El diagnóstico de la sobredispersión en modelos de análisis de datos de recuento

Jaume Vives Brosa

Tesis doctoral dirigida por el Doctor:
Josep Maria Losilla Vidal

Departament de Psicobiologia i de Metodologia
de les Ciències de la Salut

Facultat de Psicologia

Universitat
Autònoma
de Barcelona

2002

A la Bet
Als meus pares

Agradecimientos:

Son muchas las personas que me han ayudado y apoyado durante el largo camino hasta llegar aquí; ellas saben quienes son y estoy seguro que ni tan sólo esperan que las nombre, así que no lo haré. Sin embargo, sin una de esas personas el presente trabajo habría resultado, por diversos motivos, simplemente imposible y, por tanto, no quiero desperdiciar esta oportunidad de agradecérselo. Me refiero a mi director de tesis, el Dr. Josep Maria Losilla.

El diagnóstico de la sobredispersión en modelos de análisis de datos de recuento

Jaume Vives Brosa

Tesis doctoral dirigida por el Doctor:
Josep Maria Losilla Vidal

Departament de Psicobiologia i Metodologia
de les Ciències de la Salut

Facultat de Psicologia

Universitat Autònoma de Barcelona

2002

Contenidos

1
Justificación y objetivos

Las variables de recuento son un tipo de información que cuenta con una amplia presencia en diversos ámbitos de investigación aplicada tanto en las Ciencias Sociales como en las Ciencias de la Salud. En este sentido, encontramos ejemplos recientes de investigaciones aplicadas con variables de recuento en disciplinas como Demografía (Melkersson y Rooth, 2000; Wang y Famoye, 1997), Farmacología (Lindsey, Jones y Jarvis, 2001), Relaciones Laborales (Sturman, 1999), Criminología (Osgood, 2000) o Biología (Cox, Heyse y Tukey, 2000), por citar algunos.

Mucho más numerosas resultan las investigaciones que contienen variables de recuento en Medicina (Biggeri, Marchi, Lagazio, Martuzzi y Böhning, 2000; Navarro, Utzet, Puig, Caminal y Martín, 2001), Ciencias Políticas (King, 1988) y Ciencias Económicas (García-Crespo, 2001; Meliciani, 2000). De hecho, las tres disciplinas merecen ser consideradas a parte puesto que no sólo cuentan con una extensa aplicación de investigaciones con variables de recuento, sino que han hecho valiosas aportaciones en el tratamiento estadístico de este tipo de variables (Cameron y Trivedi, 1986, 1990, 1998; Gourieroux, Monfort y Trognon, 1984a; Hausman, Hall y Griliches, 1984; King, 1989a; 1989b; King y Signorino, 1995; Winkelmann, 2000).

En cuanto a la presencia de las variables de recuento en Psicología, el presente trabajo toma como punto de partida y como primer objetivo el estudio de la frecuencia con que intervienen este tipo de variables en la investigación aplicada en Psicología. Para ello se presenta en el capítulo 2 un estudio bibliométrico que describe la frecuencia de uso de variables de recuento en diversas especialidades de nuestra disciplina. Este mismo estudio bibliométrico también permite cubrir el segundo objetivo de nuestro trabajo, que es conocer el tipo de análisis estadístico que se aplica habitualmente en la investigación en Psicología cuando la variable de respuesta a modelar es de tipo recuento.

Los recuentos se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación durante un intervalo temporal o espacial definido (Lindsey, 1995b). De esta definición se desprenden dos de las principales características idiosincrásicas de las variables de recuento que permiten diferenciarlas de, por ejemplo, las variables cuantitativas continuas: su naturaleza discreta y no negativa.

El conocimiento acerca de la naturaleza de una variable, así como la identificación de sus características distribucionales, constituyen la base a partir de la cual se justifica la aplicación de un modelo estadístico determinado. En este sentido, Long (1997, p. 3) advierte de las consecuencias de modelar ignorando la naturaleza de la variable objeto de estudio:

«*Una vez se determina el nivel de la variable dependiente, es importante que el modelo usado y el nivel de medida coincidan. Si el modelo seleccionado asume un nivel de medida erróneo, el estimador puede resultar sesgado, ineficiente, o simplemente inapropiado*».

Esta coincidencia entre modelo estadístico aplicado y nivel de medida puede conseguirse a través de diferentes estrategias de análisis de datos. De acuerdo con Sturman (1999, p. 415), podemos clasificar tales estrategias en dos grandes alternativas:

- cambiar las características de los datos de forma que se ajusten a los supuestos de los «*métodos estadísticos clásicos*», esto es, ajustar los datos al modelo, o bien,
- aplicar el procedimiento estadístico que se ajuste mejor a los datos, es decir, ajustar el modelo a los datos.

La primera de las alternativas pasa por adoptar estrategias que hagan posible la aplicación de procedimientos estadísticos a datos para los que no fueron concebidos. En el ámbito de las variables de recuento, la más frecuente de tales estrategias es la transformación de los datos y el objetivo es, a menudo, la aplicación del modelo lineal general. En este sentido, Winkelmann (2000, p. 1), señala que aunque la regresión lineal es usada generalmente como herramienta de regresión multipropósito, «*para datos discretos en general, y para recuentos en particular, la regresión lineal normal presenta ciertos problemas que hacen que sus uso sea dudoso y lógicamente insatisfactorio*».

Por otro lado, tal como indican diversos autores (Ato, Losilla, Navarro, Palmer y Rodrigo, 2000b; King, 1988; Lindsey, 1995b), la vía de ajustar el modelo a los datos mediante transformaciones no tan sólo comporta problemas nuevos como, por ejemplo, sesgos en la estimación o dificultad de interpretación de los resultados, sino que no resuelve los problemas de ajuste a las condiciones de aplicación del modelo lineal general. En este sentido, tal como indica Sturman (1999), cuando se cumplen las asunciones del modelo estadístico empleado, habitualmente los coeficientes describen correctamente la relación. Sin embargo, cuando se violan dichas asunciones, como por ejemplo al emplear regresión lineal para analizar datos procedentes de distribuciones no normales, las estimaciones pueden no ser válidas. Concretamente, pueden identificarse relaciones inexistentes (Errores Tipo I) o infravalorar relaciones existentes (Errores Tipo II).

La recomendación de Gardner, Mulvey y Shaw (1995, p. 394) en relación a la transformación de datos es clara:

«*En lugar de transformar los datos para que cumplan las asunciones del modelo de regresión lineal, deberían considerarse modelos de regresión que capturen las características naturales de los datos*».

En esta misma línea se expresan también Ato et al. (2000b), al señalar que una estrategia de modelado óptima debe respetar la naturaleza distribucional de la variable a modelar. Esta estrategia hace referencia a la segunda de las dos grandes alternativas apuntadas anteriormente, a saber, seleccionar aquel modelo que mejor se ajuste a las características de los datos.

En este sentido, es importante recordar que el objetivo del análisis de datos es, en palabras de Bartholomew (1995, p. 13), identificar «*las estructuras comunes que subyacen a diferencias superficiales*», es decir, capturar las invariancias de los datos para de esta forma poder inferir el mecanismo generador de datos subyacente (Lindsey, 1995a). Sin embargo, difícilmente se podrá inferir el mecanismo generador de unos datos cuya naturaleza ha sido modificada. Dos de los principales problemas con los que topa dicho proceso de inferencia es que, por un lado, no existe un conjunto de soluciones cerradas y únicas que especifiquen, a modo de algoritmo, los pasos a seguir ante cualquier patrón de datos para identificar su correspondiente mecanismo generador; y por otro lado, tampoco existe una herramienta estadística multipropósito aplicable a cualquier tipo de datos. No obstante, es imprescindible contar con a) un procedimiento de análisis de datos, así como b) herramientas específicas, es decir, modelos basados en distribuciones que representen adecuadamente los datos:

a) El procedimiento de análisis de datos debe ser lo suficientemente genérico para que sea posible adaptarlo a las características de los datos a analizar. Tal adaptación requiere que el análisis de datos se conciba, por parte del investigador, como un proceso:

   - activo, puesto que se trata de un proceso de toma de decisiones que está continuamente guiado por el investigador;
   - flexible, en el sentido de que la estrategia del análisis de datos basado en «*recetario de cocina*» (Judd y McClelland, 1989, p. v), restringe el patrón de datos al que es aplicable el análisis;
   - y recursivo, ya que se trata más bien de un proceso iterativo.

   Este es, de forma genérica, el concepto de análisis de datos en que se basa el proceso de *modelado estadístico*.

b) En cuanto a la disponibilidad de modelos adecuados, la introducción del modelo lineal generalizado (Nelder y Wedderburn, 1972), su posterior popularización (McCullagh y Nelder, 1989), y la progresiva incorporación al mismo de múltiples aportaciones procedentes de diferentes autores, ha hecho posible integrar en un mismo entorno inferencial una gran cantidad de modelos específicos para diversos tipos de datos.

El presente trabajo se basa en el modelado estadístico como procedimiento general para el análisis de datos y en el modelo lineal generalizado como marco teórico inferencial. Puesto que, tal como indican diversos autores (Hutcheson y Sofroniou, 1999; Lindsey, 1998; Winkelmann, 2000), este entorno teórico y de procedimiento son aún dos grandes ausentes en Ciencias Sociales y de la Salud en general, y en Psicología en particular, ambos conceptos son presentados en los primeros apartados del capítulo 3.

En la confluencia entre el modelo lineal generalizado y el estudio de las variables de recuento, se encuentran los modelos para las variables de recuento. De entre estos últimos destaca, entre otras cosas, por su papel como modelo de referencia en el estudio de las variables de recuento, el modelo de regresión de Poisson. En efecto, tal como indican Winkelmann y Zimmermann (1995, p. 2), «*en muchos sentidos el modelo de Poisson es tan relevante para los datos de recuento como los mínimos cuadrados ordinarios para los datos continuos: propiedades óptimas análogas al teorema de Gauss Markov pueden ser derivadas para el modelo de Poisson*». La principal bondad del modelo de regresión de Poisson es que es capaz de capturar la naturaleza discreta y no negativa de los datos de recuento, en especial cuando tales datos de recuento proceden de eventos raros. El modelo de regresión de Poisson se presenta en la segunda mitad del capítulo 3, junto con una discusión acerca de las características de la naturaleza de las variables de recuento y de su distribución de probabilidad de referencia, la distribución de Poisson.

Si bien es cierto que el modelo de regresión de Poisson presenta indudables mejoras con respecto al modelo lineal general, no es menos cierto el hecho de que puede resultar inapropiado en otros aspectos. En este sentido, tal como expone Winkelmann (2000), «*es común encontrar en el trabajo aplicado con datos de recuento* (...) *que ciertas asunciones del modelo de regresión de Poisson son sistemáticamente rechazadas por los datos*».

La limitación en la aplicabilidad del modelo de regresión Poisson proviene de la restrictividad impuesta por sus propios supuestos. La violación de tales supuestos, aunque de origen diverso (Winkelmann, 2000), tiende a un desenlace común, a saber, la ausencia de equidispersión (Cameron y Trivedi, 1998). A su vez, la violación del supuesto de equidispersión resulta en estimaciones ineficientes y en errores estándar sesgados (Winkelmann y Zimmermann, 1995). Por otra parte, es importante resaltar que, tal como indican diversos autores (Hauer, 2001; Long, 1997; McCullagh y Nelder, 1989; Mullahy, 1997; Yau y Lee, 2001), en la práctica la equidispersión es la excepción, y de sus posibles consecuencias, a saber, infradispersión y sobredispersión, es esta última la que adquiere un mayor el protagonismo debido a la frecuencia con que aparece. De esta forma, dada su importancia en los estudios de variables de recuento, el resto de contenidos de este trabajo giran entorno a la sobredispersión. Concretamente, en el capítulo 4 se exponen:

- los principales métodos para la detección y evaluación de la sobredispersión, así como,

- las soluciones al modelado en presencia de sobredispersión, tanto específicas –que implican la derivación de modelos específicos o extensiones para tratar causas identificadas de sobredispersión-, como inespecíficas –como son el uso de modelos más generales derivados de otras distribuciones, o métodos de estimación semiparamétricos.

