resultados comprueban el incumplimiento del supuesto de equidispersión.

Se comprueba la adecuación del ZIP, respecto al modelo de Poisson, por medio de la prueba de Voung. Los resultados de esta prueba son: en bebidas fermentadas $z = 6,52$ ($p < 0,001$); en bebidas destiladas $z = 12,59$ ($p < 0,001$); en cannabis $z = 7,28$ ($p < 0,001$) y en tabaco $z = 10,03$ ($p < 0,001$).

En las tablas se muestran las variables que han aparecido como significativas en la explicación del consumo de cada una de las sustancias. Como estamos trabajando con datos de recuento, la interpretación no se puede hacer directamente sobre los coeficientes del modelo por falta de linealidad, por ello usamos la transformación mediante la exponenciación de los coeficientes exp(b), siendo la interpretación de esta transformación similar a las *odds ratio*, siempre manteniendo constante el resto de variables.

Si vemos lo que ocurre en función de las sustancias podemos señalar que:

A. En consumidores:

En el consumo de bebidas fermentadas (tabla 1), aparece como variable explicativa el hecho de estar satisfecho consigo mismo, que hace que los sujetos que lo están consuman un 46,4% menos respecto de los que no lo están. Los sujetos que señalan que les gusta el desenfreno y la desinhibición consumen un 71% más, mientras que los sujetos que hacen cosas prohibidas e ilegales y que rompen, queman o deterioran cosas de otros, indicadores de conducta antisocial, presentan un consumo de bebidas fermentadas del 93,4 y 89,9% más a la semana, respectivamente, que los que no señalan estos ítems. La edad se asocia con un incremento del 16,8% de consumo por año de aumento.

En bebidas destiladas (tabla 2), los sujetos que dicen que les gusta hacer cosas que impliquen peligro consumen un 30,1% más de este tipo de bebidas; sin embargo, los sujetos que dicen que contestan mal a las personas mayores consumen un 24,9% menos.

En cannabis (tabla 3), los sujetos que dicen que les gusta practicar deportes y actividades de riesgo consumen un 89% más y los que señalan que suelen hacer cosas prohibidas e ilegales consumen 3 veces más. La edad de nuevo predice consumo, ya que por cada año de aumento en la edad se consume un 38,3% más.

En el consumo de tabaco (tabla 4), los sujetos que señalan que les resulta difícil estarse quietos consumen un 63,1% más. Los que dicen que hacen cosas prohibidas e ilegales consumen un 87,1% más y los que señalan que suelen pelearse e insultar a otros consumen el doble. En cuanto a la edad, por cada año que aumenta la edad, el consumo aumenta un 18%, mientras que respecto al sexo, el ser mujer predice un consumo de un 59,4% más que en varones.

B. En no consumidores:

El modelo aplicado, ZIP, permite analizar el papel explicativo y predictivo de las variables en aquellos sujetos no consumidores. Así, podemos comprobar

Tabla 1. Consumo de bebidas fermentadas, medido en UBE, en una semana

Bebidas fermentadas	Exp(b)	p > z	IC del 95%	
Consumidores				
Edad	0,1554569	0,023	1,021987	1,335313
Perso2	–0,6296173	0,018	0,3163904	0,8972182
Perso14	0,5364313	0,011	1,129153	2,589318
Perso16	0,6597447	0,015	1,135921	3,293811
Perso17	0,641172	0,041	1,027176	3,509702
No consumidores				
Edad	–0,2145215	0,002	–0,3521274	–0,0769155
Perso2	–0,5627946	0,023	–1,047413	–0,0781765
Perso6	–0,4669422	0,032	–0,8948768	–0,0390077
Perso17	–0,8028918	0,031	–1,531086	–0,0746979

IC: intervalo de confianza; exp(b): exponenciación de los coeficientes.

Tabla 2. Consumo de bebidas destiladas, medido en UBE, en una semana

Bebidas destiladas	Exp (b)	p >z	IC del 95%	
Consumidores				
Perso13	0,2630345	0,011	1,060932	1,595075
Perso19	–0,2857554	0,035	0,5763589	0,9797223
No consumidores				
Edad	–0,5143184	0,000	–0,6210203	–0,4076166
Perso3	0,3545247	0,022	0,0521556	0,6568939
Perso4	0,4753441	0,023	0,0651328	0,8855553
Perso11	–0,3474365	0,031	–0,6624757	–0,0323973
Perso12	–0,7923896	0,000	–1,152911	–0,4318681
Perso14	–0,6387928	0,000	–0,9816917	–0,295894
Perso15	–0,3794528	0,028	–0,7169316	–0,0419741

IC: intervalo de confianza; exp(b): exponenciación de los coeficientes.

Tabla 3. Consumo de cigarrillos de cannabis en una semana

Cannabis	Exp(b)	p > z	IC del 95%	
Consumidores				
Edad	0,3195136	0,000	1,198994	1,580189
Perso11	0,6367947	0,001	1,283932	2,78337
Perso16	1,379888	0,000	2,78505	5,671822
No consumidores				
Perso4	0,7813845	0,004	0,2498487	1,31292
Perso8	0,4885174	0,028	0,0536962	0,9233386
Perso12	–1,376608	0,000	–2,01516	–0,7380561
Perso15	–0,4804475	0,037	–0,9317174	–0,0291776
Perso16	–1,437324	0,000	–1,91944	–0,955208

IC: intervalo de confianza; exp(b): exponenciación de los coeficientes.

Tabla 4. Consumo de cigarrillos de tabaco en una semana

Tabaco	Exp (b)	p > z	IC del 95%	
Consumidores				
Edad	0,1735059	0,011	1,041129	1,358942
Sexo	0,4660112	0,034	1,03476	2,454326
Perso7	0,4892801	0,025	1,064085	2,500385
Perso16	0,626552	0,015	1,131444	3,094446
Perso18	0,7038184	0,006	1,226733	3,331032
No consumidores				
Edad	–0,1686438	0,003	–0,2801577	–0,0571299
Sexo	–0,7458229	0,000	–1,099974	–0,3916715
Perso12	–0,8844517	0,000	–1,312343	–0,4565603
Perso14	–0,3724147	0,045	–0,7362063	–0,0086231
Perso16	–0,9809596	0,000	–1,42762	–0,5342997

IC: intervalo de confianza; exp(b): exponenciación de los coeficientes.

que: en bebidas fermentadas (tabla 1), los sujetos no consumidores que señalan los indicadores: «estoy satisfecho conmigo mismo», «hago cosas impulsivamente» y «rompo, quemo o deterioro cosas de los otros» tendrán menor probabilidad de seguir siendo no consumidores.

En bebidas destiladas (tabla 2), según aumenta la edad disminuye la probabilidad de seguir siendo no consumidor. Los sujetos no consumidores que dicen gustarse físicamente y no considerarse un fracaso, presentan mayor probabilidad de seguir siendo no consumidores. Por otra parte, los que señalan que les

gusta practicar deportes y actividades de riesgo, tener experiencias nuevas, el desenfreno y la desinhibición y dicen ser personas que alborotan y montan jaleo, tendrán menor probabilidad de seguir siendo no consumidores.

En cannabis (tabla 3), señalar los indicadores: «creo que tengo buenas cualidades» y «digo las cosas sin pensarlas» aumentan la probabilidad de seguir siendo no consumidor. Por otra parte, los sujetos que apuntan que les gusta tener experiencias nuevas, que se consideran personas que alborotan y montan jaleo, y que dicen hacer cosas prohibidas e ilegales, tendrán menor probabilidad de seguir siendo no consumidores.

En tabaco (tabla 4), a medida que aumenta la edad aumenta la probabilidad de pasar a ser consumidor, y ser mujer aumentará la probabilidad de ser consumidor. Los sujetos que dicen que les gusta tener experiencias nuevas, el desenfreno y la desinhibición y que suelen hacer cosas prohibidas e ilegales, tendrán menor probabilidad de seguir siendo no consumidores.

Discusión

Los indicadores de personalidad y de autoconcepto, tanto en consumidores como no consumidores, son una valiosa información de cara al diseño de planes de prevención y tratamiento. Estos indicadores pueden explicar el consumo realizado de las diferentes sustancias y predecir el riesgo de consumo por parte de aquellos sujetos actualmente no consumidores.

El análisis global de los resultados permite mostrar cómo los sujetos que no consumen, pero señalan indicadores de búsqueda de sensaciones y de conducta antisocial[4], tienen una predisposición al consumo y una alta probabilidad de pasar a ser consumidores. Se puede observar, a raíz de los resultados, como la conducta antisocial y la búsqueda de sensaciones parecen ir siempre unidas en la conducta de consumo. Así, mientras que la búsqueda de sensaciones es lo que hace que los sujetos pasen de no consumidores a consumidores, la conducta antisocial es la que hace que se dé un mayor consumo cuando ya son consumidores. El autoconcepto, en general, no parece influir en los consumidores, pero sí que aparece como importante en los no consumidores, asociándose un autoconcepto positivo con una protección del consumo de bebidas destiladas y cannabis. La impulsividad[1,2,10], por su parte, influye en el paso al consumo de no consumidores de alcohol y aumenta el consumo de tabaco en consumidores, no obstante, ni explica ni predice el consumo del resto de sustancias.

En la presente investigación, podemos señalar que se han encontrado diferencias en función del sexo en el consumo de tabaco, y esta variable aparece como explicativa de la cantidad de sustancia consumida en una semana. La edad aparece como explicativa, junto con indicadores de personalidad, en las diferentes sustancias analizadas, tanto en los consumidores como en los no consumidores, observándose un incremento en las cantidades consumidas a medida que aumenta la edad, en todas las sustancias.

No aparece una personalidad adictiva, pero sí que hay indicadores de personalidad que parecen explicar mejor o predecir el consumo futuro de los adolescentes y estos indicadores influyen de forma diferente en función de la sustancia.

Es importante remarcar la importancia de la aplicación de los análisis adecuados a los datos a analizar. Así, se puede señalar que los análisis aplicados a los datos de la presente investigación permiten unir en un mismo análisis a todos los sujetos independientemente de la sustancia que consumen y de la cantidad consumida. Si se observan las investigaciones en sustancias adictivas, encontramos que hay muchos problemas en la clasificación del tipo de consumo que hacen las personas, obligándose en muchas ocasiones a hacer una descripción previa a las investigaciones de lo que se considera un consumidor y un no consumidor, o bien, agrupando la información para poder aplicar análisis. Si nos basamos en la cantidad consumida en una semana de cada una de las sustancias, sin tener en cuenta la etiqueta a la que se puede unir ese sujeto, esto nos permitirá trabajar con toda la información de la que se dispone, sin perder esta información por agrupaciones bajo etiquetas creadas por el investigador.

En la presente investigación, los análisis muestran que ciertos indicadores de personalidad, de autoconcepto, así como variables personales influyen de forma diferente en los consumidores y en los no consumidores, y esta distinción también se da en función de las sustancias.

Muchos estudios intentan comprobar las diferencias en función de la edad, del sexo y de la etnia[14-16], pero pocos estudios se enfocan para ver las diferencias en función de las sustancias. Si es importante, como se ha demostrado, tener en cuenta las características personales de los sujetos, igual o más importante es tener en cuenta las sustancias a las que está expuesto. Los resultados obtenidos en este estudio abren una puerta al futuro en cuanto permiten ver la probabilidad de consumo futuro de los sujetos en función de sus características personales, mostrando que incluso los indicadores de personalidad van a influir de forma dife-

rente en las diferentes sustancias y en el tipo de consumo que se realice.

Como conclusión, señalar que los planes de prevención se deberían enfocar en función de las sustancias y en función del tipo de consumo de la población diana. Los resultados obtenidos en este trabajo aportan evidencias de que los indicadores de personalidad y autoconcepto nos están dando información de los adolescentes y su posible consumo futuro. Por tanto, los tratamientos enfocados al abandono del consumo, o a la prevención del inicio de la conducta adictiva, deberían tenerlos en cuenta para ser más efectivos y eficaces.

Bibliografía

1. Hayaki J, Stein MD, Lassor JA, Herman DS, Anderson BJ. Adversity among drug users: relationship to impulsivity. Drug Alcohol Depend. En prensa 2004.
2. Allen TJ, Moeller FG, Rohades HM, Cherek DR. Impulsivity and history of drug dependence. Drug Alcohol Depend. 1998;50:137-45.
3. Hawkins JD, Catalano RF, Millar JY. Risk and protective factors for alcohol and other drug problems in adolescenceand early adulthood: implications for substance abuse prevention. Psychol Bull. 1992;112:64-105.
4. Adalbjarnardottir S, Rafnsson FD. Adolescent antisocial behavior and substance use longitudinal analyses. Addict Behav. 2002;27:227-40.
5. Zuckerman M. Sensation seeking and psychopathology. En: Hare RD, Schalling D, editors. Psychopathic behavior: approaches to research. New York: John Wiley; 1978.
6. Zucker RA, Gomberg ESL. Etiology of alcoholism reconside ed: the case for a biopsychological process. Am Psychol. 1986;41:783-93.
7. Alterman AI, Tarter RE. Examination of selected typologies: hyperactivity, familial, and antisocial alcoholism. En: Galanter M, editor. Recent developments in alcoholism. Vol. 4. New York, NY: Plenum Press; 1986. p. 169-89.
8. Donovan JE, Jessor R, Jessor L. Problem drinking in adolescence and young adults. J Stud Alcohol. 1983;44:109–37.
9. Knop J. Risk factors in alcoholism. Lancet. 1985;2:387–8.
10. Cox WM. Identifying and measuring alcoholic personality characteristics. San Francisco: Jossey-Bass; 1983.
11. Cox WM. Personality correlates of substance abuse. En: Galicio M, Maisto SA, editors. Determinants of substance abuse: biological, psychological and environmental factors. Nueva York: Plenum Press; 1985. p. 209-46.
12. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman & Hall; 1989.
13. Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the Poisson model. J Econometrics. 1990;46:347-64.
14. Sprujt-Metz D, Gallear PE, Unger JB, et al. Meanings of smoking and adolescent smoking across ethnicities. J Adolesc Health. 2004;35:197-205.
15. Sprujt-Metz D. Adolescent, affect and health. London: Psychology Press; 1999.
16. Mermelstein R. Explanations of ethnic and gender differences in youth smoking: a multi-site, qualitative investigation. Nic Tob Res. 1999;1:S91-8.