Puesto que en la práctica no existe un consenso acerca de cuales son los criterios de aplicación de las pruebas diagnósticas de la sobredispersión, y además existen dudas sobre su correcto funcionamiento en algunas situaciones, en el capítulo 5 nos proponemos, como primer objetivo, realizar un estudio comparativo de las pruebas que habitualmente más 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 realiza a partir de los resultados de cuatro experimentos de simulación Monte Carlo en el entorno R que permiten comparar, concretamente, la tasa nominal de error y la potencia de estas pruebas diagnósticas.

El segundo objetivo que nos fijamos en la parte empírica de nuestro trabajo consiste en valorar los efectos de la sobredispersión sobre las estimaciones de los coeficientes y sus errores estándar obtenidos al ajustar los modelos de regresión de Poisson, binomial negativo y «quasi-Poisson», en las mismas situaciones definidas para la valoración de los tests diagnósticos de sobredispersión.

Por último, y puesto que tampoco existe un consenso claro en la literatura sobre cuál es el remedio más eficaz para tomar en cuenta la sobredispersión cuando no existe una hipótesis acerca de su causa probable, un quinto experimento de simulación Monte Carlo permite responder a nuestro tercer objetivo empírico: comparar la eficiencia y la precisión de los principales métodos para la corrección de la infraestimación del error estándar de los coeficientes de regresión que se obtienen con el modelo de regresión de Poisson en presencia de sobredispersión, a saber, el producto de los errores estándar por diferentes estimaciones del parámetro de dispersión ($\sqrt{(\chi^2/gl)}$, $\sqrt{(D/gl)}$ y $\sqrt{\varphi}$), y la estimación no paramétrica por remuestreo jackknife y bootstrap de dichos errores estándar.

Finalmente, en el último capítulo dedicado a las conclusiones generales de nuestro trabajo se sintetizan todos los resultados obtenidos, y se presentan las directrices de los próximos estudios que nos planteamos desarrollar en la línea iniciada con este trabajo.

2
Las variables de recuento en la investigación en Psicología

2.1 Presentación del estudio bibliométrico

El estudio bibliométrico que se presenta a continuación, que fue objeto de una comunicación en el VII Congreso de Metodología de las Ciencias Sociales y de la Salud (Vives y Losilla, 2001), constituye en realidad, la semilla de la que ha nacido el presente trabajo. Previamente a la realización de dicho estudio, habíamos observado de forma asistemática que en la literatura científica en Psicología había una cierta presencia –no sabíamos aún hasta que punto–, de investigaciones en las que la variable de respuesta era una variable de recuento. Sin embargo, no habíamos detectado la aplicación de análisis estadísticos basados en modelos específicos para variables de recuento como, por ejemplo, el modelo de regresión de Poisson (MRP en adelante), excepto en los casos en los que todas las variables medidas eran de naturaleza categórica y, por tanto, su análisis se realizaba a partir de tablas de contingencia y de modelos de regresión logit y/o modelos log-lineales.

Aunque el desarrollo del MRP no es reciente, sus propias limitaciones lo han puesto de actualidad a través del estudio de extensiones y generalizaciones del mismo, así como de modelos distribucionales alternativos que permitan relajar su restrictividad. La atención creciente que se está prestando a los modelos de recuento proviene principalmente, aunque no únicamente, del ámbito de la economía, concretamente de la econometría (Cameron y Trivedi, 1998; Winkelmann, 2000). Sin embargo, parece que la publicitación de los modelos de recuentos no encuentra el amplificador adecuado en diversos ámbitos; la Psicología es uno de ellos.

La falta de aplicación de modelos específicos para variables de recuento no es, sin embargo, exclusiva de la Psicología. Tal como ha sido reportado por diversos autores, resulta muy poco frecuente en la literatura científica de algunas disciplinas analizar las variables de recuento utilizando el MRP. Así, por

ejemplo, Lindsey (1998, p. 1746) afirma que «*es sorprendentemente común encontrar tales recuentos, aunque sean pequeños, tratados como variables continuas aplicando regresión lineal o análisis de la variancia*»; en esta misma línea, Long (1997, p. 217), constata que aunque «*existen diversos modelos que se ajustan explícitamente a las características de un recuento, …las variables de recuento son tratadas a menudo como si fueran continuas de forma que se aplica la regresión lineal*».

El conjunto de consideraciones expuestas anteriormente, han motivado la necesidad de conocer el estatus que ostentan las variables de recuento en la investigación en Psicología. La hipótesis de partida es que el poco uso que se hace de los modelos para datos de recuento en la literatura científica psicológica no es debido a su falta de aplicabilidad. Es decir, aunque en la investigación en Psicología se usan los recuentos como variable de respuesta, el modelo estadístico aplicado no es específico para datos de recuento. Concretamente, los objetivos planteados en el estudio bibliométrico son:

- Conocer la frecuencia de uso de variables de respuesta de tipo recuento en diferentes ámbitos de aplicación en Psicología, y
- Evaluar la adecuación del análisis estadístico aplicado a tales variables.

2.2 Procedimiento

Para poder identificar de forma unívoca la presencia de recuentos, se ha establecido una definición operativa de recuento, que se basa en la propuesta de Lindsey (1995b): un recuento se define como el número de eventos de una misma variable que ocurren al mismo sujeto o unidad de observación en un intervalo temporal o espacial definido.

La definición de variable de recuento es importante porque, entre otras cosas, permite diferenciarla de una variable de frecuencia. Las variables de frecuencia comparten ciertas características con las variables de recuento como el hecho de que sólo pueden tomar valores enteros positivos. Sin embargo, hablamos de «frecuencia» cuando los eventos ocurren de forma independiente en diferentes sujetos o unidades (ampliaremos esta distinción más adelante, en el apartado 3.2.1).

Puesto que la distribución de referencia en el ámbito de los recuentos, la distribución de Poisson, resulta especialmente adecuada para el modelado eventos raros (Winkelmann, 2000), se presenta también el análisis del uso de las variables de recuento en función de la frecuencia (alta /baja) del recuento. En términos de frecuencia porcentual de aparición del evento, y puesto que no existe un consenso claro en la literatura a este respecto, hemos utilizado el 15% como un valor criterio, que consideramos conservador, para clasificar los recuentos en las categorías de alta y baja frecuencia.

Para obtener la frecuencia de uso variables de respuesta de tipo recuento en Psicología, hemos revisado una muestra representativa compuesta por 40 artículos procedentes de los últimos números de 2 revistas seleccionadas al azar entre las 10 con mayor índice de impacto asignado por el ISI (Índice JCR-SCI, 1999) para cada ámbito de investigación en Psicología (según la clasificación en ámbitos que realiza este prestigioso instituto norteamericano para el análisis de la producción científica).

La revisión bibliográfica y análisis bibliométrico se centró en aquellos estudios cuyo objetivo era el pronóstico o la explicación de una variable de respuesta a partir de un conjunto de variables.

Hay que señalar que se excluyeron del análisis los siguientes ámbitos de investigación (de nuevo según la clasificación del ISI):

- Psicología Matemática, por la alta frecuencia de contenidos no aplicados o de modelado matemático (no estadístico).
- Psicología Aplicada, por la disparidad de sus contenidos y por la falta de uso de análisis estadísticos.
- Psicoanálisis, por ausencia de aplicación de análisis estadísticos.
- Psicología General, por contener en su mayor parte artículos de revisión.

De esta forma, una vez centrado el objetivo y tras aplicar los criterios de exclusión, la muestra inicial de 240 artículos quedó reducida a los 168 artículos la muestra final.

En la Tabla 1 se muestran las revistas seleccionadas dentro de cada ámbito de investigación establecido por el ISI, así como el factor de impacto de cada una de ellas.

Para responder a nuestros objetivos, para cada artículo revisado se registraron los siguientes indicadores:

- El tipo de variable de respuesta utilizada: recuento / no recuento.
- En caso de que la variable de respuesta fuera de recuento, se registró:
  - El modelo estadístico aplicado.
  - El grado de aparición del evento de interés en el recuento: alta frecuencia ($p > 0.15$) vs. baja frecuencia ($p \leq 0.15$).

Tabla 1. Revistas seleccionadas

Ámbito de investigación	Revista	Factor impacto
Psicología del desarrollo	▪ Development and Psychopathology ▪ Psychology and Aging	2.89 2.66
Psicología educacional	▪ Reading Research Quarterly ▪ School Psychology Review	1.95 1.10
Psicología social	▪ Journal of Personality and Social Psychology ▪ Journal of Experimental Social Psychology	2.72 1.74
Psicología clínica	▪ Journal of Consulting and Clinical Psychology ▪ Journal of Abnormal Psychology	3.92 3.17
Psicología experimental	▪ Journal of Experimental Psychology: General ▪ Cognition	4.18 3.39
Psicología biológica	▪ Psychophysiology ▪ Journal of Experimental Psychology: Animal Behavioural Processes	3.00 1.44

2.3 Resultados

En la Figura 1 se encuentran los diagramas de barras correspondientes a los porcentajes totales y por ámbito de variables de recuento. En primer lugar, es importante destacar que un gran número de los artículos revisados recogen como variable de respuesta una variable de recuento: el 38.1%.

En cuanto al uso de variables de recuento en los diferentes ámbitos de investigación psicológica observamos que dónde se utilizan con mayor frecuencia es en Psicología experimental y Psicología biológica. De hecho, son los únicos ámbitos en los que el porcentaje de variable de respuesta de recuento es superior a las de no recuento (57% y 61%, respectivamente). Cabe señalar que en Psicología biológica, el uso de variables de recuento presenta cierta variabilidad entre revistas. En general, son las revistas de temática conductual las que registran y analizan con mayor frecuencia variables de respuesta de recuento, mientras que en aquellas de cariz más fisiológico (dónde es más común el uso de medidas continuas como EEG, resistencia electrodermal, etc.) los recuentos son menos frecuentes.

Por otro lado, es en las investigaciones que se enmarcan dentro de las áreas de Psicología del desarrollo y Psicología educacional dónde hemos encontrado un menor uso de las variables de recuento (18% y 25%, respectivamente).

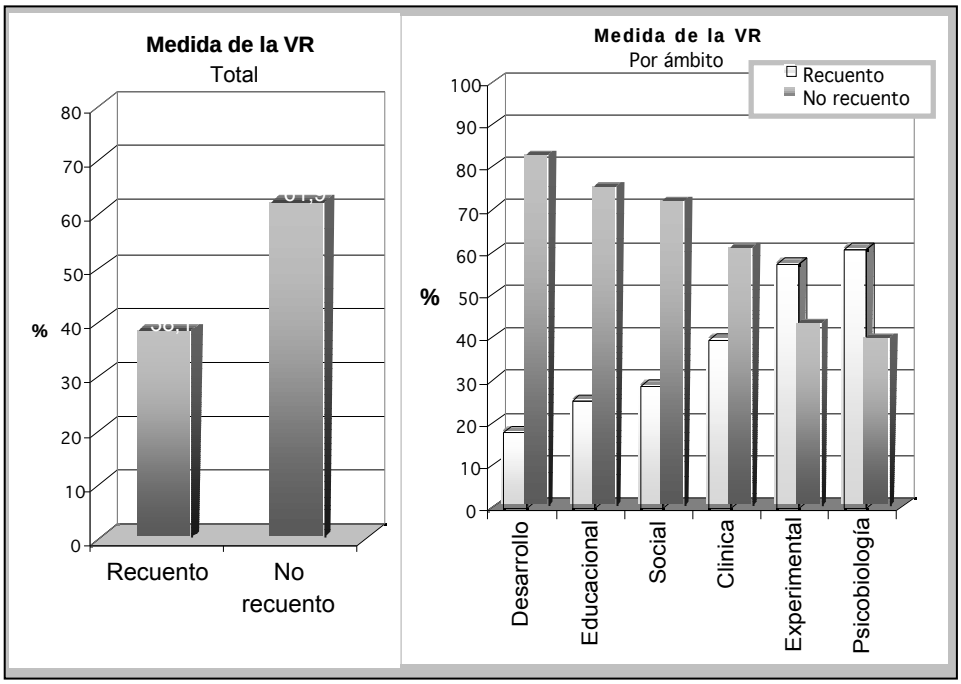

Figura 1. Porcentajes de variables de recuento

En segundo lugar, en las investigaciones cuya variable de respuesta es de recuento, e independientemente que éste sea de alta o baja frecuencia, en el 92% de los casos los modelos estadísticos que se aplican son el análisis de la variancia y la regresión lineal, tratando, por tanto, la variable registrada como si su escala de medida fuera continua en lugar de discreta. Es importante destacar que no hemos observado ni una sola investigación en la que se aplique el modelo de regresión de Poisson u otros modelos desarrollados específicamente para el análisis de datos de recuento.

Como es sabido, a medida que los valores de los recuentos aumentan, la distribución de Poisson, que es la distribución de referencia en el ámbito de los recuentos, converge a la distribución normal (Long, 1997), tal como apunta la Figura 2 (volveremos sobre esta convergencia en el apartado 3.2.2).

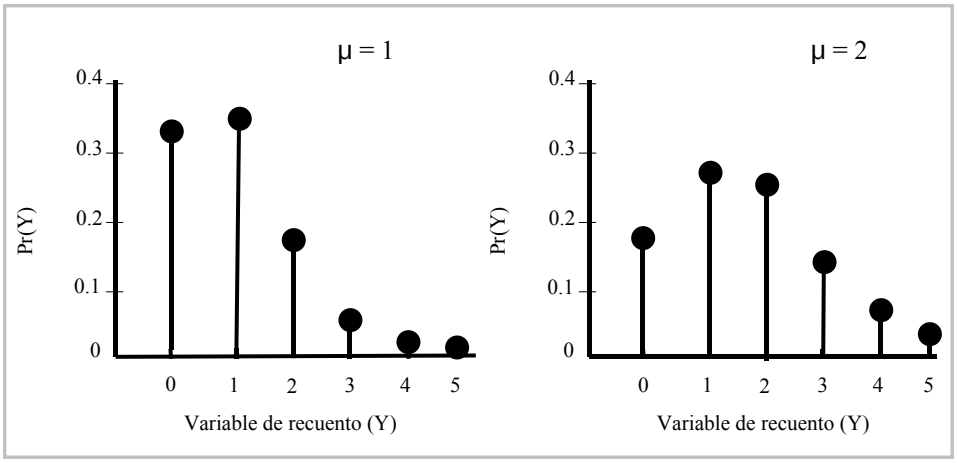

Figura 2. Aproximación de una distribución de Poisson a una distribución Normal. (Adaptada de Long, 1997)

De esta forma, teniendo en cuenta las condiciones que permiten la convergencia de la distribución de Poisson a la normal, una posible explicación del uso tan extendido del modelo lineal general podría ser que la mayor parte de los artículos revisados contienen una variable de respuesta de recuento de alta frecuencia. Pero la observación de los resultados representados en el diagrama situado a la izquierda de la Figura 3 advierte que la realidad no es esa: el 51.6% de los artículos registran al menos una variable de recuento con una frecuencia de ocurrencia baja, es decir, modelan eventos o suceso raros, mientras que el 48.4% restante trata con variables de recuento de alta frecuencia, en el sentido indicado antes, de forma que admitiría, en principio, aunque como se expondrá más adelante no está exento de problemas, que la distribución de probabilidad de dichas variables puede ajustarse a una distribución normal y en consecuencia, podría ser aplicable el modelo lineal general si se dan las condiciones adecuadas para estos modelos.

En algunas ocasiones nos ha resultado difícil clasificar una variable de recuento en función de la frecuencia de ocurrencia del evento medido, debido a que los pocos índices descriptivos mostrados en la mayor parte de artículos tratan sólo la distribución marginal y no a las distribuciones condicionadas a los niveles y/o categorías de las variables explicativas o pronósticas registradas, por lo que el porcentaje de recuentos de baja frecuencia de ocurrencia podría ser mayor que el que hemos indicado antes.

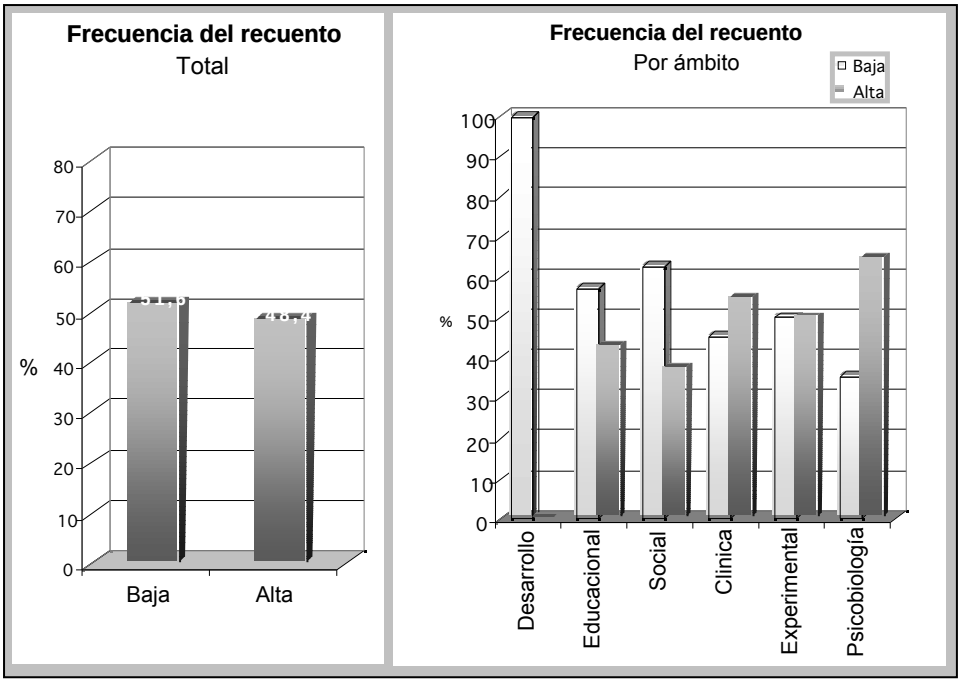

Figura 3. Frecuencia de los recuentos

Por otro lado, tal como se muestra en la Figura 3, los ámbitos dónde resulta menos frecuente el uso de variables de respuesta de recuento es precisamente donde éstas son mayoritariamente de baja frecuencia.

Un aspecto destacable que también hemos observado en nuestra revisión, es la práctica habitual de la transformación de la variable de respuesta de recuento mediante la aplicación de una raíz cuadrada, la función arcoseno, etc. En otros casos, se recodifica la escala de la variable de respuesta, como, por ejemplo, en un estudio predictivo acerca de la adherencia terapéutica (MacNaughton y Rodrigue, 2001), en el cual el terapeuta realiza una serie de recomendaciones a los pacientes y se recuenta el número de recomendaciones que realmente son seguidas por éstos. Así, en lugar de modelar el recuento de recomendaciones seguidas, se transforma el recuento en una variable binaria, con la consiguiente pérdida de información, y se aplica una regresión logística.

También es una práctica extendida la conversión del recuento en una proporción o porcentaje cuando para cada una de las unidades de observación dicho recuento se basa en períodos de tiempo, espacio, o tamaños muestrales distintos; una vez realizada la transformación, una vez más se analiza como una variable continua, en lugar de aplicar, por ejemplo, un MRP incluyendo el denominador de esas proporciones o porcentajes como variable de exposición («*offset*») (véase apartado 3.3.2).

En la misma línea de tratamiento indebido de las informaciones procedentes de recuentos, tampoco se tiene en cuenta en ningún caso la presencia excesiva de ceros ni su ausencia, situaciones estas que deberían conllevar la aplicación de modelos de análisis específicos desarrollados en el contexto del análisis de recuentos, como extensiones o alternativas distribucionales, y que serán presentados más adelante en este trabajo. Así, por ejemplo, una de las investigaciones revisadas tiene como objetivo la comparación de dos tratamientos contra el insomnio, para lo cual los autores seleccionan sujetos que presentan al menos 3 episodios de insomnio a la semana (Lichstein, Riedel, Wilson, Lester y Aguilard, 2001); siendo de una semana el período de registro de episodios de insomnio durante la investigación, resulta imposible el registro de un número de episodios inferior a 3, tratándose, por tanto, de un caso de truncamiento por la izquierda para el cual existen modelos de análisis específicos.

En el polo opuesto se encontraría el problema de la presencia excesiva de ceros en las medidas de recuento, como sucede, por ejemplo, en un estudio lingüístico acerca de la representación espacial y el lenguaje espacial en el que se observa que las distribuciones marginales de recuentos de términos axiales básicos presenta una alta frecuencia de ceros que sobrepasa el valor esperado en una distribución de Poisson (Munnich, Landau y Dosher, 2001).

2.4 Discusión

Los datos presentados en este estudio constituyen un indicador fiable del hecho que las variables de recuento son de uso frecuente en Psicología. Así pues, en nuestra disciplina adquiere una gran importancia el conocimiento acerca del tratamiento estadístico apropiado para este tipo de variables. Más aún cuando no se ha detectado un sólo análisis específico para datos de recuento. Lo cierto es que los datos presentados nos dan suficientes indicios para pensar que la afirmación que desde el ámbito de la medicina hace Lindsey (1998) es totalmente extrapolable a la Psicología: en presencia de variables de recuento el análisis estadístico más utilizado es el del modelo lineal general.

Tal como señala Long (1997) el uso del modelo de regresión lineal para recuentos puede producir estimaciones ineficientes, inconsistentes y sesgadas. Una solución habitual en estos casos consiste, tal como se ha indicado anteriormente, en realizar una transformación de los datos, como la raíz cuadrada. Sin embargo, tal como advierte Lindsey (1998), esta estrategia tiene al menos dos inconvenientes conocidos:

- El modelo resultante, en términos de, por ejemplo, raíz cuadrada de recuentos, es difícil de interpretar y de comunicar;
- La transformación aplicada no tiene, generalmente, un significado psicológico claro.

Además, como señalan McCullagh y Nelder (1989), en el contexto del análisis de una variable de recuento mediante el modelo lineal general se topa a menudo con la presencia de heterocedasticidad, relaciones no lineales entre la respuesta y las variables explicativas, y/o predicciones absurdas (valores fuera de rango, como por ejemplo, valores negativos); para cada uno de estos problemas se debería de aplicar una transformación diferente.

Siguiendo con la cuestión de las transformaciones de los datos, Long (1997) advierte que una vez el nivel de la variable de respuesta se ha determinado, es importante adecuar el modelo estadístico al nivel de medida. Si el modelo escogido parte de un nivel de medida erróneo, el estimador puede presentar diversos problemas o ser simplemente inapropiado.

En definitiva, y como una cuestión más de fondo, la vía de las transformaciones busca ajustar los datos al modelo, enfoque éste del análisis de datos que, tal como se apuntaba en el capítulo anterior, no compartimos.