IV.- REFERENCIAS BIBLIOGRÁFICAS

Aitchison, J. y Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika, 76*(4), 643-653.

Andersen, B. (n.d.). Regression III. Advanced Methods. Recuperado el 10 de Febrero de 2005 desde http://socserv.mcmaster.ca/andersen

Anderson, D.R., Burnham K.P. y White G.C. (1994). AIC model selection in overdispersed capture-recapture data. *Ecology 75(6)*:1780-1793.

Anscombe, F.J. (1953). Contribution to teh discusión of H. Hotelling's paper. Journal of the Royal Statistical Society. B(15), 229-230.

Ato, M. y López, J.J. (1996). *Análisis estadístico para datos categóricos.* Madrid: Síntesis.

Ato, M., Losilla, J.M., Navarro, J.B., Palmer, A.L. y Rodrigo, M.F. (2000a). *Del contraste de hipótesis al modelado estadístico.* Terrassa: CBS.

Ato, M., Losilla, J.M., Navarro, J.B., Palmer, A.L. y Rodrigo, M.F. (2000b). *Modelo lineal generalizado.* Terrassa: CBS.

Bae, S., Famoye, F., Wulu, J.T., Bartolucci, A.A. y Singh, K.P. (2005). A rich family of generalized Poisson regression models with applications. Mathematics and Computers in Simulation.

Baringhaus, L. y Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistics and Probability Letters,* 13, 269-274.

Belsley, D.A., Kuh, E. y Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* New York: Wiley.

Bickel, P.J. y Freedman, D.A. (1981). Some asymptotic Theory for the bootstrap. *Annals of Statístics,9*, 1196-1217.

Böhning, D. (1994). A note on a test for Poisson overdispersion. *Biometrika*, 81, 418-419.

Box, G.E.P. y Cox, D.R. (1964). An analisis of transformations. *Journal of the Royal Statistical Society-Series B, 26,* 211-246.

Box, G.E.P., Hunter, W.G, y Hunter, J.S. (1988). *Estadística para investigadores. Introducción al diseño de Experimentos, Análisis de Datos*

y Construcción de Modelos (Trad.). Barcelona: Reverté. (Traducción del original, 1978)

Brännäs, K. y Rosenqvist, G. (1994). Semiparametric estimation of hetrogeneous count data models. *European Journal of Operational Research, 76*, 247-258.

Breslow, N. (1990). Test of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association, 85*(410), 565-571.

Breslow, N. (1996). Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata, 8*, 23-41.

Cameron, A.C. y Trivedi, P.K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests [Versión electrónica]. *Journal of Applied Econometrics, 1*, 29-53.

Cameron, A.C. y Trivedi, P.K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cameron, A.C. y Trivedi, P.K. (1998). Regression Analysis of Count Data. *Econometric Society Monographs*, *30*. Cambridge: Cambridge University Press.

Campbell, D.B. y Oprian, C.A. (1979). On the Kolmorogov-Smirnov test for the Poisson distribution with unknown mean*. Biometrical Journal, 21*, 17-24.

Cochran, W. (1954). Some methods of strengthening the X^2 goodness of fit test. *Biometrics, 10,* 417-451.

Cohen, A.C. (1960). Estimation in a Truncated Poisson Distribution When Zeros and Ones are missing. *Journal of the American Statistical Association, 55*, 342-348.

Cohen, A.C. (1963). Estimation in Mixtures of Discrete Distribution In Proceedings of the International Symposium on Discrete Distributions, Montreal, Quebec.

Consul, P.C. (1989). *Generalized Poisson distributions.* New York: Marcel Dekker.

Consul, P.C. y Famoye, F. (1992). Generalized Poisson regression model. *Communications in Statistics - Theory and Methods, 21*, 89-109.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15-18.

Cox, D. R. y Snell, E. J. (1989). *Analysis of Binary Data*. London: Chapman and Hall.

Cox, D.R. y Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B, 30*, 248-265.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Análisis, 45*, 215-233.

Cribari-Neto, F. y Zarkos, S. (2001). Heteroskedasticity-Consistent Covariance Matrix Estimation: White's Estimator and the Bootstrap. *Journal of Statistical Computation and Simulation,68,* 391-411.

Cribari-Neto, y Zarkos, S. (1999). Bootstrap Methods for Heteroskedastic Regression Models: Evidence on estimation and testing. *Econometric Reviews,*18, 211-228.

Dalrymple, M.L., Hudson, I.L. y Ford, R.P.K. (2003). Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS. *Computational Statistics and Data Análisis, 41, 491-504*.

Davidson, R. y Mackinnon, J.G. (1993). *Estimation and Inference in Econometrics,* Oxford, Oxford University Press.

Davison, A.C. y Gigli, A. (1989). Deviance residuals and normal scores plots. *Biometrika, 76,* 211-221.

Dean, C.B., Eaves, D.M. y Martinez, C.J. (1995). A comment on the use of empirical covariance matrices in the analysis of count data *Journal of Statistical Planning and Inference, 48, 197-205.*

Dean, C.B., Lawless, J.F. y Willmot, G.E. (1989). A mixed Poisson-inverse Gaussian regression model. *Canadian Journal of Statistics, 17*(2), 171-181.

Deb, P. y Trivedi, P.K. (1997). Demand for medical care by the elderly in the United States: a finite mixture approach. *Journal of Applied Econometrics, 12,* 313-326.

Deb, P. y Trivedi, P.K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics, 21*, 601-625.

Dunn, P.K. y Smith, G.K. (1996). Randomized Quantile Residuals. Journal of Computational. and Graphical Statistics, 5(3), 236-244.

Eberhardt, L.L. (1978). Appraising variability in population studies. *Journal of Wildlife Management 42*, 207-238.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics,7*,1-26.

Efron, B. (1982a). *The jackknife, the bootstrap and another resampling plans*. CBMS Regional Conference Series in Applied Mathematics 38. Philadelphia: SIAM Publications.

Efron, B. y Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science,1* ,54-57.

Efron, B. y Tibshirani, R. (1993). *An introduction to the Bootstrap*. Nueva York: Chapman and Hall/CRC.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley , CA: University of California Press, pp. 59-82.

Fahrmeir, L. y Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* (2ª ed.). New York: Springer-Verlag.

Famoye, F. (1993). Restricted generalized Poisson regression. *Communications in Statistics - Theory and Methods, 22*, 1335-1354.

Famoye, F. y Wang, W. (2004). Censored generalized Poisson regression model. *Computational Statistics and Data Analysis,46(3),* 547-560.

Feller,W. (1943). On a general class of “Contagious” distributions. *Annals of Mathematical Statistics, 14,* 389-400.

Friendly, M. (2001). *Visualizing Categorical Data*. SAS Publishing.

Ganio, L.M. y Schafer, D.W. (1992). Diagnostics for overdispersion. *Journal of the American Statistical Association, 87*, 795-804.

Gardner, W., Mulvey, E. y Shaw, E. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392-404.

Gart, J. y Pettigrew, H. (1970). On the conditional Moments of the K-statistics for the Poisson distribution. *Biometrika,* 57, 661-664.

Gill, J. (2000). Generalized linear models: A unified approach. *Sage University Papers on Quantitative Applications in the Social Sciences*, *07-134*. Thousand Oaks, CA: Sage.

Gourieroux, C., Monfort, A. y Trognon, A. (1984a). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica, 52*, 701-720.

Gourieroux, C., Monfort, A. y Trognon, A. (1984b). Pseudo-maximum likelihood methods: theory. *Econometrica, 52*, 681-700.

Green, P. J. (1994). *Contribution to the discussion of paper by Grenander and Miller* (1994). Journal of the Royal Statistical Society B, 56, 589--590.

Greene, W. H. (2000). *Econometric Analysis.* (4ª ed.). New York: Prentice Hall.

Grogger, J.T. y Carson, R.T. (1991). Models for truncated counts. *Journal of Applied Econometrics, 6*, 225-238.

Gupta, A.K., Mori, T.F. Y Szekely, G.Z. (1994). Testing for Poissonity-normality vs other infinite divisibility. *Statistics and Probability Letters, 19*, 245-248.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business and Economic Statistics, 9*(2), 215-222.

Gurmu, S. y Trivedi, P.K. (1992). Overdispersion tests for truncated Poisson regression models. *Journal of Econometrics, 54*, 347-370.

Gurmu, S., Rilstone, P. y Stern, S. (1998). Semiparametric estimation of count regression models. *Journal of Econometrics, 88*(1), 123-150.

Halekoh, U. (2002). Residuals in Genelalized Linear Models. Recuperado el 25 de marzo de 2005 desde http://genetics.agrsci.dk/biometry/courses/statmaster/course.

Hamilton, L. (n.d.) Regression with graphics. Recuperado el 06 de Octubre 2004 desde http://www.ats.ucla.edu/stat/stata/examples/rwg/rwgstata5/rwgstata5.htm.

Hardin, J. y Hilbe, J. (2001). *Generalized Linear Models and Extensions.*College Station, TX: Stata Press.

Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in empirical bayes estimation [Versión electrónica]. *Accident Analysis and Prevention, 33*(6), 799-808.

Hausman, J., Hall, B.H. y Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship [Versión electrónica]. *Econometrica, 52*(4), 909-938.

Heilbron, D. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal, 36*, 531-547.

Henze, N. (1996). Empirical distribution function goodness of fit tests for discrete models. *Canadian Journal of Statistics*, 24, 81-93.

Henze, N. y Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Australian Journal of Statistics*, 38, 61-74.

Hinde, J. y Demétrio, C.G.B. (1998). Overdispersion: Models and estimation. *Computational Statistic and Data Analysis, 27,*151-170.

Hines, R.J.O. y Carter, E.M. (1993). Improved added variable and partial residual plots for detection of influential observations in generalized linear models. *Applied Statistics, 42*, 3-20.

Hinkley, D.V. (1977). Jackknifing in Unbalanced Situations. *Technometrics, 19,* 285-292.

Hoffmann, J.P. (2004). *Generalized Linear Models. An applied approach.* New York: Pearson Education.

Huber, P.J. (1967). The Behavior of Maximum Likelihood Estimates Under Non-standard Conditions. *In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley: University of California Press, pp.221-233.

Johnson, N.L. y Kotz, S. (1969). Discrete Distributions, Boston, Houghton Mifflin.

Johnson, N.L., Kotz, S. y Balakrishnan, N. (1994). *Continuous univariate distributions. Vol. I.* (2ª ed.). New York: John Wiley.

Karlis, D. y Xekalaki, E. (1998). Robust inference for finite Poisson mixtures. *Technical Report,* 49. Department of Statistics, Athens University of Economics and Business, Athens.

Karlis, D. y Xekalaki, E. (2000). A simulation comparison of several procedures for testing the Poisson assumption. *The Statistician,* 49, 355-382.

Kauermann, G. y Carroll, R.J. (2001). A note on the efficiency of Sandwich Covariance Matrix Estimation. *Journal of the American Statistical Association, 96*(456), 1387- 1400.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator [Versión electrónica]. *American Journal of Political Science, 33*(3), 762-784.

Klar, B. (1999). Goodness of fit tests for discrete models based on the integrated distribution function. *Metrika,* 49, 53-69.

Kleinbaum, D.G., Kupper, L.L. y Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariate Methods.* (2ª ed.). Belmont, CA: Duxbury Press.

Kocherlakota, S. y Kocherlakota, K. (1986). Goodness of fit for discrete distributions. *Communications in statistics-theory and methods*, 15, 815-829.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics, 34*, 1-14.

Lee, L.F. (1986). Specification test for Poisson regression models. *International Economic Review, 27*, 689-706.

Lee, S. (1998). Detecting departures from a Poisson model. *Communications in Statistics-Theory and Methods, 25*, 1201-1210.

Lee, Y. y Nelder, J.A. (2000). Two ways of modelling overdispersion in non-normal data. *Applied Statistics, 49*, 591-598.

Liang, K.Y. y Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,*73 (1), 13-22.

Liao, T.F. (1994). *Interpreting Probability Models. Logit, Probit and other Generalized Linear Models.* London:Sage

Lindsey, J. K. (1998). Counts and times to events. *Statistics in Medicine, 17*(15-16), 1745-1751.

Lindsey, J.K. (1995). *Introductory Statistics: A Modelling Approach.* New York: Oxford University Press.

Long, J.S y Ervin, L.H. (2000). Using heteroskedasticity consistent standard errors in the Linear Regression Model. Forthcoming, *The American Statistician, 54*, 217-224.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Long, J.S. y Ervin, L.H, (1998). Correcting for Heteroscedasticity with Heteroscedasticity Consistent Standard Errors in the Linear Regression Model: Small Sample Considerations. Working Paper.

Losilla, J.M. (2002). Computación intensiva para el Análisis de Datos en el siglo XXI. *Metodología de las ciencias del comportamiento*, 4(2), 201-221.

Lunneborg, C. E. (1994). *Modelling Experimental and Observational Data.* Belmont, CA: Duxbury Press.

Lunneborg, C. E. (1999). *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove, CA: Brooks-Cole.

Lunneborg, C.E. (1985). Estimating the correlation coefficient: the bootstrap approach. *Psychological Bulletin, 98*, 209-215.

MacKninnon, J.G. and White, H. (1985). Some Heteroskedasticity-consistent covariance matrix estimators with improvement finite samples properties. *Journal of Econometrics,29*, 305-325.

McCullagh, P. y Nelder, J.A. (1989). *Generalized linear models.* (2ª ed.). London: Chapman & Hall.

Melkersson, M. y Rooth, D.O. (2000). Modelling female fertility using inflated count data models [Versión electrónica]. *Journal of Population Economics, 13*, 189-203.

Miller, R.G. (1974). The Jackknife – a review. *Biometrika,* 61(1), 1-15.

Mullahy, J. (1986). Specificaton and testing of some modified count data models. *Journal of Econometrics, 33*, 341-365.

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models [Versión electrónica]. *Journal of Applied Econometrics, 12*, 337-350.