La alternativa más adecuada al modelo de regresión lineal pasa por considerar la distribución de probabilidad de la variable de recuento, variable ésta de naturaleza discreta y no negativa que se ajusta, bajo ciertas condiciones a la distribución de Poisson. En consecuencia, el modelo más adecuado para evaluar el efecto de un conjunto de variables explicativas sobre la respuesta de recuento, será un modelo de datos de recuento, como puede ser el MRP o bien algunas de sus extensiones o generalizaciones. Se trata, en última de instancia, de adecuar el modelo estadístico al nivel de medida o, lo que es lo mismo, de ajustar el modelo a los datos y no a la inversa.

Las explicaciones tentativas al problema de la falta de aplicación de modelos específicos para datos de recuento pueden ser de diversa índole. A continuación, proponemos algunas:

- Los modelos de recuento, contienen relaciones no lineales entre las variables explicativas y la variable de respuesta, lo cual comporta cierta dificultad en su interpretación (Long, 1997).
- La naturaleza idiosincrásica de las variables de recuento no ha sido asumida por la mayor parte de ámbitos de investigación, de forma que éstas son tratadas como variables cuantitativas continuas.
- Se identifica correctamente a las variables de recuento pero se desconocen las implicaciones asociadas al uso de un modelo estadístico no pertinente como el modelo lineal general.
- Se desconoce la existencia de modelos estadísticos para modelar específicamente este tipo de variables, o se piensa que dichos modelos están todavía en una fase de desarrollo teórico y/o son muy difíciles de aplicar y de interpretar (Long, 1997).
- Se conoce la existencia de modelos estadísticos específicos para recuentos, pero se desconocen las herramientas informáticas que los implementan.

Relacionada con la última propuesta de hipótesis explicativa, creemos que hoy en día, y más aún en el ámbito aplicado, saber de la existencia de una técnica estadística es una condición necesaria pero no suficiente para que ésta sea empleada. Para su aplicación, el procedimiento estadístico debe estar implementado en paquetes estadísticos, a ser posible en aquellos de uso más extendido.

3
Modelado estadístico de respuestas de recuento

Un problema fundamental en la investigación científica radica en el hecho de que la forma en la que tratamos de resolver un problema afecta al tipo de resultados que observamos (Kuhn, 1962). En esta misma línea, Sturman (1999) señala, refiriéndose a la investigación en Ciencias Sociales, que el método estadístico usado para analizar los datos afecta al tipo de relaciones que somos capaces de detectar. Así pues, dada la importancia que adquiere la validez del enfoque a través del cual se analizan los datos, procederemos, antes de discutir los modelos para el análisis de datos de recuento, a situar el enfoque de análisis de datos en el que se circunscribe el presente trabajo. Esto es, el modelado como entorno epistemológico, el modelado estadístico como procedimiento general de análisis y el Modelo Lineal Generalizado (MLG, en adelante) como marco teórico inferencial.

3.1 La perspectiva del modelado estadístico

3.1.1 Modelado y modelado estadístico

Tal como señala Jáñez (1989), la búsqueda de la solución a un problema en una ciencia es una actividad que da como resultado un modelo formalizado del sistema real en unos casos y de una teoría en otros; teoría esta que para su verificación será concretada en un modelo. De esta forma, elaborar un modelo no es un fin sino un medio para aprehender mejor el fenómeno objeto de estudio (Losilla, 1995). Esta concepción instrumental del modelo queda claramente reflejada en la siguiente definición de Cortés y Martínez (1996): «*un modelo es un intermediario cuya utilidad o función se explica por la analogía que mantiene con aquello para lo que es intermediario*». De esta forma, es posible concebir un modelo como un interfaz o intermediario entre la teoría y la realidad que pretende explicar.

Aunque se considera que cualquier teoría constituye de por sí un modelo de la realidad, al mismo tiempo una construcción teórica utiliza los modelos como un

medio para acercar lo abstracto a lo concreto, lo conceptual a lo empírico. En este sentido, Anguera (1989, p. 564) sitúa el modelo «*… a mitad de camino de la explicación completamente teórica y los datos puramente empíricos*». Por su parte, Lunneborg (1994, p. 4) indica que «*modelar requiere un ir y volver entre lo que ya conocemos y lo que nuestros datos tienen que decirnos*».

El proceso de modelado no sólo resulta provechoso por su papel de mediador entre teoría y datos, sino que presenta, además, otras virtudes (Jáñez, 1989; Losilla, 1995; Lunneborg, 1994; Sejnowski, Koch y Churchland, 1990; Suárez, 1996):

- Utilidad para la construcción de teorías. Desde un punto de vista inductivo, el potencial de un modelo para la construcción de teorías es el resultado de, en palabras de Jáñez (1989, p. 36), «*colocar los huevos de la observación en el nido de la Matemática y dejar que ésta los incube*». En este sentido, los modelos ayudan a organizar los datos y a motivar la investigación, y sugieren cuál es el proceso empírico por el que los datos se ajustan a la explicación teórica de un fenómeno.
- Utilidad para la evaluación de teorías. Por supuesto, los científicos siempre parten de una teoría de la cual derivan hipótesis, y ello constituye el marco interpretativo de los resultados de sus investigaciones. Sin embargo, es frecuente que este marco interpretativo resulte un tanto vago en cuanto a su especificación. La contribución del modelado viene por la vía de su capacidad de detección de ambigüedades, falta de especificación o falta de coherencia interna de las teorías. De esta forma, la especificación de un modelo supone un paso decisivo hacia la contrastabilidad de la teoría en el que se enmarca.
- Valor heurístico de los modelos. Las inferencias generadas a partir del modelo pueden motivar nuevas hipótesis.
- Facilidad para la comparación de explicaciones alternativas de un fenómeno. A través del modelado es posible la comparación entre modelos que rivalizan en la explicación del fenómeno bajo estudio.

Aunque existen diversos criterios de clasificación de los modelos –en función del tipo de analogía que guarda con la realidad (formal / material), en función de la relación de equivalencia con el sistema modelado (isomorfo / homomorfo), etc.-, resulta de especial interés en el ámbito del modelado estadístico, la distinción entre modelos deterministas y no deterministas (Lindsey, 1997). Los modelos deterministas son aquellos con capacidad para dar cuenta –describir, predecir, explicar- de un fenómeno determinado, de forma que contienen información suficiente para representar sin error la parcela de realidad que modelan (Kleinbaum, Kupper y Muller, 1988; Reese, 1986). Por el contrario, los modelos probabilísticos o estocásticos tratan «*con la variabilidad y la incertidumbre*» (Bartholomew, 1995, p. 5) Estos modelos, asumen la existencia de un error resultante de la desviación entre el fenómeno observado y su representación mediante el modelo. Los modelos que incluyen un componente de error aleatorio

se denominan también modelos estadísticos (Ato, Losilla, Navarro, Palmer y Rodrigo, 2000a).

Un modelo estadístico se formaliza, generalmente, mediante una ecuación que expresa la relación entre las variables medidas y los supuestos sobre el proceso aleatorio que da lugar a los datos que se analizan (Rodrigo, 2000). Por otra parte, tal como indica Lunneborg (1994), el modelado estadístico se refiere al proceso por el que los modelos son construidos, evaluados y modificados. En esta misma línea, Judd y McClelland (1989) afirman que los procedimientos estadísticos pueden ser considerados como herramientas para la generación de modelos.

El enfoque del modelado estadístico como herramienta para el análisis de datos inició su auge aproximadamente a partir de 1980, gracias a la confluencia de diversos factores de entre los cuales destacan los nuevos desarrollos en la teoría estadística y la evolución en la capacidad de computación de los ordenadores, así como su popularización (Lunneborg, 1994).

El modelado estadístico es un proceso polietápico intrínsecamente iterativo, en el que se procede a reducir el error de forma progresiva, mediante la comparación de múltiples pares de modelos que pretenden dar cuenta de la misma realidad empírica. Los procedimientos para la comparación de modelos son de naturaleza inferencial y se centran en la significación estadística de algún índice indicador del aumento o la disminución del error al comparar dos modelos, evaluada mediante la técnica clásica del contraste de la hipótesis nula, de tal manera que cuando se comparan dos modelos, uno de ellos se corresponde con la formulación de la hipótesis nula y el otro con la hipótesis alternativa (Judd y McClelland, 1989; Krzanowski, 1998).

El objetivo básico del modelado estadístico es derivar, a partir de la variabilidad observada, un modelo donde la proporción de variabilidad sistemática sea relativamente grande respecto a la variabilidad aleatoria, esto es, un modelo que represente óptimamente la relación entre una variable de respuesta y un número de variables explicativas, minimizando el componente aleatorio. Así, mientras que la variabilidad observada responde a la expresión (Lindsey, 1995a):

$$\textit{Variable de respuesta} = \textit{componente sistemático} + \textit{componente aleatorio}$$

el modelo constituye una aproximación a los datos, de forma que (Judd y McClelland, 1989):

$$\textit{Modelo} = \textit{Datos} - \textit{Error}$$

El componente sistemático resume cómo la variabilidad en la respuesta es explicada por los valores de ciertas variables y es descrita, generalmente, mediante un modelo de regresión. Por su parte, el componente aleatorio o error constituye la variabilidad no explicada por el componente sistemático y describe,

mediante una distribución de probabilidad, la discrepancia entre la respuesta observada y la respuesta esperada o predicha por la parte sistemática del modelo.

El objetivo final que persigue el proceso de modelado estadístico, al igual que en el modelado de la mayor parte de sistemas, es el de encontrar el modelo «óptimo». El modelo óptimo es aquel que explica el máximo de variabilidad con el mínimo número de parámetros. Modelar la complejidad de un sistema o de un fenómeno pasa por llevar a cabo un proceso de simplificación, evitando de esta forma que el número excesivo de parámetros dificulte su análisis e interpretación (Judd y McClelland, 1989). El proceso de simplificación se rige por el principio de parsimonia, que se basa en el conocido postulado atribuido a Guillermo de Occam (1285-1347) denominado «navaja de Occam» («*Occam's razor*»), según el cual, a igualdad de condiciones es preferible un modelo simple a un modelo complejo. Así, el principio de parsimonia proporciona una guía importante para la simplificación de modelos, basada en la ponderación del criterio de máximo ajuste por el criterio de mínimo número de parámetros. Es decir, los efectos sistemáticos sólo deberían ser incluidos en el modelo si hay una evidencia convincente de su necesidad (Aitkin, Anderson, Francis y Hinde, 1989).

Sin embargo, tal como indican McCullagh y Nelder (1989), aunque se han propuesto estrategias dirigidas a obtener un balance óptimo entre la bondad de ajuste y la evitación de la complejidad innecesaria, difícilmente puede existir un patrón de toma de decisión puesto que el «*mejor modelo será un conjunto de alternativas buenas todas ellas y estadísticamente indistinguibles*» (op. cit., p. 23). En esta misma línea, Lindsey (1995a, p. 188) señala que «*no existe un modelo intrínsecamente verdadero*». Así, puesto que no existe un modelo válido único para una realidad concreta, sino un conjunto de modelos apropiados entre los cuales las diferencias de ajuste en relación a la realidad observada serán mínimas, el criterio para seleccionar uno de ellos no se basará en criterios exclusivamente estadísticos sino que la significación sustantiva –teórica, práctica, etc.– asumirá un papel relevante (Ato et al., 2000a).

3.1.2 El modelado estadístico en el entorno del MLG

Tal como indican Hutcheson y Sofroniu (1999) una de las mayores contribuciones en el ámbito de la estadística en el último cuarto del siglo pasado, ha sido la introducción por parte de J. Nelder y R. W. M. Wedderburn (1972) de los modelos lineales generalizados. Las dos grandes aportaciones del MLG son (Hutcheson y Sofroniou, 1999):

- La introducción del concepto de modelado estadístico como procedimiento general para el análisis de datos.
- La integración del modelado de datos categóricos y cuantitativos en un mismo entorno.