Nakamura, M. y Perez-Abreu, V. (1993). Use of an empirical probability generating function for testing a Poisson model. *Canadian Journal of Statistics*, 21, 149-156.

Nakashima, E. (1997). Some methods for estimation in a negative-binomial model [Versión electrónica]. *Annals of the Institute of Statistical Mathematics, 49*(1), 101-115.

Nass, C. (1959). The X^2 test for small expectations in contingency tables with special reference to accidents and absenteism. *Biometrika, 46,* 365-385.

Nelder, J. A. y Wedderburn, W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society-Series A, 135*, 370-384.

Osgood, D.W. (2000). Poisson-based regression analysis of aggregate crime rates [Versión electrónica]. *Journal of Quantitative Criminology, 16*(1), 21-43.

Palmer, A., Losilla, J.M., Llorens, N., Sesé, A., Montaño, J. J.,Jimenez, R. y Cajal, B. (2002). La sobredispersión en el Modelo Lineal Generalizado. Aplicación a datos de recuento. *Metodologia de las Ciencias del Comportamiento, volumen especial,* 433-436.

Pierce, D.A. y Schafer, D.W. (1986). Residuals in Generalized Models. *Journal of the American Statistical Association, 81*, 977-986.

Poisson, S.D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en materière civile, précédés des règles générales du calcul des probabilités*. Paris: Bachelier

Poortema, K. (1999). On modelling overdispersion of counts. *Statistica Neerlandica, 53*(1), 5-20.

Pregibon, D.(1981). Logistic regression diagnostics*. Annals of Statistics, 9,* 705-724.

Preisser, J.S. y Garcia, D.L. (2005). Alternative computational formulae for generalized linear model diagnostics: identifying influential observations with SAS software*. Computational Statistics and Data Analysis, 48*, 755-764.

Quenouille,M. H. (1949). Aproximate test of correlation in time series. *Journal of the Royal Statistical Society – Series B*, 11, 68-84.

Quenouille,M. H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

Rayner, J.C.W. y Best, D. J. (1988). Smooth goodness of fit tests for regular distributions. *Communications in statistics-theory and methods, 17,* 3235-3267.

Rayner, J.C.W. y McIntyre, R. (1985). Use of the score statistic for testing goodness of fit of some generalized distributions. *Biometrical Journal*, 27, 159-165.

Read, T. y Cressie, N. (1988). *Goodness of fit Statistics for Discrete Multivariate Data.* New York: Springer.

Rider, P. (1961a). The method of moments applied to a mixture of two exponential distributions. *Annals of Mathematical Statistics, 32,* 143–147.

Rider, P. (1961b). Estimating the parameters of mixed Poisson, Binomial and Weibull distributions by methods of moments. Bulletin de l'Institut International de Statistiques 38, part 2.

Rodríguez, G. (2002). *Lecture Notes on Generalized Linear Models*. Recuperado el 28 de mayo de 2002 desde http://data.princeton.edu/wws509/notes.

Rueda, R., Perez-Abreu, V. y O'Reilly, F. (1991). Goodness of fit for the Poisson distribution based on the probability generating function. *Communications in Statistics-Simulation and Computation, 28*, 259-274.

Saha, A. y Dong, D. (1997). Estimating nested count data models [Versión electrónica]. *Oxford Bulletin of Economics and Statistics, 59*(3), 423-430.

Shankar, V., Milton, J. y Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry [Versión electrónica]. *Accident Analysis and Prevention, 29*(6), 829-837.

Spinelli, J.J. y Stephens, M.A. (1997). Cramer Von-Mises test of fit for the Poisson distribution. *Canadian Journal of Statististics, 25*, 257-268.

Sturman, M.C. (1999). Multiple approaches to analysing count data in studies of individual differences: The propensity for Type I errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement, 59*(3), 414-430.

Svetliza, C.F. y Paula, G.P. (2001). On diagnostic in log-linear negative binomial models. *Journal of Statistical Computation and Simulation,71,* 231-244.

Tang, N.S., Wei, B.C. y Wang, X.R. (2000). Influence diagnostics in nonlinear reproductive dispersion models. *Statistics and Probability Letters, 46*, 59-68.

Titterington, D.M., Smith, A. F. M. y Markov, U. E. (1985). *Statistical Analysis of Finite Mixtures Distributions*. New York: Wiley.

Tukey, J.W. (1958). Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics, 29,* 614.

UCLA, Stata Learning Module on Regresión Introduction to Regression Diagnostics. Recuperado el 10 de Marzo de 2005 desde http://www.ats.ucla.edu/stat/stata/modules

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307-334.

Wang, p., Puterman, M.L., Cockburn, I. (1998). Analysis of patent data- a mixed Poisson-regression-model approach. *Journal of Business and Economic Statistics,16(1)*, 27-41.

Wang, W. y Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression [Versión electrónica]. *Journal of Population Economics, 10*, 273-283.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and Gauss-Newton methods. *Biometrika,*61(3), 439-447.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity, *Econometrica,* 48(4), 817-838.

Wilcox, R.R., y Muska, J. (1999). Tests of hypotheses about regression parameters when using a robust estimator. *Communications in Statistics Theory and Methods* 28, 2201–2212

Winkelmann, R. (2000). *Econometric Analysis of Count Data.* (3ª ed.). Berlin: Springer-Verlag.

Winkelmann, R. y Zimmermann, K.F. (1991). A new approach for modeling economic count data. *Economic Letters, 37*, 139-143.

Winkelmann, R. y Zimmermann, K.F. (1995). Recent developments in count data modelling: theory and application. *Journal of Economic Surveys, 9*(1), 1-24.

Wu, C.F.J. (1986). Jackknife, bootstrap and another resampling methods in regression analysis. *The Annals of. Statistics,* 14(4), 1261-1295.

Yen, S.T. (1999). Gaussian versus count-data hurdle models: cigarrette consumption by women in the US [Versión electrónica]. *Applied Economic Letters, 6*, 73-76.

Zeger, S.L. y Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics,*42, 121-130.

Zelterman, D. (1988). Homogeneity test against central-mixture alternatives. *Journal of American Statistical Association*, 83,179-182.

Zorn, C. (1996). Evaluating Zero-inflated and Hurdle Poisson Specifications. Working Paper. Departament of Political Science, Ohio State University, Columbus, OH.

III.- RESUMEN DE RESULTADOS Y CONCLUSIONES

Los principales resultados obtenidos de los diferentes trabajos que componen la tesis se pueden dividir en función de las líneas señaladas en los objetivos. Así podemos ver que:

El primer paso cuando nos encontramos con datos de recuento es la comprobación de los supuestos del modelo. Por ello y teniendo en cuenta la importancia de la sobredispersión en este tipo de datos planteamos el estudio de aquellas pruebas más utilizadas en la literatura para detectar este tipo de problema. Debido a la gran cantidad de pruebas utilizadas y el hecho de la aparición de discrepancias entre ellas a la hora de señalar la existencia de sobredispersión en los datos, consideramos la realización de un estudio de simulación, capaz de comprobar el comportamiento de las pruebas más utilizadas bajo diferentes condiciones de sobredispersión y tamaños de muestra.

En el estudio 1 “Overdispersion diagnostics in count data analysis models” (Vives, Losilla y Llorens, en revisión) nos propusimos el estudio comparativo de algunas de las pruebas más utilizadas en la literatura, para comprobar la existencia de sobredispersión. Para alcanzar el objetivo propuesto realizamos un estudio de simulación en el que se comprobaban la tasa de error y la potencia de cada una de las pruebas analizadas, en diferentes condiciones de sobredispersión y tamaño de muestra. Concretamente se realizaron dos estudios, en el primero de los cuales se comparó la tasa nominal de error de las diferentes pruebas. Los resultados apoyaron la utilización de la por presentar una aproximación más estable al error nominal fijado. El resto de pruebas mostraron consistencia en los diferentes grados de sobredispersión analizados, a excepción de la prueba de la Discrepancia, que presentó una gran divergencia respecto el valor α fijado. En el segundo estudio se comparó la potencia de las diferentes pruebas, la y la prueba de razón de verosimilitud mostraron una potencia superior en la mayoría de las situaciones analizadas.

Hardin y Hilbe (2001) señalan la utilización de la prueba de regresión de Cameron y Trivedi (1990) y la prueba de LM como posibles pruebas para el estudio de la sobredispersión. Debido a la fácil implementación de la prueba de regresión de Cameron y Trivedi (1990) en diferentes programas estadísticos, en los artículos sustantivos hemos aplicado esta prueba para la comprobación de la existencia de sobredispersión.

Posteriormente planteamos la necesidad de reunir en un mismo artículo todos aquellos índices y estimadores de los errores estándar utilizados para corregir el problema de la sobredispersión. Creamos un estudio de simulación en el que comparamos diferentes situaciones que podían inducir a resultados diferentes.

Los resultados indicaron que las estimaciones de los coeficientes en el MRP eran insesgadas, aunque las estimaciones de los errores estándar presentaban un sesgo hacia la infravaloración, bajo cualquier mecanismo generador de sobredispersión, influyendo no obstante el tamaño muestral y el grado de sobredispersión, que incidía directamente sobre el nivel de infraestimación de los EE de los parámetros del modelo de regresión.

En el estudio 2 "Ajuste y estimación de los errores estándar de los parámetros del modelo de regresión de Poisson en presencia de sobredispersión" (Llorens, Palmer y Losilla, 2004) se mostró que el ajuste que realiza HC1 sobre el HC0 mejora su comportamiento en muestras pequeñas. De los índices directos comparados, el índice $\sqrt{\chi^2 / gl}$ funcionó mejor en casi todas las condiciones, excepto en las condiciones de muestras grandes donde no aparecieron diferencias. En cuanto al Bootstrap y al Jackknife, cabe subrayar, que el funcionamiento de los dos estimadores fue prácticamente el mismo, observando una aproximación casi perfecta al EE, dando estimaciones eficientes. Solo podían señalarse pequeñas diferencias cuando se daban las condiciones más adversas, esto es, tamaño pequeño, media pequeña y grado de dispersión grande. En esta situación aunque los dos sobrestimaban el EE, el Bootstrap se acercaba más al valor real, dando valores menores del EE.

En general el Bootstrap fue el que presentó una mejor estimación del EE en casi todas las ocasiones, tanto en las condiciones adversas como en las más favorables. El HC1 funcionó correctamente a partir de medias moderadas, independientemente del grado de dispersión y del tamaño de la muestra. Con tamaño de muestra grande todos los métodos comparados funcionan de forma correcta, presentando todos ellos un buen ajuste al EE real. Debido a los resultados del artículo, consideramos de interés la ampliación de los estimadores incluidos en la investigación previa, así como el análisis específico de éstos. Para ello realizamos el estudio 3 "Overdispersion in the Poisson

regression model: A comparative simulation study" (Llorens, Palmer, Losilla y Vives, en revisión).

En esta investigación mostramos la importancia del tamaño de la muestra y del grado de dispersión de los datos, estimando el EE original así como corrigiéndolo. Cuando el valor del parámetro de escala aumentaba, la conducta media del grupo de índices era aproximadamente igual, manteniendo el valor de lambda y el tamaño. Un incremento en el tamaño de la muestra, disminuía el error estimado e incrementaba la precisión.

Dean, Eaves y Martinez (1995) y Breslow (1996), encontraron que el estimador sandwich infraestimaba la variabilidad real en tamaños de muestra pequeña y media. La literatura al respecto mostraba el comportamiento del HC3 (Davidson y Mackinnon, 1993; Long y Erving, 1998, 2000) y del HC4 (Cribari-Neto, 2004) como las mejores opciones al HC0. En nuestra investigación estos estimadores se adecuaban a las situaciones analizadas. No obstante, en todos los casos, los estimadores sandwich eran menos exactos que los estimadores de remuestreo y que el índice basado en el ji-cuadrado.

En cuanto a corregir y estimar los EE de los coeficientes del modelo de regresión de Poisson, los resultados indicaron que la estimación del remuestreo era menos parcial que los índices basados en la corrección directa del EE, básicamente el ji-cuadrado, o el estimador robusto; sin embargo, también se mostró que estos estimadores presentaban el RMSE más alto y por consiguiente menos preciso en casi todas las situaciones, sobre todo en tamaño de muestra pequeña. Así, nuestros resultados siguen las líneas sugeridas por Cameron y Trivedi (1998), quienes recomiendan el uso de las técnicas de remuestreo para las muestras pequeñas.

No obstante, creemos que deben realizarse varios bootstraps con las muestras pequeñas, analizar su estabilidad y así poder disminuir el efecto de su baja precisión; en este sentido, consideramos la necesidad de seguir realizando investigación para hacer bootstraps más precisos.

En muestras grandes, no aparecieron prácticamente diferencias entre los valores proporcionados por los diferentes índices y estimadores, exceptuando los índices directos cuya precisión supero al resto de estimadores. Concretamente en nuestro estudio el índice del ji-cuadrado presentó un RMSE un 65% más bajo que el del remuestreo con el mismo tamaño de muestra.

En general, los índices directos eran más precisos, seguidos de los estimadores sandwich y de las técnicas de remuestreo en tercer lugar; sin embargo, todos los índices mantuvieron un cierto sesgo comparados con el valor real, sobre todo en las muestras pequeñas y medianas.

El índice basado en el ji-cuadrado, pareció ser un buen corrector del EE, porque sus valores medios estaban cerca del valor real y porque era uno de los índices más preciso de todos los comparados.

Este segundo estudio nos permitió ser más exactos en las conclusiones a las que habíamos llegamos en el primer estudio. Así, en el primer estudio la conclusión a la que llegamos fue:

- En cuanto a los procedimientos de corrección y estimación del error estándar de los coeficientes de regresión de Poisson, claramente los resultados indicaron la superioridad de las estimaciones no paramétricas Bootstrap y Jackknife, sobre la corrección directa del error estándar infraestimado mediante su producto por la raíz de alguna forma de estimación del parámetro de dispersión.

Sin embargo después de haber realizado el segundo estudio podemos ser más precisos y añadir:

- En cuanto a los procedimientos de corrección y estimación del error estándar de los coeficientes de regresión de Poisson, los resultados indicaron que una buena opción cuando se ha de modificar el EE sería utilizar tanto el ji-cuadrado como las técnicas de remuestreo (el bootstrap y jackknife) y evaluar su coincidencia. Cuando coincidan, ambos podrían escogerse, pero con preferencia las técnicas de remuestreo. En el caso de que no coincidan recomendamos el uso del ji-cuadrado, porque su valor medio está más cercano al valor real.

El estudio 4 “Modelado del número de días de consumo de cannabis” (Palmer, Llorens y Perelló, en prensa) intentó mostrar la importancia de ajustar el modelo a los datos y no los datos al modelo. En Sturman (1999) se recogen algunos de los problemas que podemos encontrar al utilizar un procedimiento no adecuado a los datos con los que trabajamos.

Puesto que la variable número de días de consumo es una variable cuantitativa, tiende a ser modelada mediante el modelo de regresión lineal, un

hecho paradigmático de que el modelo de regresión lineal no es adecuado para datos de recuento es que hace predicciones negativas para una variable que, claramente, no admite valores negativos.

En general, en estas situaciones se acostumbra a realizar una transformación logarítmica en la variable respuesta, ya que esto proporciona una distribución cercana a la Normal, lo que posibilita su manejo mediante mínimos cuadrados ordinarios (MCO). El modelo de regresión con transformación logarítmica de la variable respuesta ajusta mejor que el modelo de regresión lineal ya que el índice AIC (Akaike Information Criterion) disminuye. Sin embargo, con la transformación logarítmica, existen por un lado problemas de estimación ya que un valor y=0, frecuente en una variable de recuento, necesita ser transformado para poder ser utilizado, en general sumándole una pequeña cantidad, y por otra parte existen problemas de interpretación ya que, aunque se cumpla que exp[log(y)] sea igual a y, el valor predicho por la ecuación viene dado por exp[E(log(y))] el cual es diferente al valor de E(y).

Así pues, será necesario elegir el modelo adecuado a este tipo de variable. Cuando trabajamos con el MRP es importante comprobar el supuesto básico de equidispersión y en su defecto la aparición de sobredispersión. Un primer indicio de la existencia de sobredispersión se tiene a partir de los resultados del MRP en los que se comprueba que el valor del cociente entre la Discrepancia y sus grados de libertad es superior a 1. Posteriormente se comprueba la sobredispersión de los datos a través de la prueba basada en la regresión (Cameron y Trivedi, 1990).

Ante la sobredispersión existen dos opciones, como señalan entre otros Hardin y Hilbe (2001): se puede realizar un ajuste post hoc de los errores estándar, utilizando para ello los diferentes índices que existen para tal fin, o por otro lado modelar con un modelo que sea más tolerante con la falta de equidispersión, como el modelo de regresión de la Binomial Negativa (MRBN) (Lindsey, 1995).

El modelo más ampliamente utilizado en situaciones de sobredispersión es el MRBN ya que es capaz de recoger la sobredispersión causada por heterogeneidad no observada. En este modelo la variancia viene dada por $V(y) = \mu + \alpha\mu^2$, por lo que se necesita una estimación de la constante alfa que se obtiene por máxima verosimilitud. En nuestro estudio este valor fue 2.706.

La no-adecuación del MRP en nuestro estudio se vio a través de los índices BIC y AIC ya que en este modelo sus valores fueron de –68.12 y 6.95, siendo –1493.21 y 3.79 en el MRBN, lo que indicó un mejor ajuste del MRBN.

Una manera de comparar la eficacia de ambos modelos fue por medio de los residuales de discrepancia producidos en cada modelo. Las observaciones ajustadas correctamente por un modelo deberían presentar unos residuales que se movieran en el intervalo –2 a +2. Así, se observó que todas las observaciones ajustadas por Poisson también eran bien ajustadas por la Binomial Negativa, pero observaciones mal ajustadas por MRP eran ajustadas correctamente por el MRBN.

Otra manera de evaluar la diferencia entre MRP y MRBN fue por medio de las probabilidades predichas por cada modelo respecto a los valores observados. Ambos modelos proporcionan una tasa predicha de 1.39 lo que indicó que las estimaciones del modelo de Poisson eran consistentes, aun en presencia de sobredispersión.

La existencia de un número excesivo de ceros, es decir un número superior al predicho por el modelo, nos podía indicar la existencia de una mezcla de distribuciones. En nuestro caso era habitual encontrar esta situación, ya que en la muestra podían haber sujetos que no sean consumidores de cannabis por lo que su consumo en el último mes era cero, aunque eran sujetos expuestos a la posibilidad de consumir.

El test de Vuong (Vuong, 1989) permitió comparar el modelo ZIP frente al modelo de Poisson, es decir modelos no anidados, proporcionando un valor $z=6.97$ ($P<0.0001$) según el cual el modelo ZIP proporcionaba un mejor ajuste que el modelo de Poisson.

Otra opción ante el incumplimiento de la equidispersión fue la corrección de los errores estándar a través de los índices y estimadores mostrados en los dos primeros artículos.

En el 5º estudio ”Las estrategias de afrontamiento: factores de protección en el consumo de alcohol, tabaco y cannabis” (Llorens, Perelló y Palmer, 2004) se llevó a cabo la comprobación empírica, en una muestra de datos reales, sobre el cumplimiento de los supuestos del MRP. Se utilizó la prueba de hipótesis basada en la regresión (Cameron y Trivedi, 1998) para evaluar la existencia de sobredispersión.

En las bebidas destiladas el grado de dispersión medido a través de la discrepancia del modelo fue de 3,48 (t=2,82 ; P=0.006), en las bebidas fermentadas 2,84 (t=1,46 ; P=0.148), en el tabaco fue de 17,47 (t=7,96 ; P<0.001) y en el cannabis el grado de dispersión fue de 3,57 (t=4,49 ; P<0.001). Así pues, se comprobaba el incumplimiento del supuesto de equidispersión en bebidas destiladas, en tabaco y en cannabis, mientras se cumplía el supuesto en bebidas fermentadas.

La corrección de los EE se podía realizar por medio de diferentes índices señalados en las investigaciones previas, entre ellos los estimadores Sandwich (Eicker 1967; Huber 1967; White 1980), la Discrepancia y la χ^2 (Hardin y Hilbe, 2001), el Bootstrap y el Jackknife (Efron y Tibshirani, 1993). Optamos por la aplicación de la corrección del Sandwich Robusto (HC0) también denominado estimador robusto de la matriz de variancias-covariancias. La selección de la corrección se basó en el estudio de adecuación de estas correcciones en función del grado de dispersión presentado en la muestra. Como señalan Long y Ervin (2000) a medida que el tamaño muestral aumenta, los estimadores sandwich estiman mejor el error estándar, presentando todos ellos un comportamiento similar a partir de muestras de 100 sujetos. Con tamaño de muestra grande, la estimación de los diferentes sandwich converge al valor teórico, independientemente del grado de sobredispersión, por ello hemos aplicado la corrección Sandwich en su versión HC0.

Lo que ocurrió al corregir los errores estándar es que variables que aparecían como explicativas de la conducta de consumo en el modelo original, dejaban de serlo.

Demostramos así con datos reales la importancia tanto de aplicar el modelo adecuado a los datos, como de comprobar los supuestos de esos modelos.

En una segunda fase de la investigación se planteó la opción como señala Lindsey (1998) de modelar con otros modelos más permisivos con el exceso de variabilidad de los datos. El primer modelo que se planteó fue el de la BN, posteriormente se compararon con modelos que se adecuaban mejor a la naturaleza de los datos. Los modelos se aplicaron a datos reales del campo de las conductas adictivas, en función de las características que éstos presentaban.

En el estudio 6º, “Activity levels and drug use in a sample of Spanish adolescents” (Llorens, Perelló y Palmer, en prensa) se comprobaron directamente los supuestos de aplicación del MRP y se compararon los diferentes modelos a aplicar en función de las características de los datos. En este caso después de comprobar los posibles modelos se optó por el MRBN. Este modelo representa de forma más fidedigna las características de los datos, esto es, datos de recuento con sobredispersión, donde no aparece un exceso de ceros y, por tanto, lo que requieren este tipo de datos es un modelo más permisivo en cuanto al supuesto de equidispersión.

Por el contrario, en el estudio 7º “Características de personalidad en adolescentes como predictores de la conducta de consumo de sustancias psicoactivas” (Llorens, Palmer y Perelló, 2005), nos encontramos con datos de recuento con sobredispersión y con un exceso de ceros. Después de aplicar diferentes modelos y evaluarlos, llegamos a la conclusión de que el modelo adecuado era el ZIP. Los datos analizados presentaban dos tipos de ceros, los debidos a sujetos que no consumían nunca y los debidos a sujetos que aunque si que consumían de forma habitual, no habían consumido en la última semana.

Como señalan Gardner, Mulvey y Shaw (1995) en una situación de sobredispersión hay dos alternativas a utilizar. En primer lugar, el investigador puede corregir las inferencias estadísticas obtenidas desde el modelo de regresión de Poisson, estimando un parámetro de dispersión y usando éste para corregirlas. Por otra parte, pero en la misma línea, en lugar de corregir los EE estimados por el modelo puede ser preferible estimar por otra vía estos EE. En este sentido existen diferentes formas para llevar a cabo esta estimación, como puede ser la estimación robusta o las estimaciones por remuestreo. Esta primera alternativa puede ser el método de elección cuando el investigador está interesado principalmente en la prueba de hipótesis sobre los coeficientes del modelo de regresión de Poisson. Sin embargo, el modelo de Poisson con sobredispersión no especifica la distribución de probabilidad de los datos, por lo que en esta situación puede ser recomendable cambiar de distribución.

Feller (1943) y Mullahy (1997) han mostrado como la probabilidad del cero en un modelo mixto de la distribución de Poisson es mayor que la probabilidad de cero en una distribución de Poisson ordinaria con la misma media.

Sobre los artículos aplicados a las conductas adictivas, se realizaron las comprobaciones pertinentes, señaladas en la introducción y pertenecientes a la etapa de evaluación, pero las características de las revistas a las que iban dirigidos estos artículos nos impidieron extendernos en la parte metodológica, obligándonos a omitir bastante información. Por ello adjuntamos un anexo en el que se realiza un estudio de residuales, de valores extremos y valores influyentes en datos de recuento.

Discusión:

En esta tesis se han descrito diferentes enfoques del estudio de datos de recuento. En primer lugar se ha señalado la importancia de tener en cuenta y comprobar el cumplimiento de supuestos y la aplicación del modelo adecuado a los datos.

La existencia de multitud de pruebas para la comparación de modelos nos permite valorar la adecuación de un modelo específico a los datos con los que trabajamos. No obstante no debemos caer en sobreajustes de modelos que no nos permitan ir más allá de la muestra con la que trabajamos. Es importante como señala Sturman (1999) tener claros los objetivos que pretendemos.

La etapa de evaluación, en muchos casos olvidada, es un punto clave en el proceso de análisis de datos, de manera que si es importante aplicar el modelo teórico adecuado a unos datos, igual de importante es comprobar la adecuación de ese modelo a los datos con los que trabajamos.

Los datos de recuento son muy utilizados en el campo de las conductas adictivas, pero debido al desconocimiento que existe actualmente del análisis de este tipo de datos, es habitual comprobar la aplicación de modelos lineales o la transformación de los datos como método de análisis. El hecho de poder unir en un mismo análisis a todos los sujetos independientemente de la sustancia que consumen y de la cantidad consumida rompe una barrera que parecía infranqueable en el mundo de las conductas adictivas. Si se observan las investigaciones en sustancias adictivas, encontramos que hay muchos problemas en la clasificación del tipo de consumo que hacen las personas, obligándose en muchas ocasiones a hacer una descripción previa a las investigaciones de lo que se considera un consumidor y un no consumidor, o bien, agrupando la información para poder aplicar análisis. Si nos basamos en

la cantidad consumida en una semana de cada una de las sustancias sin tener en cuenta la etiqueta a la que se puede unir ese sujeto, esto nos permitirá trabajar con toda la información de la que se dispone, sin perder esta información por agrupaciones bajo etiquetas creadas por el investigador.

Esperamos que las aportaciones de la presente tesis haga cambiar de opinión a aquellos que se hayan planteado aplicar un modelo no adecuado a los datos de recuento. Hemos intentado dar una visión general de las diferentes opciones que tenemos cuando trabajamos con datos de recuento. Cómo comprobar los supuestos y cómo comparar los diferentes modelos para escoger el más adecuado a las características de los datos.

Nos hemos encontrado con lagunas de conocimiento en la etapa de evaluación en datos de recuento. Concretamente hemos comprobado que en el caso de no haber información sobre algo concreto, lo que los investigadores hacen es aplicar lo utilizado en regresión lineal.

Hemos intentado dar suficiente información para que un investigador que trabaje con datos de recuento tenga una visión general de cómo analizar los datos y como realizar la evaluación de ese análisis.

Hemos demostrado cómo el hecho de que se pueda aplicar un modelo no nos da ninguna seguridad de que esté bien hecho. Una cosa es que podamos aplicar un modelo que en principio es adecuado por la naturaleza de los datos y otra cosa es que debido a las características concretas de los datos ese modelo sea el adecuado. Así, es adecuado aplicar el modelo de Poisson a datos de recuento, pero si estos datos por ejemplo, presentan un exceso de ceros, será más adecuado aplicar un ZIP, si debido a las características de los datos, no pueden haber ceros, será más adecuado un modelo de Ceros truncados, si presenta sobredispersión puede ser adecuado una corrección o la aplicación de un modelo más permisivo en cuanto a este supuesto, como el MBN.

Se habla de la adecuación de aplicar Poisson a datos de recuento, y en principio es así, de hecho es siempre más adecuado aplicar Poisson que regresión lineal, como se ha demostrado en el tercer artículo, pero la etapa de evaluación nos permite ser más rigurosos y más exactos en el análisis de los datos.

En vista de los resultados, la investigación futura puede orientarse hacia:

- Encontrar los factores de corrección para los estimadores e índices que corrigieran el sesgo producido por la sobredispersión.
- Realizar un estudio más amplio comparativo de los diferentes residuales en el estudio de datos de recuento.
- Estudiar la adecuación de modelos más generales, como los modelos de clase latente.