El MLG incluye, además de los modelos con componente aleatorio distribuido normalmente, aquellos cuyo componente aleatorio pertenece a la familia

exponencial de distribuciones. Al mismo tiempo, integra mecanismos para abordar la presencia tanto de relaciones lineales como no lineales, entre variables explicativas y variables de respuesta. De esta forma, el MLG integra como caso particular al modelo lineal general (Hutcheson y Sofroniou, 1999; Krzanowski, 1998).

El modelo lineal general se caracteriza por los siguientes aspectos:

- Los valores observados y_i son independientes y siguen una distribución normal con media μ y variancia σ^2 constante.
- Las variables explicativas proporcionan un conjunto de predictores lineales $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij}$.
- Los valores esperados y predichos se encuentran en la misma escala. Esto es, $\mu_i = \eta_i$

Existen básicamente dos situaciones en las que el modelo lineal general puede resultar insatisfactorio (Krzanowski, 1998):

- cuando la distribución de la variable de respuesta no es normal.
- cuando los valores esperados no presentan una correspondencia directa con el predictor lineal, sino que es una función del mismo. Esto es, $g(\mu_i) = \eta_i$

En presencia de estas situaciones es necesaria la aplicación de modelos lineales generalizados basados en distribuciones no normales (McCullagh y Nelder, 1989).

3.1.2.1 Supuestos básicos del modelo

Tal como indica Krzanowski (1998, p. 243) «*todos los modelos contienen asunciones, y es importante que dichas asunciones coincidan tanto como sea posible con las características físicas del problema*».

Un supuesto fundamental del MLG, de la misma forma que en el modelo lineal general, es la independencia de observaciones. Sin embargo, a diferencia del modelo lineal general, en los MLG con componente aleatorio no normal no se requiere el supuesto de homogeneidad de variancias. Es más, en algunos modelos como el modelo de regresión de Poisson o el modelo de regresión logística, la variancia de los datos viene determinada por el valor esperado μ, de forma que es de esperar que al variar μ varíe también la variancia. La distribución de Poisson o la de Bernoulli tienen un solo parámetro ajustado, la media μ, que determina la distribución entera. En cambio, la distribución Normal tiene dos parámetros ajustables, la media y la variancia, de forma que la media no determina por sí sola la distribución (McCullagh y Nelder, 1989).

Un problema que se da con cierta frecuencia en los modelos de regresión de Poisson o logit es que la relación media-variancia de los datos empíricos no se ajusta a la relación media-variancia que caracteriza la distribución subyacente teórica. De esta forma, cuando la variancia observada es mayor que la variancia

nominal, es decir, la variancia definida por la distribución de referencia, existe un problema de «sobredispersión» («*overdispersion*»), mientras que en el caso contrario, cuando el problema reside en que la variancia observada es inferior a la nominal, existe «infradispersión» («*underdispersion*»). Ambas situaciones serán tratadas con mayor detalle más adelante, en el apartado 3.3.6 y el capítulo 4.

3.1.2.2 Etapas del modelado

No parece haber un consenso acerca de la denominación y delimitación de las etapas que componen el proceso de modelado estadístico, aunque existen diversas propuestas (Box y Jenkins, 1976; McCullagh y Nelder, 1989). De entre ellas, destaca la realizada por Ato, Losilla, Navarro, Palmer y Rodrigo (2000a) por su aplicabilidad al entorno MLG y por constituir una síntesis comprehensiva de las propuestas antes citadas. De acuerdo con dicha propuesta, el proceso de modelado estadístico puede dividirse en las siguientes etapas:

1. Especificación
2. Selección
3. Evaluación
4. Interpretación

3.1.2.2.1 Especificación

En esta etapa, a partir de la teoría sustantiva y del análisis exploratorio de datos:

- se evalúan los supuestos del *componente aleatorio*,
- se establece la función del *componente sistemático*, y
- se determina cómo los dos componentes son combinados en el modelo mediante la *función de enlace*.

3.1.2.2.1.1 Componente aleatorio

El valor observado y_i, cuyo valor esperado viene dado por $E(y_i) = \mu_i$, es una realización de la variable aleatoria Y cuya distribución de probabilidad $f(y)$ forma parte de la familia exponencial de distribuciones. Algunos miembros de la familia exponencial son las distribuciones normal, binomial, Poisson, gamma o binomial negativa.

La distribución de una variable aleatoria Y, caracterizada por los parámetros θ y φ, pertenece a la familia exponencial si presenta la forma (Fahrmeir y Lang, 2001; Krzanowski, 1998; Lindsey, 1997; McCullagh y Nelder, 1989):

$$f(y, \theta, \varphi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y, \varphi)\right\} \qquad (3.1)$$

donde θ es el parámetro de localización o canónico, φ el parámetro de dispersión o de escala y $a(\varphi)$, $b(\theta)$ y $c(y,\varphi)$ son funciones conocidas (Rodríguez, 2002).

En cuanto a los dos primeros momentos de las distribuciones de la familia exponencial, se demuestra que (Rodríguez, 2002):

$$E(y) = \mu = b'(\theta) \tag{3.2}$$

$$\text{Var}(y) = \sigma^2 = b''(\theta)\, a(\varphi) \tag{3.3}$$

donde $b'(\theta)$ y $b''(\theta)$ son, respectivamente, la primera y la segunda derivadas de $b(\theta)$.

En relación a la variancia, puesto que μ depende de θ pero no de φ, podemos escribir la variancia como (McCullagh y Nelder, 1989)

$$\text{Var}(y) = a(\varphi)\, V(\mu) \tag{3.4}$$

donde $V(\cdot)$ se denomina *función de variancia*. Esta función captura la relación entre $E(y)$ y Var (y).

Por otro lado, a menudo $a(\varphi)$ tendrá la forma (McCullagh y Nelder, 1989):

$$a(\varphi) = \varphi / w \tag{3.5}$$

donde φ, también simbolizado por α y σ^2, es el denominado parámetro de dispersión y w un peso conocido que varía de observación en observación.

En la Tabla 2 se resumen los elementos principales que caracterizan a algunas de las distribuciones más utilizadas de la familia exponencial (Gill, 2001; McCullagh y Nelder, 1989).

Tabla 2. Elementos principales de diversas distribuciones de la familia exponencial

Distribución	Rango de Y	θ	$a(\varphi)$	$b(\theta)$	$V(\mu)$
Binomial B(m,p)	[0,n]	$\log\left(\frac{p}{1-p}\right)$	1	$m\_\log(1+\exp(\theta))$	$mp(1-p)$
Binomial negativa NB (μ, k)	Ent [0,∞)	$\log\left(\frac{\mu}{k+\mu}\right)$	1	$\frac{-\log[1-(\exp(\theta))]}{k}$	$\mu+k\mu^2$
Gamma G (μ,ν)	(0,∞)	$-1/\mu$	$1/\nu$	$-\log(-\theta)$	μ^2
Normal N (μ,σ^2)	(−∞,∞)	μ	σ^2	$\theta^2/2$	1
Poisson P(λ)	Ent [0,∞)	$\log(\lambda)$	1	$\exp(\theta)$	λ

3.1.2.2.1.2 Componente sistemático

$\theta_i = x_i\beta$ es el componente sistemático que recoge la variabilidad de Y explicada por X a partir de un combinación lineal de x_i que incluye los parámetros β. El componente sistemático, también denominado predictor lineal, se simboliza por η:

$$\eta_i = x_i\beta \tag{3.6}$$

Su estimación puntual a partir de una muestra es:

$$\hat{\eta}_i = x_i\hat{\beta} \tag{3.7}$$

El componente sistemático del MLG puede incluir términos tales como las variables explicativas originales, potencias y transformaciones de estas variables, interacciones entre las variables explicativas originales así como variables de control (Ato et al., 2000b).

3.1.2.2.1.3 Función de enlace

Tal como se indicaba anteriormente, en el modelo de regresión lineal $\eta_i = \mu_i$. Sin embargo, en otros modelos del MLG el valor predicho η_i y el valor esperado μ_i se encuentran en escalas de medida diferentes, de forma que $\eta_i \neq \mu$. En esta

situación, es necesaria la inclusión de una función que relacione los valores predichos y los esperados. Esta función se denomina función de enlace y se simboliza por $g(\mu_i)$.

La función de enlace que transforma el valor esperado a la escala del predictor lineal es:

$$g(\mu_i) = \eta_i = x_i\beta \tag{3.8}$$

La inversa de la función de enlace realiza el proceso inverso, es decir, transforma el predictor lineal η_i a la escala del valor esperado μ_i, que se halla en la escala de la variable de respuesta:

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i\beta) \tag{3.9}$$

La elección de la función de enlace no siempre resulta obvia (Aitkin et al., 1989; Czado y Munk, 2000). En este sentido, Krzanowski (1998) indica que pueden existir diferentes funciones de enlace aplicables a un problema particular, de forma que el investigador debe decidir cuál de ellas es la más apropiada en cada caso. Para simplificar el proceso, es habitual utilizar el parámetro canónico de la distribución como función de enlace, de forma que esta última recibe el nombre de «*función de enlace canónica*». Cuando se utiliza la función de enlace canónica se tiene que (McCullagh y Nelder, 1989):

$$\eta_i = \theta_i = x_i\beta \tag{3.10}$$

En la Tabla 3 se presentan las funciones de enlace canónicas, así como sus inversas, para las mismas distribuciones incluidas en la Tabla 2 (Fahrmeir y Lang, 2001; Gill, 2001; McCullagh y Nelder, 1989).

Tabla 3. Funciones de enlace canónicas.

Distribución	Función de enlace canónica $\theta = \eta = g(\mu_i)$		Inversa de la función de enlace canónica $\mu = g^{-1}(\eta_i)$
Binomial B(n,π)	Logit	$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$	$g^{-1}(\eta_i) = \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)$
Binomial negativa NB (μ, k)		$g(\mu_i) = \log\left(\frac{\mu}{k+\mu}\right)$	$g^{-1}(\eta_i) = \frac{\exp(\eta_i)}{k(1-\exp(\eta_i))}$
Gamma G (μ,ν)	Recíproca	$g(\mu_i) = -\frac{1}{\mu_i}$	$g^{-1}(\eta_i) = -\frac{1}{\eta_i}$
Normal N (μ,σ^2)	Identidad	$g(\mu_i) = \mu_i$	$g^{-1}(\eta_i) = \eta_i$
Poisson P(μ)	Logarítmica	$g(\mu_i) = \log(\mu_i)$	$g^{-1}(\eta_i) = \exp(\eta_i)$

Obsérvese que el modelo de regresión lineal constituye un caso particular en cuanto a la aplicación de la función de enlace debido a que, tal como se indicaba anteriormente, $\mu_i = \eta_i$ de forma que la función de enlace es la función de identidad $g(\mu_i) = \mu_i$.