El objetivo futuro más importante que planteamos de la presente tesis es difundir el conocimiento que hemos alcanzado para que en el campo sustantivo se apliquen de forma rigurosa los modelos adecuados a los datos con los que se trabajan, siendo estrictos en la evaluación que se hacen de los modelos.

V.- ANEXO:

“Residuales y gráficos en la etapa de evaluación”

Como ilustración del uso de residuales en la etapa de evaluación del modelo, utilizaremos una matriz de datos de las manejadas en los artículos.

La variable de respuesta es la cantidad de bebidas fermentadas consumidas en la última semana, y las variables explicativas introducidas en el modelo son las estrategias de afrontamiento, habilidades propias y habilidades sociales. El modelo aplicado es el modelo de regresión de Poisson. En la tabla 1 se muestran la salida dada por el Stata al aplicar el modelo a los datos.

Tabla 1: Salida del Modelo de Regresión de Poisson.

```
 Generalized linear models                         No. of obs      =       159
Optimization     : ML: Newton-Raphson              Residual df     =       144
                                                   Scale parameter =         1
Deviance         =   519.883849                    (1/df) Deviance =  3.610305
Pearson          =  681.3362483                    (1/df) Pearson  =  4.731502
Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]
Standard errors  : OIM
Log likelihood   = -510.1551536                    AIC             =  6.605725
BIC              = -210.0383561
------------------------------------------------------------------------------
      cant1f |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        afr1 |  -.0398938   .1186568    -0.34   0.737    -.2724569    .1926693
        afr2 |  -.5512503   .1859945    -2.96   0.003    -.9157929   -.1867077
        afr3 |   -.305054   .1147282    -2.66   0.008    -.5299171   -.0801909
        afr4 |   1.838029   .5211908     3.53   0.000     .8165139    2.859544
    hpropia1 |   .1998759    .093805     2.13   0.033     .0160214    .3837304
    hpropia2 |  -.3278218   .1203659    -2.72   0.006    -.5637347   -.0919089
    hpropia3 |   .1319046   .1683773     0.78   0.433    -.1981089     .461918
    hpropia4 |   .2843753   .3740043     0.76   0.447    -.4486597     1.01741
    hsocial1 |   -.054179   .1663123    -0.33   0.745    -.3801451    .2717871
    hsocial2 |  -.0907907   .1100339    -0.83   0.409    -.3064532    .1248717
    hsocial3 |  -.3019776      .1946    -1.55   0.121    -.6833867    .0794314
    hsocial4 |   .3178738   .2321263     1.37   0.171    -.1370853    .7728329
    hsocial5 |  -.1318848   .1100909    -1.20   0.231    -.3476591    .0838895
    hsocial6 |   1.159542   .2476012     4.68   0.000     .6742522    1.644831
       _cons |   1.576799   .4390139     3.59   0.000     .7163478    2.437251

------------------------------------------------------------------------------
```

Se comprueba el cumplimiento del supuesto de equidispersión del modelo, mediante la prueba de hipótesis basada en la regresión (Cameron y Trivedi, 1998) obteniéndose una t=1.71 (p=0.089).

El modelo de regresión de Poisson aplicado parece presentar un ajuste adecuado de los datos. En este anexo vamos a comprobar la adecuación de este modelo mediante la utilización de residuales y gráficos.

En la tabla 2 se presenta un resumen estadístico de diferentes definiciones de residuales aplicadas al modelo analizado.

Estos residuales son:

Ans: residual de Anscombe

Ans2: residual de Anscombe estandarizado

Dev: residual de discrepancia

Dev2: residual de discrepancia estandarizado

Dev3: residual de discrepancia ajustado

Pear: residual de Pearson

Pear2: residual de Pearson estandarizado

Resp: residual de respuesta

Resp2: residual de respuesta estandarizado

Work: residual de trabajo

Work2: residual de trabajo estandarizado

Score: residual de puntuaciones

Score2:residual de puntuaciones estandarizado

Tabla 2:Resumen de los principales estadísticos de los residuales obtenidos.

VARIABLE	MEDIA	VARIANCIA	ASIMETRÍA	MAX	MIN	P90	APUNTAM.	P10
ans	-.2488713	3.399326	1.600015	11.28211	-3.937941	1.644078	11.84845	-2.115776
ans2	-.2614085	3.712368	1.447862	11.37148	-4.343148	1.704645	10.68742	-2.165246
dev	-.2334901	3.235541	1.66087	11.00395	-3.712726	1.641443	11.77648	-2.097198
dev2	-.2449806	3.53276	1.511433	11.09112	-4.094759	1.701913	10.61903	-2.153744
dev3	-.1602828	3.237785	1.656593	11.07086	-3.649241	1.711462	11.74094	-2.0208
pear	.0041605	4.312237	4.254495	17.58272	-2.625294	1.820866	34.1208	-1.753938
pear2	.0047965	4.619063	4.000859	17.722	-2.895432	1.887945	31.10867	-1.837645
resp	-1.14e-08	25.05944	4.676064	43.79572	-6.892168	3.79572	38.9093	-3.968119
resp2	.0046156	26.76045	4.414702	44.14265	-7.60136	3.825788	35.68313	-4.237586
work	.1869761	946.7153	5.060713	271.7209	-47.50198	23.54971	41.77355	-23.29919
work2	.3184706	1022.439	4.739728	273.8734	-52.38985	23.73626	37.7665	-25.57124
score	-1.14e-08	25.05944	4.676064	43.79572	-6.892168	3.79572	38.9093	-3.968119
score2	.0046156	26.76045	4.414702	44.14265	-7.60136	3.825788	35.68313	-4.237586

Como señala Cameron y Trivedi (1998), los residuales se acercan a la normalidad si no presentan asimetría y tienen un apuntamiento cercano a 3. Los residuales de estos datos analizados no se acercan a la normalidad, podemos observar un apuntamiento elevado, concretamente con el residual de trabajo, que también presenta una alta asimetría. Los residuales de Anscombe son los que presentan un valor más bajo, tanto en el apuntamiento como en la asimetría. Si tenemos en cuenta el criterio de normalidad, nuestros datos apuntan como residual de elección el residual de Anscombe, esta elección apoya la realizada por Cameron y Trivedi, (1998).

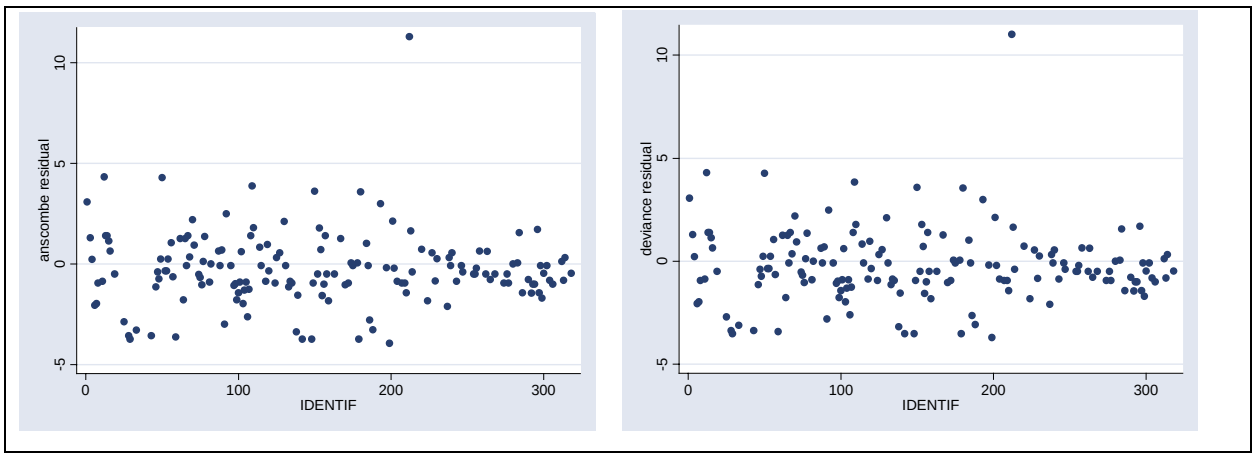
La semejanza entre los residuales se puede comprobar en la tabla 3, donde se dan las correlaciones entre los residuales y en los gráficos.

Tabla 3: Correlaciones entre los residuales analizados

	ans	ans2	dev	dev2	dev3	pear	pear2	resp	resp2	work	work2	score	score2
ans	1.0000												
ans2	0.9989	1.0000											
dev	0.9997	0.9985	1.0000										
dev2	0.9987	0.9997	0.9989	1.0000									
dev3	0.9997	0.9985	1.0000	0.9988	1.0000								
pear	0.9608	0.9552	0.9636	0.9581	0.9635	1.0000							
pear2	0.9646	0.9609	0.9674	0.9638	0.9673	0.9992	1.0000						
resp	0.9483	0.9418	0.9510	0.9445	0.9507	0.9940	0.9921	1.0000					
resp2	0.9526	0.9479	0.9552	0.9507	0.9549	0.9936	0.9934	0.9991	1.0000				
work	0.9069	0.9000	0.9094	0.9027	0.9089	0.9607	0.9583	0.9846	0.9842	1.0000			
work2	0.9056	0.9012	0.9082	0.9039	0.9076	0.9544	0.9542	0.9787	0.9811	0.9973	1.0000		
score	0.9483	0.9418	0.9510	0.9445	0.9507	0.9940	0.9921	1.0000	0.9991	0.9846	0.9787	1.0000	
score2	0.9526	0.9479	0.9552	0.9507	0.9549	0.9936	0.9934	0.9991	1.0000	0.9842	0.9811	0.9991	1.0000

Los resultados muestran también la aproximación entre los residuales, presentando una correlación entre ellos superior a 0.90 en todos los casos. Los residuales modificados presentan correlaciones superiores al 0.98 con respecto a sus residuales no modificados. Esto es debido al hecho de que el promedio de las h_i es 0.094, provocando este valor correcciones muy pequeñas.

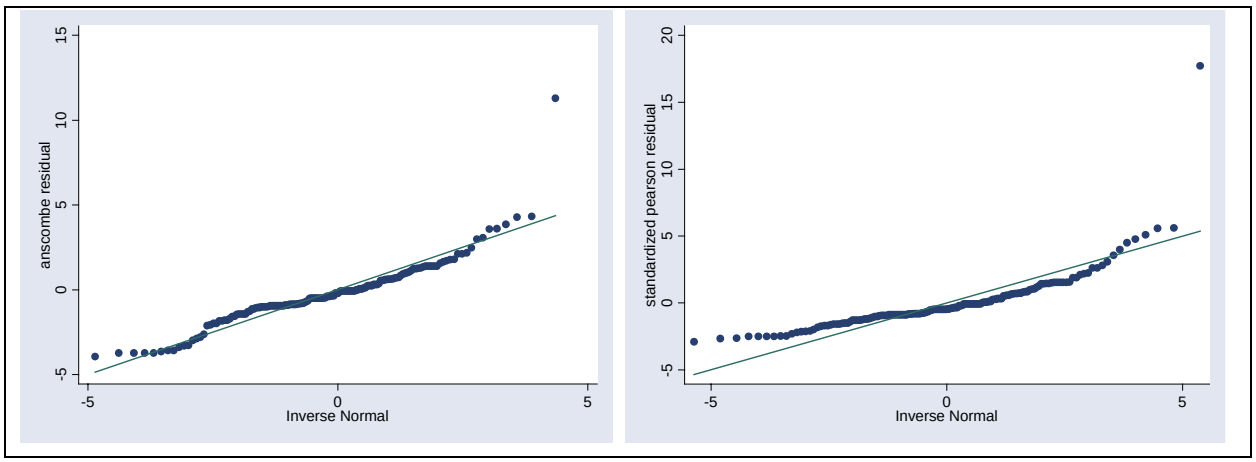
Grafico 1: Gráfico índice de residuales después de ajustar un modelo de regresión de Poisson.



En el gráfico índice de los diferentes residuales presentados, la observación 212 presenta un residual muy alto, esta observación corresponde a un sujeto que presenta un consumo semanal muy superior al resto de los sujetos (50 bebidas/semana). Los dos residuales detectan la observación anómala, presentando ambos, valores en el rango ± 5.

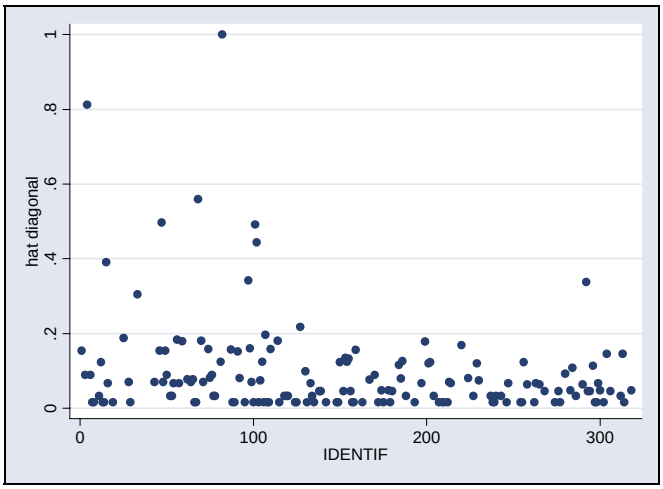
En el gráfico 2 presentamos los gráficos de probabilidad normal del residual de Anscombe y del residual de Pearson estandarizado. Como señala Breslow (1995) estos gráficos nos permiten ver la normalidad de los errores pero no la influencia de valores alejados. Aunque en el gráfico se ve la observación alejada, no podemos afirmar nada sobre su influencia. El gráfico índice de la Distancia de Cook o el gráfico de valores de influencia (leverages) nos darán esta información.

Gráfico 2: Gráfico de probabilidad normal



El siguiente gráfico índice de valores de influencia nos da información de la posible existencia de valores influyentes en los datos.

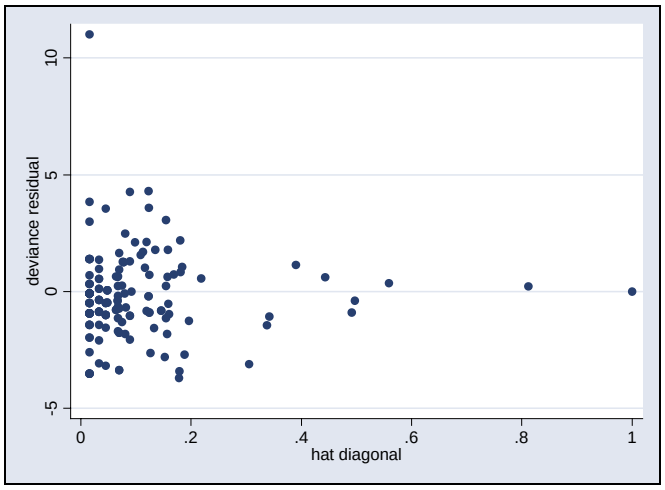
Gráfico 3: Gráfico índice de valores hat.



Para determinar qué observaciones se considerarán valores influyentes, tomamos como referencia el valor $h_{ii} > 3k/n = 0.26$. Observaciones con valores superiores a éste se considerarán influyentes. En el gráfico hay diez valores con un h_i, superior al de referencia. Como señalan Tang, Wei y Wang (2000) este gráfico nos permite ver la influencia de determinados casos, en la precisión de la estimación.

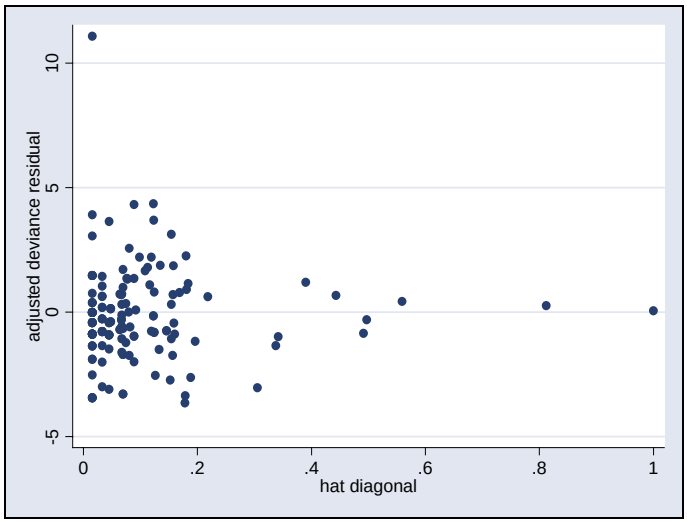
Otra forma de evaluar un modelo es presentando un gráfico de residuales frente a valores de influencia. Este gráfico (gráficos 4 y 5) permite detectar valores alejados (outliers) y observaciones influyentes al mismo tiempo.

Gráfico 4: Gráfico de residuales frente valores influyentes.



En el gráfico 4, podemos ver cómo la observación 212, que había aparecido en el gráfico 1 como un valor alejado, vuelve a aparecer y las observaciones influyentes obtenidas mediante el gráfico 3, pueden verse claramente en éste gráfico.

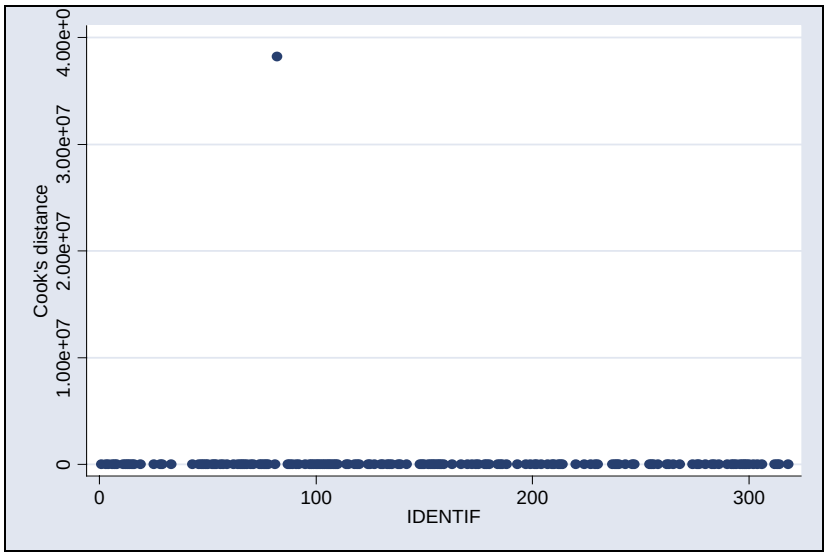
Gráfico 5: Gráfico de residuales de discrepancia ajustada frente valores influyentes.



Como señala Andersen (2004) estos gráficos son adecuados para detectar valores alejados e influyentes en los datos. La regla informal de interpretación de este gráfico nos dice que el residual será alto si supera el valor 2, por su parte hablaremos de valores influyentes si superan 2 ó 3 veces el valor promedio del hat. La media de los valores de influencia es 0.094, así consideraremos un valor influyente si supera el valor 0.282. Tanto en el gráfico 4, como en el 5, podemos observar como 10 valores superan esta barrera, del mismo modo aparecen gran cantidad de observaciones que superan la barrera señalada para el residual. Halekoh (2002) indica que podemos considerar que hay problemas en el modelo si más del 5% de los residuales superan el 2 en valor absoluto. En los datos, el 18.23% de los valores presentan un valor superior a 2 en valor absoluto. Esto puede deberse al hecho de que se suelen tomar como puntos de referencia los valores dados en regresión lineal y estos valores no siempre se ajustan a los modelos no lineales.

El gráfico índice de la Distancia de Cook nos informa principalmente de la influencia de cada observación en el conjunto de parámetros estimados. Fox (1991) sugiere como valor de referencia, para poder valorar la Distancia de Cook, una D > 4/(n-k-1), donde n es el tamaño y k el número de variables. En nuestros datos este valor sería 0.025.

Gráfico 6: Grafico índice de la Distancia de Cook



Del mismo modo que en regresión lineal, el gráfico índice de la Distancia de Cook nos informa de la influencia de una observación en el conjunto de parámetros estimados. Podemos ver en el gráfico como la observación 82 presenta un valor alejado. El valor inusualmente alto de esta observación puede estar indicando su influencia en el ajuste del modelo. La omisión de este valor puede cambiar la significación de los parámetros estudiados.

Se debe comparar este gráfico con el de valores alejados, ya que como indican Marasinghe y Duckworth (2003) una observación con un valor en la Distancia de Cook alto y un valor de influencia pequeño puede considerarse un “outlier importante”.

Resultados del modelo sin la observación 82

Tabla 4: Modelo de regresión de Poisson sin la observación 82.

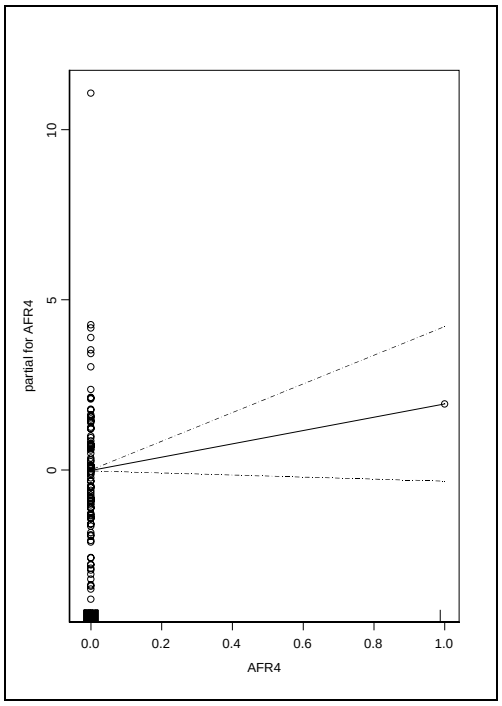
```
Generalized linear models                         No. of obs      =       158
Optimization     : ML: Newton-Raphson             Residual df     =       144
                                                  Scale parameter =         1
Deviance         =   519.883849                   (1/df) Deviance =  3.610305
Pearson          =  681.3362483                   (1/df) Pearson  =  4.731502

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -507.8766352                   AIC             =  6.606033
BIC              = -209.1298357

------------------------------------------------------------------------------
      cant1f |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        afr1 |  -.0398938   .1186568    -0.34   0.737    -.2724569    .1926693
        afr2 |  -.5512503   .1859945    -2.96   0.003    -.9157929   -.1867077
        afr3 |   -.305054   .1147282    -2.66   0.008    -.5299171   -.0801909
    hpropia1 |   .1998759    .093805     2.13   0.033     .0160214    .3837304
    hpropia2 |  -.3278218   .1203659    -2.72   0.006    -.5637347   -.0919089
    hpropia3 |   .1319046   .1683773     0.78   0.433    -.1981089     .461918
    hpropia4 |   .2843753   .3740043     0.76   0.447    -.4486597     1.01741
    hsocial1 |   -.054179   .1663123    -0.33   0.745    -.3801451    .2717871
    hsocial2 |  -.0907907   .1100339    -0.83   0.409    -.3064532    .1248717
    hsocial3 |  -.3019776      .1946    -1.55   0.121    -.6833867    .0794314
    hsocial4 |   .3178738   .2321263     1.37   0.171    -.1370853    .7728329
    hsocial5 |  -.1318848   .1100909    -1.20   0.231    -.3476591    .0838895
    hsocial6 |   1.159542   .2476012     4.68   0.000     .6742522    1.644831
       _cons |   1.576799   .4390139     3.59   0.000     .7163478    2.437251
------------------------------------------------------------------------------
```

Al eliminar la observación 82 comprobamos que la variable Afr4 desaparece del ajuste, esto es debido a la existencia de colinealidad en el nuevo modelo. Un estudio más específico de los datos nos indica que la fuente de esta colinealidad está en el

hecho de que el único sujeto que da respuesta positiva a esta variable es el sujeto eliminado. A partir de esta información la decisión sería la omisión de esta variable del modelo. Sin embargo la omisión de esta observación no produce ningún cambio ni en los coeficientes del parámetro ni en la significación del resto de variables introducidas en el modelo. El AIC del modelo aumenta respecto al modelo con todas las observaciones.

El gráfico 7 de residual parcial de la variable afr4, nos permite comprobar que únicamente tenemos una observación en la respuesta positiva.

Gráfico 7: Grafico de residual parcial para Afr4



<u>Resultados del modelo sin la observación 212.</u>

Como esta observación presenta un residual muy alto, en los gráficos mostrados y esto puede influir en el ajuste del modelo, hemos eliminado esta observación para valorar los cambios que produce en el ajuste del modelo.

Tabla 5: Modelo de regresión de Poisson sin la observación 212.

```
Generalized linear models                         No. of obs      =        158
Optimization     : ML: Newton-Raphson             Residual df     =        143
                                                  Scale parameter =          1
Deviance         =   393.725577                   (1/df) Deviance =   2.753326
Pearson          =  384.3164054                   (1/df) Pearson  =   2.687527

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM

Log likelihood   = -444.1994009                   AIC             =   5.812651
BIC              = -330.2255127

------------------------------------------------------------------------------
      cant1f |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        afr1 |  -.1049146   .1205945    -0.87   0.384    -.3412756    .1314463
        afr2 |  -.4955434   .1867829    -2.65   0.008    -.8616312   -.1294556
        afr3 |  -.2652557   .1165689    -2.28   0.023    -.4937265    -.036785
        afr4 |   1.671381   .5220611     3.20   0.001     .6481601    2.694602
    hpropia1 |   .1451249   .0945627     1.53   0.125    -.0402145    .3304643
    hpropia2 |  -.3920363   .1209874    -3.24   0.001    -.6291672   -.1549053
    hpropia3 |   .0746725   .1698185     0.44   0.660    -.2581657    .4075107
    hpropia4 |   .2474653   .3747788     0.66   0.509    -.4870876    .9820181
    hsocial1 |  -.0617398   .1679769    -0.37   0.713    -.3909685    .2674889
    hsocial2 |   -.037207   .1119328    -0.33   0.740    -.2565912    .1821772
    hsocial3 |  -.2758158   .1959559    -1.41   0.159    -.6598824    .1082507
    hsocial4 |   .2989993    .230435     1.30   0.194     -.152645    .7506436
    hsocial5 |  -.0789511   .1114141    -0.71   0.479    -.2973187    .1394165
    hsocial6 |   1.122559   .2474197     4.54   0.000     .6376253    1.607493
       _cons |   1.737294   .4405617     3.94   0.000     .8738092    2.600779
------------------------------------------------------------------------------
```

Que esta observación influye en el modelo, se puede comprobar en la mejora que se produce en el modelo al eliminarla. En primer lugar porque cambian todos los coeficientes del modelo y la significación de la habilidad propia 1, pero lo más destacado es que una única observación hace disminuir el AIC de 6.60 a 5.8.

Gráfico 8: Gráfico Distancia de Cook vs valores de influencia

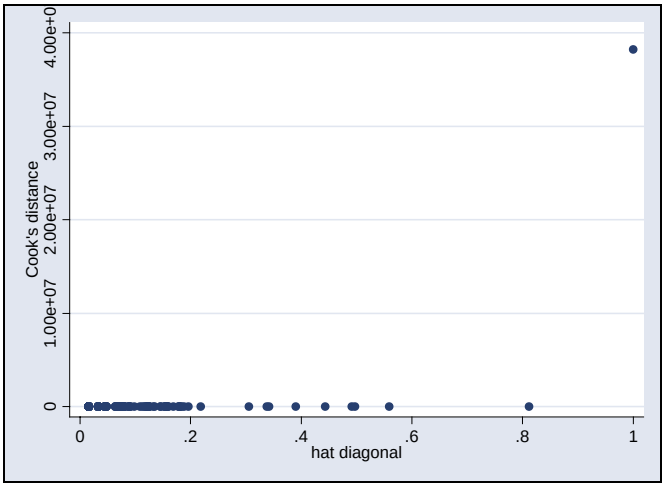


Gráfico 9: Gráfico Distancia de Cook vs valores de influencia al eliminar la observación 82.