Sin embargo, en el modelado, por ejemplo, de datos de recuento que siguen una distribución de Poisson, los valores esperados μ_i pertenecen al conjunto de números naturales, mientras que el predictor lineal η_i se encuentra en la escala de números reales. En este caso, difícilmente es aplicable la identidad como función de enlace. Los modelos para datos de recuento, caracterizados por la presencia de valores enteros positivos, implican la incorporación de efectos multiplicativos y esto es expresado mediante la función de enlace logarítmica (McCullagh y Nelder, 1989):

$$\log(\mu_i) = \eta_i = x_i\beta \tag{3.11}$$

Por otro lado, tal como se muestra en la Figura 4, la inversa de la función logarítmica, que es la función exponencial, permite obtener el valor esperado por el modelo μ_i, a partir del valor predicho η_i:

$$\mu_i = \exp(\eta_i) = \exp(x_i\beta) \tag{3.12}$$

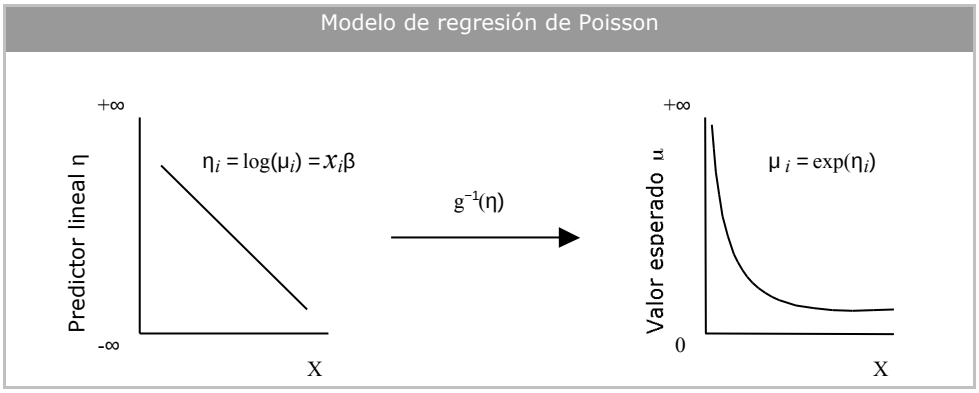

Figura 4. Transformación de valores predichos a valores esperados.
(Reproducida de Ato et al., 2000b, p. 8)

3.1.2.2.2 Selección

Tras la especificación de uno o varios modelos, se estiman, para cada modelo especificado, los parámetros del componente sistemático y se valora la precisión de las estimaciones a través del cálculo de la discrepancia entre pares de modelos, con el objetivo de seleccionar el modelo óptimo (McCullagh y Nelder, 1989).

En esta fase es importante recordar, tal como hace Lindsey (1995a, p. 188), que aunque es posible, gracias al uso de software estadístico, probar un gran número de modelos, el objetivo es conocer el «*mecanismo generador de datos, no sólo un valor P significativo*».

Por otro lado, Ato et al. (2000b) señalan que en esta fase es especialmente recomendable que las estimaciones se acompañen de medidas de precisión a través de intervalos de confianza de los parámetros del modelo, con el objetivo de clarificar su eficacia predictiva así como su interpretabilidad.

3.1.2.2.2.1 Estimación de los parámetros del modelo

Dos de las aproximaciones más comunes en la estimación estadística de parámetros son el método de Mínimos Cuadrados Ordinarios (*Ordinary Least Squares, OLS*) y el método de *Máxima Verosimilitud* (*Maximum likelihood, ML*). Mientras que el método de estimación de parámetros más utilizado en el modelo de regresión lineal es OLS, éste no resulta adecuado cuando el componente aleatorio del modelo es no normal, en cuyo caso, debe emplearse el método ML, que de hecho es el método aplicado en el MLG (Ato et al., 2000b; Dobson, 1990).

Una distinción importante entre las estimaciones OLS y ML, es que OLS se puede utilizar sin hacer hipótesis acerca de la distribución de la variable respuesta, mientras que para obtener estimaciones máximo-verosímiles se

necesita conocer la distribución de probabilidad conjunta de las *n* observaciones (Dobson, 1990).

Sea Y_1, ..., Y_n un conjunto de *n* observaciones aleatorias independientes cuya función de densidad de probabilidad $f_i(y_i;\theta)$ depende de un vector de parámetros θ. La función de densidad de probabilidad conjunta de *n* observaciones independientes $y = (y_1,...,y_n)'$ es (Long, 1997):

$$f(y;\theta) = \prod_{i=1}^{n} f_i(y_i;\theta) = L(\theta;y) \qquad (3.13)$$

La igualdad anterior, expresada como una función de los parámetros desconocidos θ dados los datos *y*, se denomina *función de verosimilitud.*

Nelder y Wedderburn (1972) demostraron que las estimaciones de máxima verosimilitud para todos los modelos incluidos en el MLG, podían ser obtenidas utilizando el algoritmo de *Mínimos cuadrados iterativamente ponderados* («*Iterative Weighted Least Squares*, IWLS»).

Las estimaciones máximo-verosímiles presentan las siguientes propiedades (Winkelmann, 2000):

- consistencia
- eficiencia asintótica
- distribución normal asintótica

3.1.2.2.2.2 Comparación de modelos

Una vez estimados los parámetros, se debe valorar la magnitud de la discrepancia entre los datos observados y los esperados por el modelo.

En el proceso de ajuste, se evalúan un conjunto de modelos que constituyen aproximaciones alternativas a los datos observados. Los modelos que pueden intervenir en las comparaciones son (Ato et al., 2000b; Lindsey, 1997):

- **Modelo saturado** (MS). En el MS el número de parámetros estimados por el modelo es igual al número de observaciones. En datos individuales, a diferencia de lo que ocurre en datos agrupados, el MS no constituye ninguna simplificación de los datos puesto que el número de observaciones es igual al tamaño de la muestra. De esta forma, la utilización de este modelo implicaría estimar un número de parámetros igual al tamaño muestral.
- **Modelo máximo** (MM). Constituye el modelo más complejo que puede ser formulado a partir de las variables explicativas registradas. Es decir, el MM puede incluir variables explicativas, variables de control y términos de interacción.

- **Modelo mínimo** (Mm). Consta del número de parámetros mínimo que, por motivos de diseño, deben ser estimados. Este aspecto está relacionado con el tipo de muestreo empleado. En este sentido y a modo de ejemplo, el modelado de diseños de grupos aleatorios con bloques requiere que cualquier modelo bajo estudio incluya el factor de bloqueo.
- **Modelo nulo** (MN). Es un modelo muy simple, habitualmente incapaz de representar adecuadamente la estructura de los datos, que se utiliza como modelo de referencia en el modelado de datos individuales. Este modelo contiene como único parámetro, el valor esperado μ, para todas las observaciones y representa, por tanto, un efecto nulo de las variables explicativas.
- **Modelo cero** (M0). Este modelo no estima ningún parámetro puesto que éstos son fijados por el investigador. El M0 es adecuado para realizar pruebas de conformidad sobre el valor de un parámetro.
- **Modelo de trabajo** (MT). Es el nombre que recibe el modelo que se está comparando en cada paso del proceso de ajuste y selección del modelo final. Consta de un número de parámetros superior al Mm e inferior al MM.

Tal como se ha expuesto anteriormente, el objetivo del proceso de modelado es la obtención de un modelo que sea capaz de representar los datos y, al mismo tiempo, de reducir la complejidad. Es decir, se trata de atender a los criterios de bondad de ajuste, parsimonia y significación sustantiva (McCullagh y Nelder, 1989). De esta forma, el modelo final deberá encontrarse en algún punto del continuo cuyos extremos son fijados por el modelo saturado y nulo: el modelo saturado asigna toda la variación al componente sistemático y ninguna al componente aleatorio, mientras que en el modelo nulo ocurre lo contrario (Hutcheson y Sofroniou, 1999).

En el contexto del MLG la selección del modelo se basa en el cálculo de la *discrepancia* (*D*) o *prueba de razón de verosimilitud* (esta prueba se trata con más detalle en el apartado 4.4.1.1). Aunque la discrepancia es una medida común a todos los modelos pertenecientes al MLG, se formula de formas diferentes según la distribución del componente aleatorio. En la Tabla 4 se muestra el estadístico discrepancia para las distribuciones Normal, Bernoulli, Binomial y Poisson. Cabe destacar que su utilidad e interpretación para valorar el ajuste de un modelo, difiere sustancialmente en función de si los datos son individuales o agrupados (Ato et al., 2000b).

Tabla 4. Estadístico discrepancia para las distribuciones Normal, Bernoulli, Binomial y Poisson

Modelo	Distribución	Discrepancia (D)
Regresión lineal	Normal	$\sum_i (y_i - m_i)^2$
Regresión logística	Bernoulli	$-2\sum_i \left[m_i \log\left(\frac{m_i}{1-m_i}\right) + \log(1-m_i) \right]$
Regresión logit	Binomial	$2\sum_i \left[y_i \log\left(\frac{y_i}{m_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - m_i}\right) \right]$
Regresión de Poisson Modelos log-lineales	Poisson	$-2\sum_{i=1}^{n} \left[y_i \log\left(\frac{m_i}{y_i}\right) + (y_i - m_i) \right]$

El proceso de selección del mejor modelo de entre un conjunto de modelos posibles se basa en la comparación entre pares de modelos anidados.

El concepto de «mejor modelo» depende de la finalidad que se persiga (Ato et al., 2000b). Cuando la finalidad del modelado es de tipo predictivo se seleccionan las variables que expliquen el mayor porcentaje de variabilidad de la respuesta, y para ello se emplean fundamentalmente criterios estadísticos. Por otro lado, cuando la finalidad es explicativa, son los argumentos teóricos los que deben tomar un mayor protagonismo en detrimento de los criterios estadísticos, por lo que el proceso de selección debe ser guiado por el investigador y se basa en la especificación de un modelo máximo inicial y de un conjunto sucesivo de modelos restringidos que se comparan mediante ajuste condicional. En este proceso es necesario respetar el principio de modelo jerárquico, que implica que si se incluye en el modelo un término de un orden determinado, todos los términos de orden inferior también se deberán incluir. A partir de ahí, el proceso de selección seguiría, a grandes rasgos los siguiente pasos (Kleinbaum et al., 1988):

1. Evaluar los términos de interacción a partir de su significación estadística.
2. Analizar la necesidad de mantener variables de control en el modelo. Haciendo uso de criterios de relevancia práctica más que en base a criterios estadísticos, se debe evaluar si:
   a. La eliminación de variables confundidoras o enmascaradas sesgará la estimación de los parámetros de interés o bien,
   b. en el caso de variables de ajuste, si su supresión del modelo implicaría una pérdida de precisión de las estimaciones.

3. Valorar si las variables explicativas de interés deben permanecer en el modelo. Para ello se deben emplear tanto criterios estadísticos como de relevancia substantiva.

3.1.2.2.3 Evaluación

La validez de las conclusiones extraídas de un modelo estadístico depende, a su vez, de la validez del modelo. Aunque se asume que el modelo puede no ser una representación exacta de la población, sí se requiere que reproduzca sin distorsiones las características principales de la misma (Aitkin et al., 1989). De esta forma, un examen pormenorizado de la correspondencia entre los datos y el modelo constituye una parte importante del modelado estadístico. La evaluación del modelo se centra, principalmente, en la evaluación de posibles errores de especificación (op. cit.):

- de la distribución de probabilidad del componente aleatorio,
- de la función que relaciona los componentes sistemático y aleatorio, esto es, la función de enlace y
- del componente sistemático del modelo, principalmente los errores de especificación relacionados con la selección de las variables explicativas.

Aunque será objeto de discusión posterior, es importante destacar que los errores de especificación mencionados arriba pueden provocar una violación del supuesto distribucional de relación media-variancia de diversos modelos lineales generalizados, como es el caso del modelo de regresión de Poisson (véase apartado 4.3).

3.1.2.2.4 Interpretación

Una vez se ha obtenido el modelo óptimo haciendo uso de los criterios antes mencionados de bondad de ajuste, parsimonia y significación sustantiva, el proceso de modelado se cierra con la interpretación del modelo.

Cabe recordar que la transformación provocada por la aplicación de una función de enlace (excepto en la función de enlace identidad del modelo lineal general) da lugar a una ecuación del modelo expresada en términos multiplicativos, en la que la interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado para un incremento unitario de X.