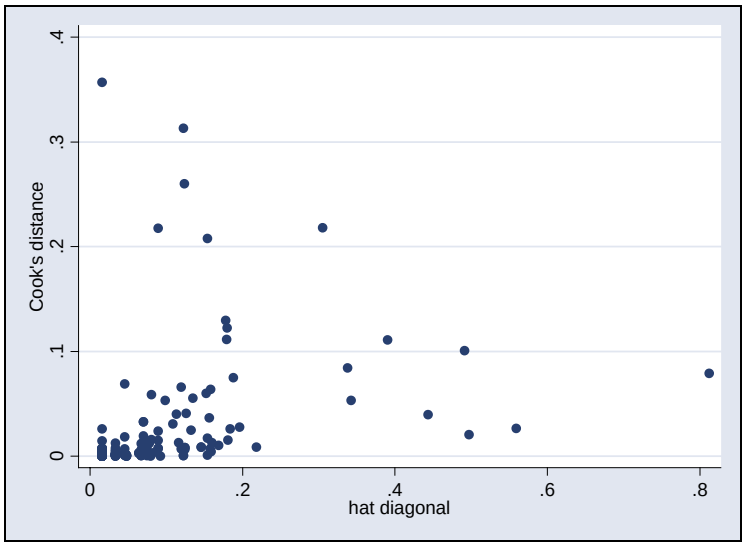
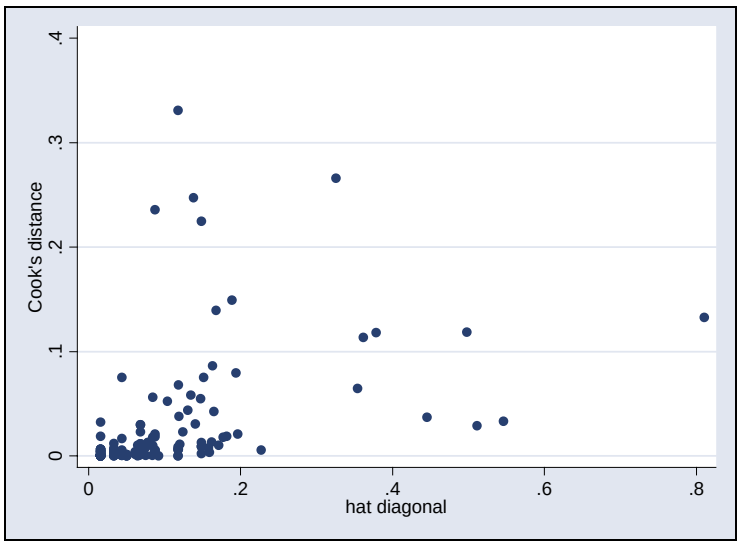


Gráfico 10: Gráfico Distancia de Cook vs valores de influencia al eliminar las observaciones 82 y 212.



Preisser y Garcia, (2005) señalan la utilización de este gráfico que enfrenta los valores de influencia a los valores de las observaciones en la Distancia de Cook. Tras la eliminación de dos observaciones con valores altos en estos índices (gráfico 10), aparecen las ocho observaciones restantes (gráfico 3) de las cuales una (observación 12) presenta un valor alto en la Distancia de Cook. Según lo señalado

por Marasinghe y Duckworth (2003), diez observaciones serían outliers en nuestros datos. La eliminación de estas observaciones podría producir cambios en el modelo, para comprobarlo eliminamos las ocho observaciones y ajustamos de nuevo el modelo.

Tabla 5: Modelo de regresión de Poisson después de eliminar las diez observaciones influyentes.

```
Generalized linear models                         No. of obs      =       148
Optimization     : ML: Newton-Raphson             Residual df     =       137
                                                  Scale parameter =         1
Deviance         =  500.2498322                   (1/df) Deviance =  3.651459
Pearson          =  664.3576647                   (1/df) Pearson  =  4.849326
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]
Standard errors  : OIM
Log likelihood   = -481.9534156                   AIC             =  6.661533
BIC              = -184.3682493
------------------------------------------------------------------------------
      cant1f |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        afr1 |  -.0912528   .1221273    -0.75   0.455    -.3306179    .1481122
        afr2 |  -.3809901   .2136044    -1.78   0.074    -.7996471    .0376669
        afr3 |  -.3497752   .1185694    -2.95   0.003     -.582167   -.1173834
    hpropia1 |   .1672193    .094972     1.76   0.078    -.0189225     .353361
    hpropia2 |  -.4043134   .1267432    -3.19   0.001    -.6527255   -.1559013
    hpropia3 |  -.0275556   .1853409    -0.15   0.882     -.390817    .3357059
    hsocial1 |   .0514557   .1722962     0.30   0.765    -.2862386    .3891501
    hsocial2 |  -.1034986   .1117938    -0.93   0.355    -.3226103    .1156132
    hsocial3 |  -.1073303   .1943793    -0.55   0.581    -.4883066    .2736461
    hsocial5 |  -.1537254   .1168417    -1.32   0.188     -.382731    .0752801
       _cons |   2.154528   .2749497     7.84   0.000     1.615636    2.693419
------------------------------------------------------------------------------
```

Al eliminar las ocho observaciones restantes, cuatro variables desaparecen del modelo por colinealidad. La fuente de esta colinealidad está en el hecho de que las observaciones eliminadas corresponden a sujetos que forman grupo en la respuesta frente al resto de sujetos. La respuesta en la variable se convierte en una constante y como consecuencia no está justificada su introducción en el modelo. En los gráficos 11 y 12 pueden observarse las distribuciones de las respuestas. En cuanto al modelo, al eliminar las 10 observaciones comprobamos que la significación de las variables cambia quedando únicamente como variables significativas Afrontamiento 3 y Habilidad propia 2 y modificándose también los coeficientes del modelo. No obstante lo más remarcado es el hecho de que el AIC aumenta, presentado este modelo un ajuste peor que el modelo con todas las variables.

Gráfico 11: Gráficos de residual parcial para Hpropia4 y Hsocial4

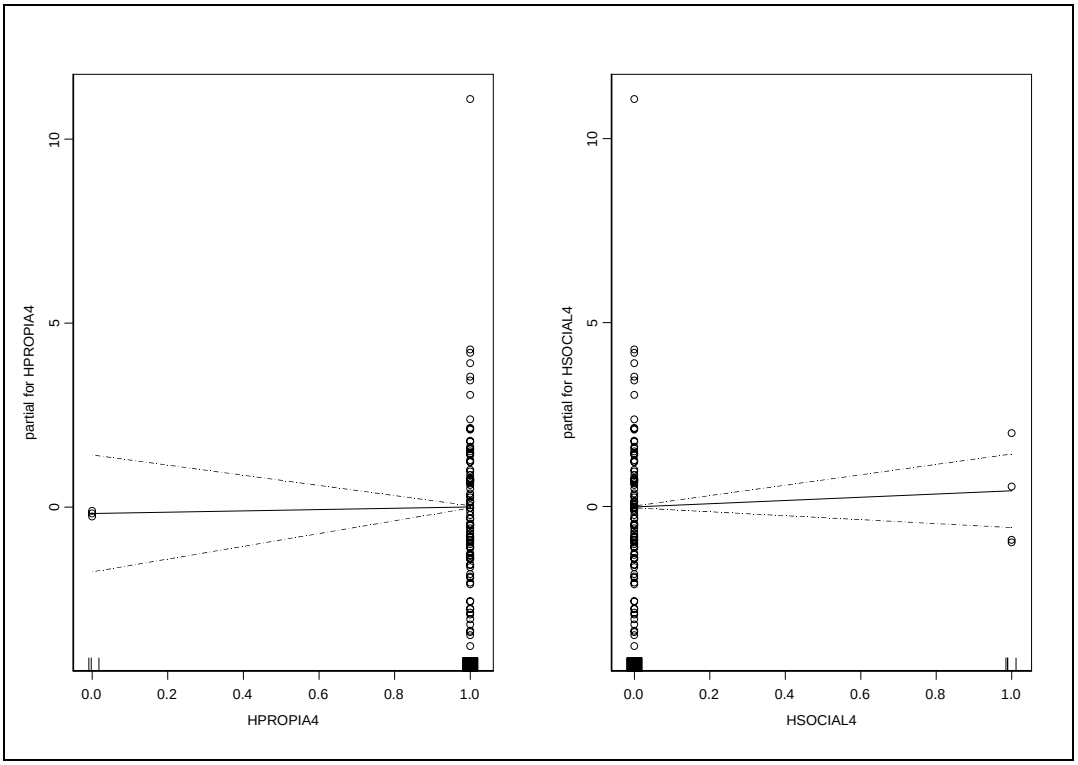
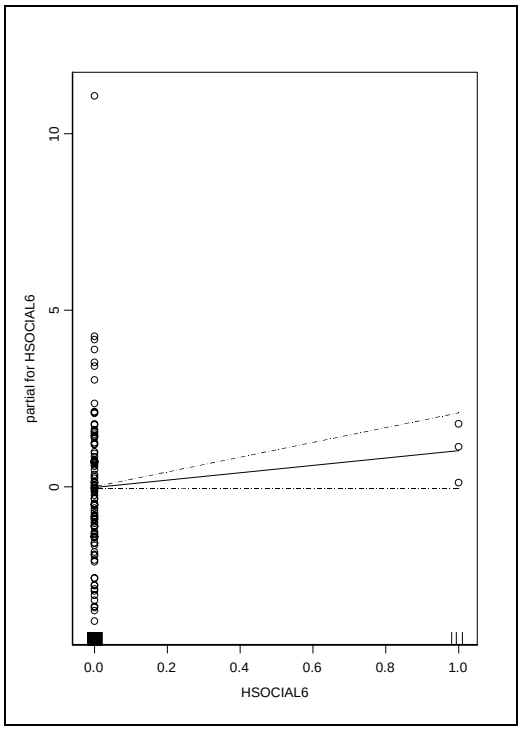


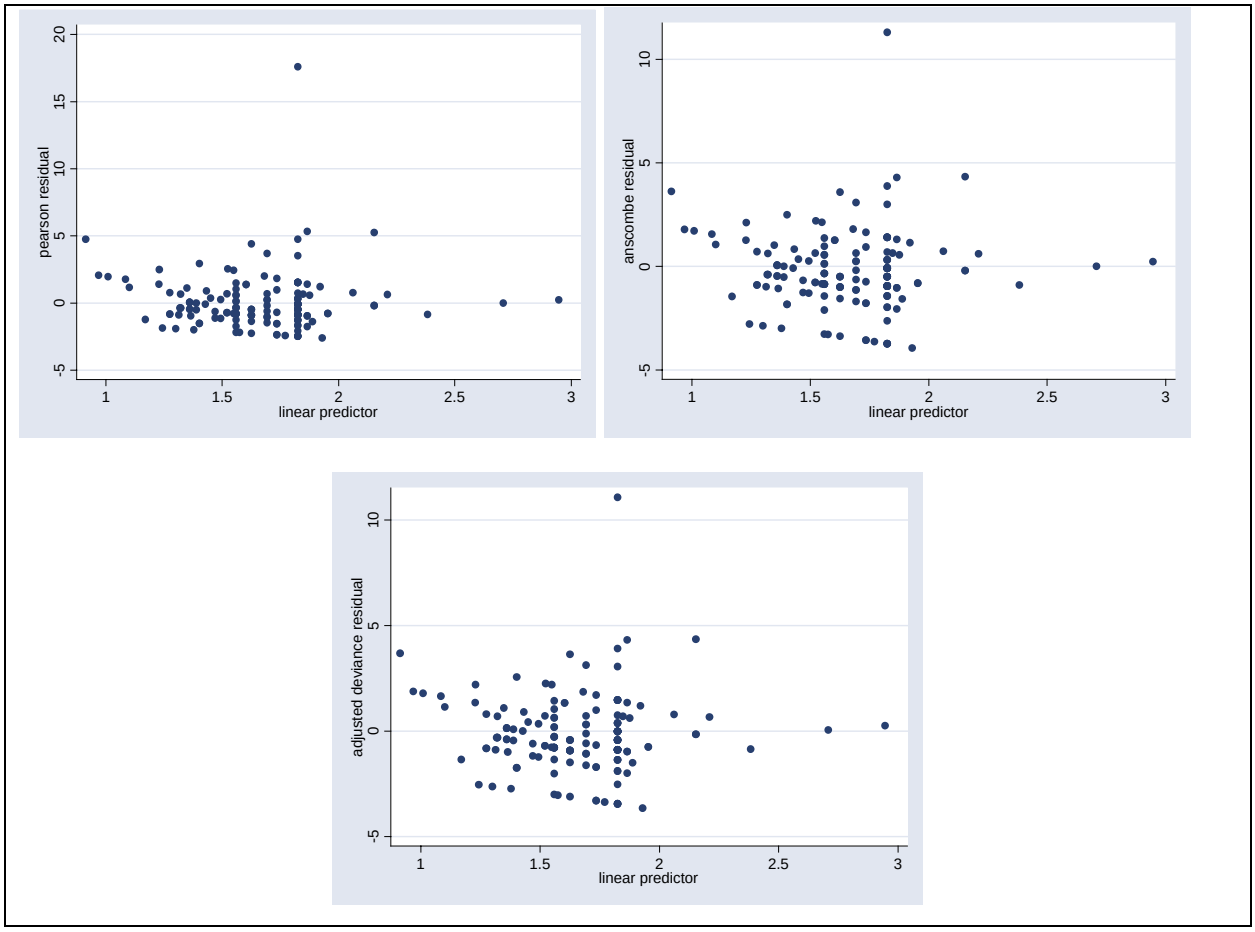
Gráfico 12: Gráficos de residual parcial para Hsocial6



Volviendo a los datos iniciales, vamos a comprobar a modo de ejemplo la adecuación de otros gráficos en la evaluación del modelo.

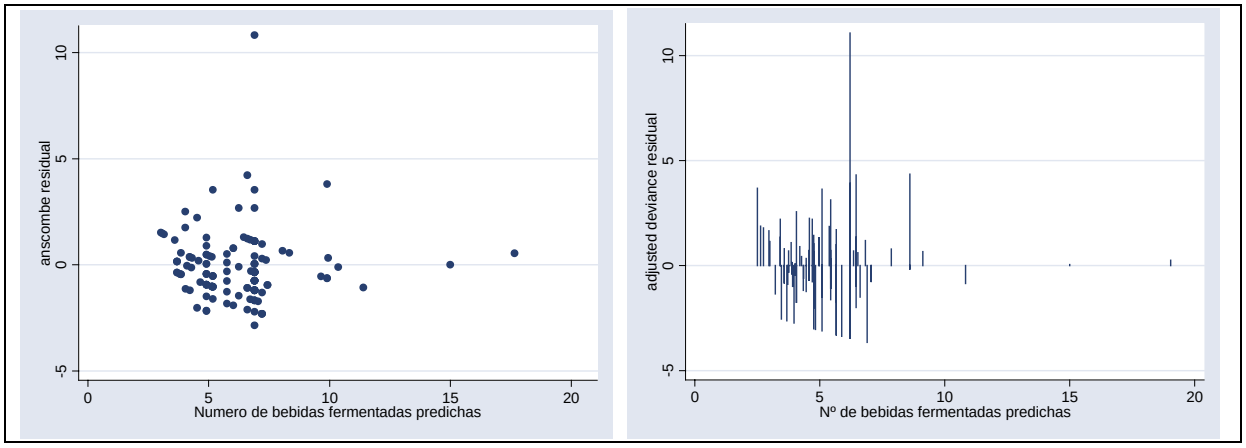
Los gráficos de los residuales frente a los valores del predictor lineal.

Grafico 13. Residual vs predictor lineal



Los gráficos de los residuales frente a los valores del predictor lineal señalan la adecuación del predictor lineal al modelo. Este gráfico debe presentar una línea recta para poder aceptar la adecuación del predictor lineal, en este caso se aceptaría la adecuación. Las curvas en este gráfico sugerirían la no-adecuación de la función de enlace. (Halekoh, 2002)

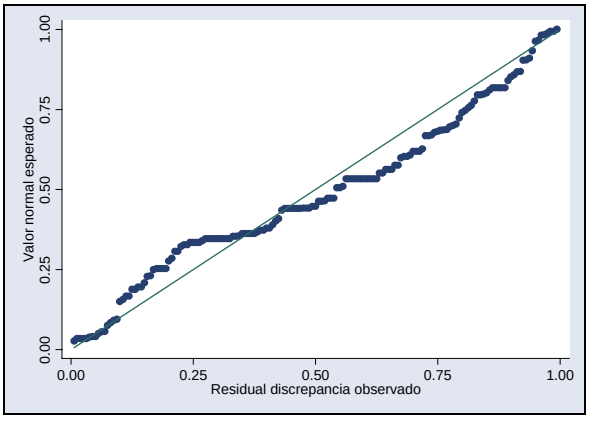
Gráfico 14: Residual vs predicción del modelo



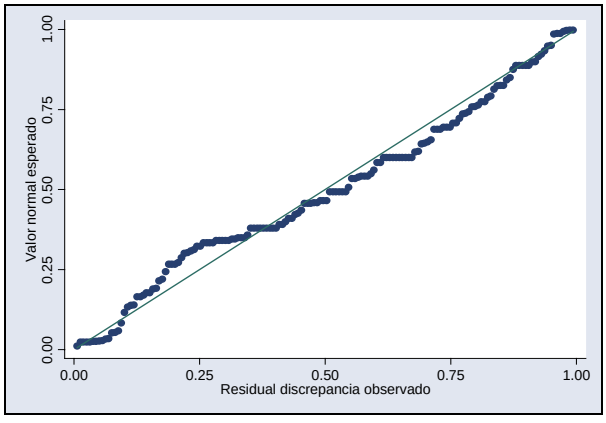
En el gráfico 14 se presentan diferentes residuales frente al valor predicho por el modelo, se puede observar una distribución aleatoria de los residuales alrededor del cero. Esto indica, como señalan entre otros Cameron y Trivedi (1998) y Halekoh (2002), la adecuación del modelo ajustado. Hoffmann(2004) recomienda utilizar este gráfico para comprobar la adecuación del modelo, pero teniendo precaución con las observaciones situadas en la parte superior del gráfico ya que son observaciones potencialmente influyentes. No obstante cuando hay poca variación en los valores, como es el caso, éste gráfico es un poco difícil de interpretar.

Grafico 15.Grafico de Probabilidad Normal

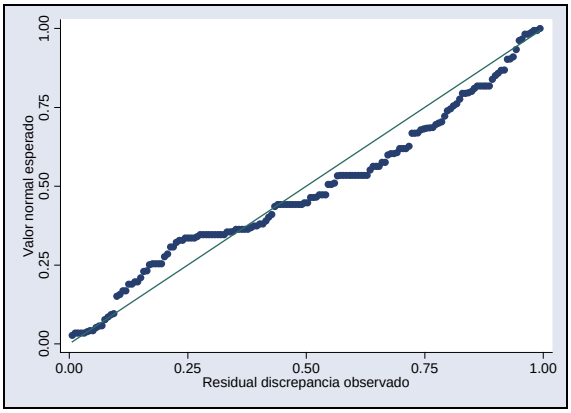
1: Todas las observaciones



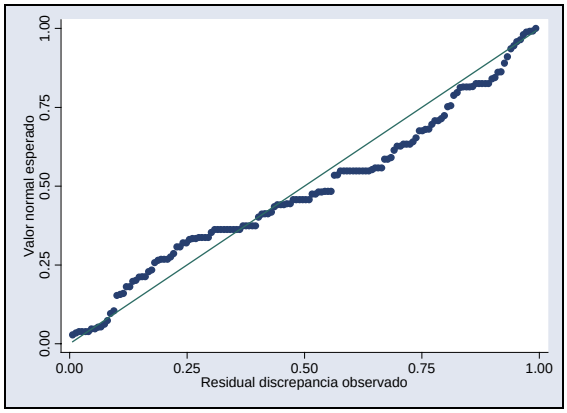
2: Eliminando la observación 212



3: Eliminando la observación 82

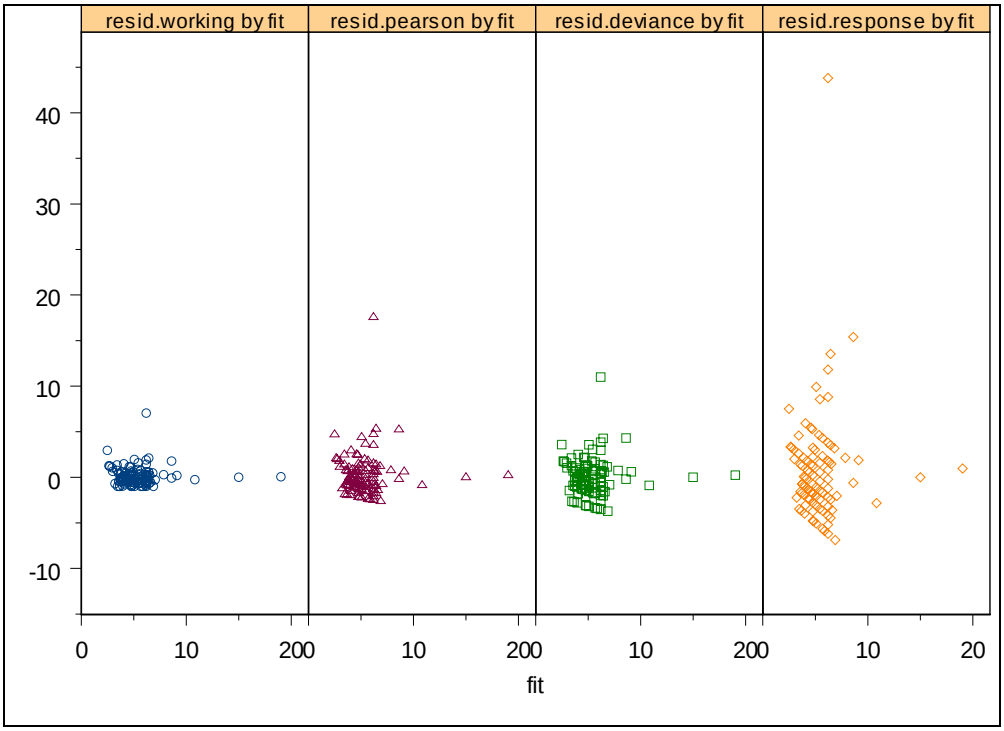


4:Eliminando las diez observaciones



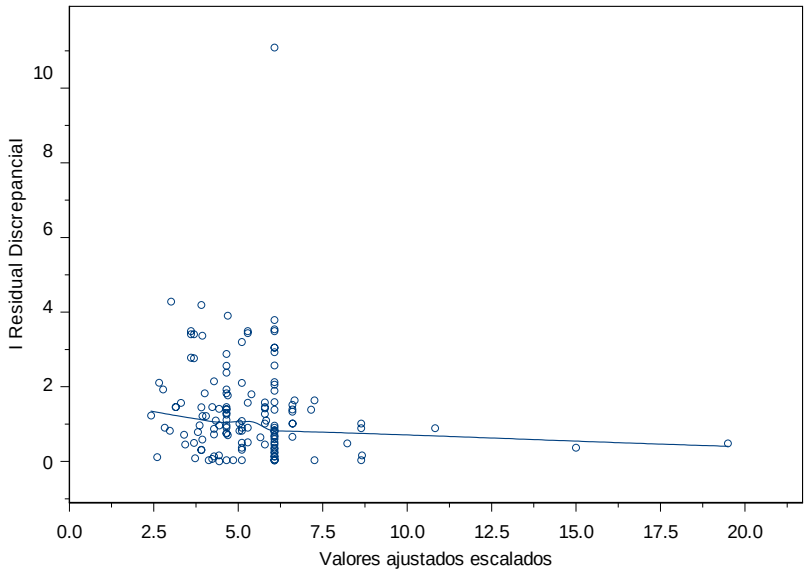
El gráfico de probabilidad normal del residual de discrepancia, (gráfico15) puede utilizarse en datos de recuento para comprobar el ajuste del modelo. Si los valores de las observaciones forman un ángulo de 45 grados podemos decir que el ajuste es correcto. En el gráfico 15(2) en el que se ha borrado la observación 212 se puede observar un mejor ajuste del modelo que en el resto de situaciones. Esto nos indica que aunque las observaciones presenten valores influyentes o incluso valores altos en la Distancia de Cook, no se justifica su eliminación del modelo, porque al eliminarlas el ajuste disminuye.

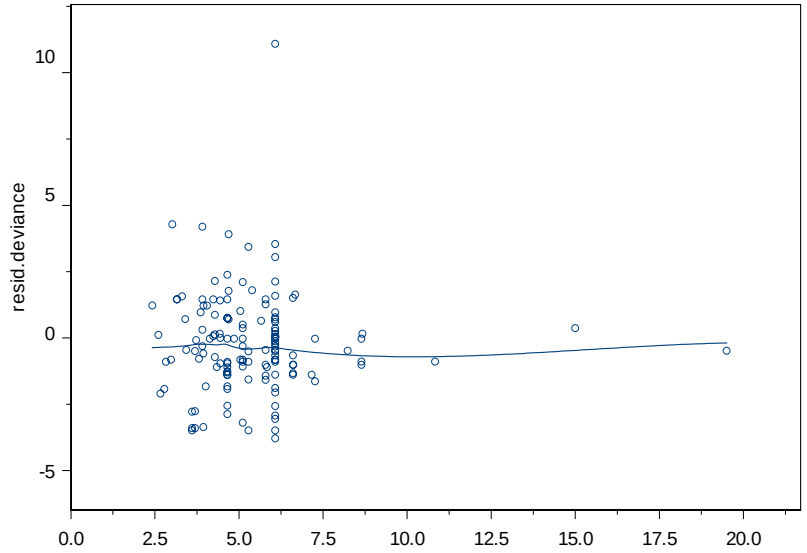
Grafico 16. Grafico de residuales vs valores ajustados



En el gráfico 16 se presentan los residuales frente a los valores ajustado por el modelo. Vemos como el que produce residuales más dispersos es el residual de respuesta, que también destaca de forma más contundente una observación con un residual muy alto. Esta observación se puede ver en todos los residuales mostrados pero en el residual de trabajo puede no detectarse al encontrarse muy cercana al resto de observaciones.

Grafico 17. Gráfico residual de discrepancia en valor absoluto y directo vs valores ajustados escalados





Este gráfico, como señalan McCullagh y Nelder (1989) y Lee y Nelder (2000) nos da información de la adecuación de la función de variancia asumida. No deben aparecer patrones, debería mostrar una línea plana y cercana a cero para poder hablar de una adecuada especificación de la función de variancia. Halekoh (2002) aconseja cambiar la función de variancia en caso de encontrar alguna tendencia positiva en

este gráfico. Davison y Gigli (1989) recomiendan utilizar este gráfico para comprobar las asunciones distribucionales.

Como señalan entre otros, Cameron y Trivedi (1998), McCullagh y Nelder (1989), lo mejor para estudiar y entender los residuales, así como para obtener mayores frutos de la información que nos ofrecen, es haciendo gráficos con ellos.

Por su parte, Ganio y Shafer (1992) recomiendan el uso de gráficos junto a índices numéricos para la evaluación del modelo, porque como ellos señalan en su trabajo, valores moderadamente alejados pueden presentar una gran influencia en los parámetros y esto únicamente se observa a través de gráficos.

Con este anexo se ha pretendido dar una visión general del uso de residuales y gráficos en la etapa de evaluación. El principal problema con el que nos hemos encontrado es la escasez de información que hay en la literatura sobre los residuales y gráficos cuando hablamos de datos de recuento y del modelo de Poisson. En modelos logit podemos encontrar algunos trabajos, entre otros Pregibon (1981); Menard(1995); Friendly (2003) al igual que en modelos como BN y el ZINB (Svetliza y Paula, 2001; Hinde y Demetrio, 1998).

En ocasiones y ante la falta de información podemos observar que se aplican los criterios de la regresión lineal o que se intentan acercar a la normalidad los modelos para justificar la aplicación de estos criterios. Ejemplo de esto es Hamilton (n.d.) y UCLA.