3.2 Eventos: recuentos y frecuencias

Los modelos para el estudio de eventos pueden clasificarse en función de la información que pretenden modelar, a saber:

- Presencia / ausencia de un evento

- Tiempo transcurrido hasta la presencia de un evento
- Recuento de eventos
- Frecuencias de eventos

A su vez, un evento puede ser clasificado en función de si es posible su agregación por unidad de observación en:

- Único: El evento sólo se da como máximo en una ocasión en una misma unidad de observación, por lo tanto, no es posible su agregación. La presencia de enfermedad crónica o la obtención del graduado escolar, son ejemplos de eventos únicos cuando la unidad de observación es el individuo.
- Múltiple: El evento se puede dar en más de una ocasión en una misma unidad de observación, de forma que es posible su agregación. Siendo el individuo la unidad de observación, son ejemplos de eventos múltiples, los episodios de insomnio o los accidentes laborales.

Es importante tener en cuenta que la unicidad o multiplicidad de un evento depende de las características tanto de la unidad de observación como de la unidad de análisis. Así, por ejemplo, mientras que la muerte es un evento de naturaleza única cuando la unidad de análisis es el individuo, esta se convierte en evento múltiple cuando el nivel de observación y/o análisis es una agrupación de individuos. Por otro lado, la conceptualización del evento de estudio puede modificar la naturaleza múltiple o única del mismo. Así, a modo de ejemplo, mientras que «el número de cigarrillos fumados» es un evento múltiple, «ser fumador» es un evento único. Teniendo en cuenta estas consideraciones, podemos seguir la propuesta de Lindsey (1998) de clasificar los modelos para el estudio de eventos en:

- **Modelos para el estudio de eventos únicos.** Tal como indica Lindsey (1998), los modelos más usados para el estudio de eventos únicos son la regresión logística y el análisis de supervivencia.

  En el caso de la regresión logística, el objetivo es el estudio de la presencia/ausencia de un evento de interés, con independencia del tiempo transcurrido hasta la ocurrencia del evento objeto de estudio.

  El modelo de regresión logística presenta los siguientes elementos.

  - Una variable dicotómica descrita mediante la proporción de observaciones que pertenecen a la categoría de referencia.
  - Un conjunto de predictores de naturaleza cuantitativa o categórica.

  En el modelo de supervivencia, también denominado análisis histórico de eventos («*event history analysis*») en el ámbito de las Ciencias Sociales, y análisis de duración («*duration analysis*») en Economía, no sólo se tiene en cuenta la ocurrencia de un evento, frecuentemente considerado inevitable, sino también el tiempo transcurrido hasta la ocurrencia del evento (Goldstein y Harrell, 1998; Lemeshow y Hosmer, 1998; Long, 1997). Concretamente, el

análisis de supervivencia presenta tres características principales (Rodríguez, 2002):

- Una variable de respuesta que es el tiempo transcurrido hasta la ocurrencia del evento objeto de estudio.
- La presencia de datos censurados procedentes de aquellas unidades de observación que tras finalizar el estudio no presentan el evento de interés.
- Un conjunto de variables, cuantitativas o categóricas, predictoras o explicativas del tiempo transcurrido hasta la ocurrencia del evento.

· **Modelos para el estudio de eventos múltiples.** El modelo de regresión de Poisson es el modelo de referencia en el ámbito de estudios de variables de recuento (Cameron y Trivedi, 1998; Winkelmann, 2000). Es un modelo que, debido a su distribución de probabilidad, resulta especialmente adecuado para modelar las ocurrencias no agregadas de eventos, esto es, valores enteros no negativos, en especial cuando la frecuencia de ocurrencia de los mismos es baja (Kleinbaum et al., 1988). Sin embargo, el modelo de regresión de Poisson es aplicable sólo en situaciones en las que se cumplen ciertos supuestos; supuestos estos, que restringen de forma considerable el ámbito de aplicación del modelo. La restrictividad exhibida por el modelo de regresión de Poisson ha contribuido al desarrollo tanto de modelos específicos, en ocasiones denominados extensiones, como de modelos más generales y menos restrictivos. De entre estos últimos destaca, por su flexibilidad y frecuencia de uso, el modelo de regresión binomial negativa. Puesto que el objetivo del presente trabajo se centra en el estudio del modelado de recuentos, los modelos mencionados anteriormente serán objeto de una exposición detallada en apartados posteriores.

  Si bien modelos como el MRP o el modelo de regresión binomial negativa son adecuados para variables de respuesta de tipo recuento, los modelos log-lineales y los modelos de regresión logit lo son para variables de respuesta de tipo frecuencia (véase el apartado 3.2.1 para una discusión acerca de la diferencia entre recuentos y frecuencias) y variables explicativas de tipo categórico. La diferencia entre ambos tipos de modelos, log-lineales y regresión logit, es el tipo de relación que pretenden modelar, a saber, simétrica y asimétrica, respectivamente. De todas formas, cabe señalar que la mayor parte de modelos de regresión logit pueden ser expresados como modelos log-lineales (Ato et al., 2000b).

  El modelado basado en la comparación de modelos log-lineales constituye un enfoque más general al análisis de datos categóricos que el centrado en el análisis de tablas de contingencia (Ato y López, 1996; Rodrigo, 2000)

3.2.1 Recuentos vs. frecuencias

Las variables de recuento son un tipo de información habitual en los estudios que se realizan en el ámbito de las Ciencias Sociales, del comportamiento y de la

salud, y se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo temporal o espacial definido. De esta definición se desprende, tal como se ha indicado anteriormente, que los sucesos a recontar corresponden necesariamente a fenómenos recurrentes, a partir de los cuales es posible obtener recuentos múltiples.

Son ejemplos de variables de recuento:

- Número de errores en una tarea de reconocimiento de palabras.
- Número de asignaturas suspendidas durante una licenciatura.
- Número de veces que un fumador ha intentado dejar de fumar.
- Número de accidentes laborales durante un período de tiempo.
- Número de recaídas en pacientes bulímicos tras la aplicación de un tratamiento.

Aunque las variables de recuento comparten ciertas características con las variables de frecuencia –como el hecho de que sólo pueden tomar valores enteros positivos, o la propia distribución del error- es importante diferenciarlas. En este sentido, y siguiendo las definiciones propuestas por Lindsey (1995b, p. 3), hablamos de *frecuencia* cuando los eventos ocurren de forma independiente en diferentes sujetos o unidades, mientras que el *recuento* se define como el número de eventos de una misma variable que ocurren en el mismo sujeto o unidad de observación. Así, por ejemplo, el número de artículos firmados durante el último año por un grupo de investigación de una facultad es un recuento, mientras que los grupos de investigación de una facultad que firman un número determinado de artículos es una frecuencia.

Obsérvese en el ejemplo representado en la Figura 5, que al pasar de una variable de recuento a una de frecuencia, cada valor de recuento pasa a ser una categoría de la cual se medirá su frecuencia.

Por otro lado, de la misma forma que ocurre en el ejemplo anterior, mientras que el valor 0 es bastante habitual en los recuentos, en el caso de las frecuencias es un valor menos común puesto que una frecuencia puede considerarse una agregación de unidades de observación en función de sus valores de recuento (Ato et al., 2000b).

Recuento	
X	Y
Grupo de investigación (Unidad de análisis)	Nº de artículos (Recuento)
A	1
B	4
C	0
D	2
E	0
F	1
G	1
I	3

Frecuencia	
X	y
Nº de artículos (Categorías)	Grupos de investigación (Frecuencia)
0	2
1	3
2	1
3	1
4	1

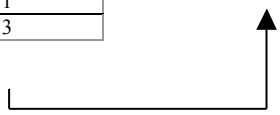

Figura 5. Datos de recuento y datos de frecuencia.

3.2.2 Distribución de Poisson

La distribución de Poisson debe su nombre al matemático Simeon Denis Poisson, quien publicó en 1837 (citado en King, 1988) un trabajo de investigación en el que presentaba una nueva distribución para el cálculo de probabilidades aplicado al ámbito penal.

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 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 siguen una distribución de probabilidad conocida: la distribución de Poisson. De hecho, tal como indica King (1988), habitualmente se asume que el mecanismo generador de datos que produce recuentos de eventos es, con independencia de su probabilidad de ocurrencia, Poisson.

La distribución de Poisson permite obtener la probabilidad de que se produzca un número determinado y_i de ocurrencias de un evento, y su función de probabilidad viene dada por (Winkelmann, 2000):

$$f(y_i, \lambda_i) = \Pr(Y = y_i \mid \lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \qquad (3.14)$$

$y_i = 0, 1, 2, \ldots, \infty$; λ_i $^{+}$,

donde $\lambda_i = \mu_i$ es el valor esperado de Y, E(Y)

Tal como se ha indicado anteriormente, el parámetro canónico es:

$$\theta_i = \log(\lambda_i) \tag{3.15}$$

mientras que las funciones conocidas son:

$$b(\theta_i) = \lambda_i = \exp(\theta_i) \tag{3.16}$$

$$a(\varphi) = 1 \tag{3.17}$$

$$c(y_i, \varphi) = -\log(y_i!) \tag{3.18}$$

Puesto que el parámetro canónico de la distribución de Poisson es $\log(\lambda_i)$, la función de enlace canónica viene dada por la función logarítmica:

$$\eta_i = \log(\lambda_i) \tag{3.19}$$

La Figura 6 muestra la distribución de Poisson para $m = 0.78$, $m = 1.49$, $m = 2.86$, $m = 5.47$, $m = 10.49$ y $m = 20.1$ (donde $m = \mu$) e ilustra algunas de las propiedades más importantes de la distribución de Poisson (Cameron y Trivedi, 1998; Long, 1997; Winkelmann, 2000):

- A medida que μ aumenta, la masa de la distribución se desplaza hacia la derecha. Específicamente:

$$E(y_i) = \exp(\theta_i) = \lambda_i = \mu_i \tag{3.20}$$

- La distribución de Poisson se caracteriza por la *equidispersión*, esto es:

$$\mathrm{Var}(y_i) = E(y_i) = \lambda_i = \mu_i \tag{3.21}$$

  La equidispersión constituye un supuesto básico de diversos modelos lineales generalizados. En contraste con otras distribuciones multiparámetro, como es el caso de la distribución normal, una violación de la asunción de la variancia es suficiente para violar el supuesto distribucional de Poisson (Winkelmann, 2000). Las desviaciones en relación a la equidispersión pueden resultar en:

  - *Sobredispersión*, si $\mathrm{Var}(y) > E(y_i)$
  - *Infradispersión*, si $\mathrm{Var}(y) < E(y_i)$.

- Las probabilidades son estrictamente decrecientes para $0 < \mu < 1$ y la moda es 0. Para $\mu > 1$ las probabilidades se incrementan hasta $y \leq \mathrm{ent}[\mu]$, y luego decrecen. Así, a medida que μ aumenta, la probabilidad de recuentos con valor cero disminuye. Cabe destacar, aunque se expondrá con más detalles en

apartados posteriores, que se da con cierta frecuencia que las variables de recuento presentan más ceros que los predichos por la distribución de Poisson.

- La distribución de Poisson tiende a la normal a medida que aumenta su media. Así, cuando el fenómeno de estudio no sea un evento raro y, por tanto, los valores recuento tengan una frecuencia elevada, la distribución de Poisson convergirá a la normal.

La distribución de Poisson es derivada a partir de un proceso estocástico $\{N(t), t \geq 0\}$, conocido como *proceso de Poisson*, en el que $N(t)$ representa el número de veces que ocurre un evento antes de un momento t.

Un proceso de Poisson presenta las características siguientes (Rodríguez, 2002):

- La probabilidad de al menos una ocurrencia de un evento en un intervalo temporal o espacial determinado, es proporcional a la amplitud de dicho intervalo. Así, tal como se detalla más adelante, es importante igualar o bien ponderar el intervalo de observación de los recuentos.
- La probabilidad de que se presenten dos o más ocurrencias de un evento en un intervalo muy reducido es aproximadamente 0.

Además, un proceso de Poisson se basa en dos asunciones críticas (Cameron y Trivedi, 1998; Winkelmann, 2000):

- Independencia entre eventos, esto es, la ocurrencia de un evento no afecta la ocurrencia futura de ese mismo evento.

  La independencia de un evento A en relación a otro evento B implica que:

$$\Pr(A \mid B) = \Pr(A \mid 1–B)$$

- Estabilidad de los eventos, es decir, la probabilidad de un evento es constante en el espacio o el tiempo.

La restrictividad de tales asunciones provoca que, en muchos ámbitos de las Ciencias Sociales o de las Ciencias de la Salud, no sea muy frecuente observar un proceso de Poisson puro. De esta forma, la aplicabilidad de la distribución de Poisson resulta restringida, lo cual hace patente, tal como se expone más adelante, la necesidad no tan sólo de la detección de procesos de recuento no poissonianos, sino también de técnicas y modelos capaces de dar cuenta de tales procesos (Trivedi, 1997).

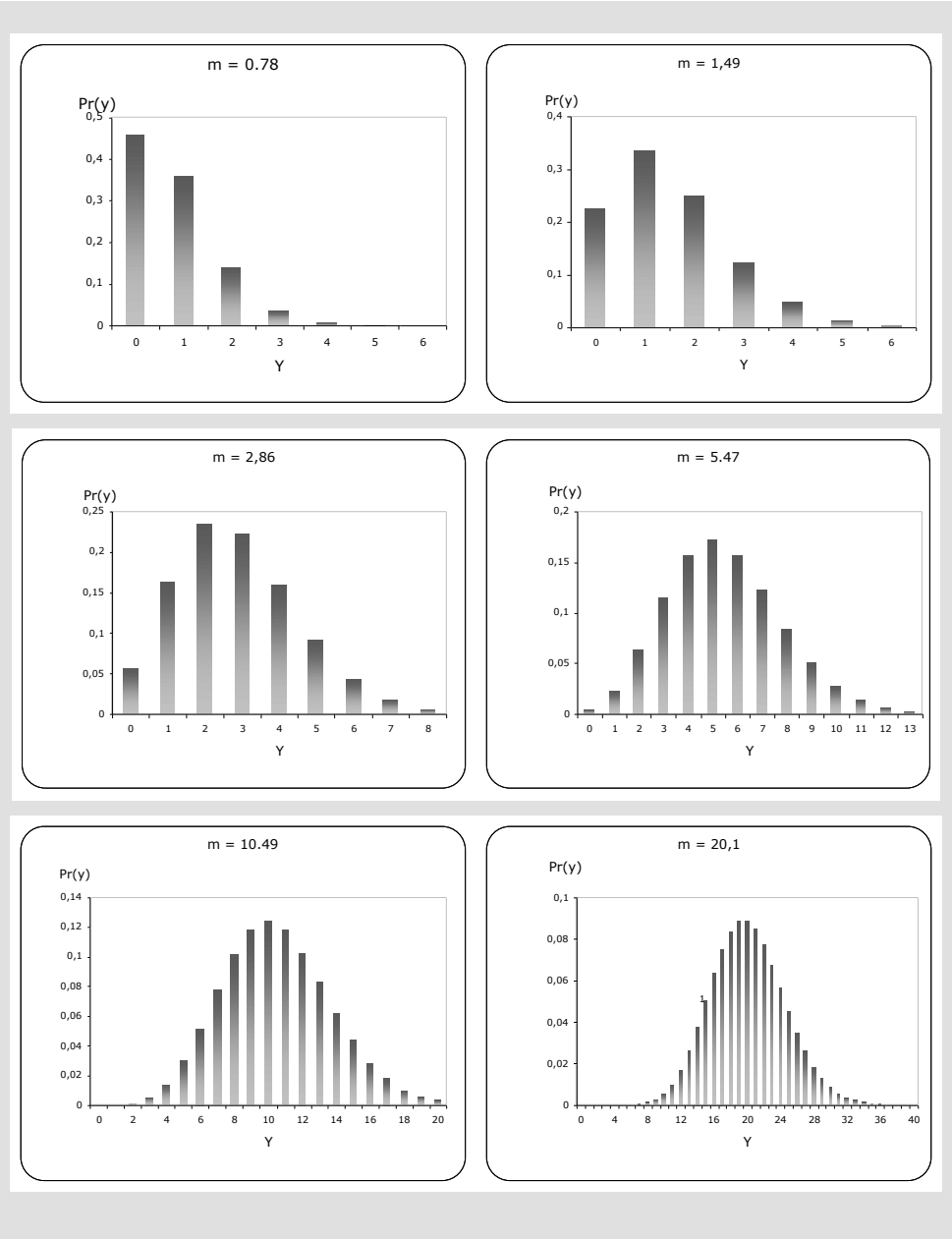

Figura 6. Convergencia de la distribución de Poisson a la distribución normal.

3.3 El modelo de regresión de Poisson (MRP)

Tal como se ha expuesto en apartados anteriores, los datos de recuento son frecuentes en Ciencias Sociales o Ciencias de la Salud, y Psicología no es una excepción. También se ha visto que resulta frecuente analizar los datos de recuento alterando ciertas características de tales datos para que se ajusten al modelo aplicado, habitualmente el modelo lineal general. Las estrategias que se usan con mayor frecuencia para ajustar los datos al modelo son (Sturman, 1999):

- **Agregación de recuentos**. Al aumentar el nivel de agregación de los recuentos se obtiene:
  - Un cambio en la forma de la distribución de los valores observados, puesto que, por un lado, existe un mayor rango de valores de recuento y por otro, disminuye la probabilidad de valores de recuento bajos, mientras que aumenta la probabilidad de valores de recuento altos.
  - Una «*mejora el poder explicativo de los modelos*», esto es, se obtiene una mayor magnitud del efecto (Sturman, 1999, p. 415).

  En relación a este último punto, cabe recordar que es importante prestar especial atención al tamaño del efecto, puesto que tiene un protagonismo destacado en el apartado de resultados del informe científico, ya que no tan solo es un dato requerido por diversas revistas, sino que tanto en la cuarta como en la quinta edición de su manual de estilo, la APA recomienda que se informe del tamaño del efecto (American Psychological Association, 1994; American Psychological Association, 2001).

  Sin embargo, la agregación sigue sin hacer posible la aplicación del modelo lineal general, puesto que el recuento sigue siendo una variable que puede tomar únicamente valores enteros positivos. Es más, no tan sólo no es capaz de solucionar ciertos problemas, sino que puede producir otros que desaconsejan su uso (Sturman, 1999):

  - Puede producir un truncamiento en la distribución. Cuanto mayor es el nivel de agregación más se desplaza la distribución hacia la derecha, de tal forma que si el nivel de agregación es muy elevado, puede resultar imposible la presencia de los valores de recuento más bajos.
  - Otra consecuencia indeseable del uso de la agregación es que puede desvirtuar o incluso confundir relaciones entre variables debido a la excesiva amplitud del rango espacial o temporal dentro del cual se llevan a cabo los recuentos. Por ejemplo, en el ámbito del absentismo laboral, en una situación en que las condiciones climatológicas del invierno pueden afectar la capacidad de los trabajadores para asistir al trabajo, si en lugar de realizar el recuento durante un mes se hace durante un año, es posible que los efectos del invierno enmascaren la relación objeto de estudio.

· **Transformaciones de los datos**.

En el caso de datos de recuento de baja frecuencia, Gardner, Mulvey y Shaw (1995), advierten que cualquier transformación monotónica de los datos no resuelve el hecho de que el valor modal se sigue situando en la parte baja de la distribución, de forma que el tamaño de la variancia difiere entre valores bajos y altos de recuento. Por otra parte, estos mismos autores (op. cit. p. 394) indican que la transformación de escala que se produce hace que los datos sean «*menos interpretables para el investigador y que los datos publicados sean inservibles para el meta-análisis*».

En datos de recuento, algunas de las transformaciones aplicadas más a menudo son:

- Transformación logarítmica. Los principales resultados de esta transformación son a) conversión de valores grandes en valores mucho más pequeños y b) valores estrictamente positivos. Sin embargo, no puede haber valores iguales a 0 puesto que el logaritmo está definido sólo para valores estrictamente positivos. Esta situación obliga a sumar una constante *c* a cada valor de recuento. Tal como indica King (1988), esta estrategia resulta insatisfactoria puesto que introduce un sesgo en la estimación que varia en función del valor escogido para la constante *c*.
- Transformación a través de la raíz cuadrada. Esta transformación puede parecer apropiada puesto que (King, 1988): a) la raíz cuadrada de 0 está definida de forma que no es necesaria la inclusión de una constate, como en el caso de la transformación logarítmica, y b) estabiliza la variancia de una variable de Poisson, evitando de esta forma el problema de la heterocedasticidad. Sin embargo, esta transformación no está exenta de problemas (op. cit): la variable transformada sigue siendo asimétrica y no es aproximadamente normal.

Anteriormente se ha señalado que el objetivo de las modificaciones en algunas propiedades de los datos como las mencionadas más arriba es, con frecuencia, poder aplicar el modelo lineal general. Tal como indican, entre otros, King (1988) y Long (1997), mientras que bajo ciertas condiciones el modelo de regresión lineal puede aproximar datos de recuento, el modelo de regresión de Poisson es el modelo indicado para este tipo de variable, puesto que una variable de recuento tiene unas características que hace que la utilización del modelo de regresión lineal pueda producir «*estimaciones ineficientes, inconsistentes y sesgadas*» (Long, 1997, p. 217). Es más, los resultados de los experimentos Monte Carlo llevados cabo por King (1988, p. 859) indican que la ineficiencia, la inconsistencia y el sesgo de las estimaciones permanecen «*incluso en muestras infinitas*».

Los principales problemas derivados de la aplicación del modelo lineal de regresión al modelado de una variable de respuesta recuento son (Ato et al., 2000b; King, 1988; Winkelmann, 2000):

- Violación del supuesto de heterocedasticidad: si partimos del hecho que una variable de recuento sigue la distribución de Poisson con media y variancia iguales, y puesto que, tal como se expone más adelante, el valor predicho de Y dado X es $\exp(x_i\beta)$, en un MRP la variancia condicional de Y depende de X según

$$\text{Var}(y_i \mid x_i) = E(y_i \mid x_i) = \exp(x_i\beta) \tag{3.22}$$

  lo cual implica que la variancia de los errores depende de X, y no es constante.

- Ausencia de normalidad: la distribución de Poisson tiende a la normal a medida que su media μ aumenta. Por tanto se violará el supuesto de normalidad cuando se modelen variables de respuesta con valores pequeños.

- Predicciones absurdas: La aplicación del modelo lineal de regresión puede predecir valores de recuento negativos. A este respecto, cabe recordar que la función de enlace logarítmica aplicada en el modelo de regresión de Poisson, garantiza que el valor del predictor lineal siempre sea positivo.

- No linealidad. Tal como señala King (1988, p. 846) el modelo de regresión lineal «*hace la asunción irrealista de que la diferencia entre la ocurrencia de 0 y 1 eventos en un intervalo temporal particular es la misma diferencia que entre, digamos, 20 y 21 eventos*».

3.3.1 Componentes del modelo

Los tres componentes del MRP son:

- Componente sistemático. El predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho es:

$$\eta_i = \log(\mu_\iota) = x_i\beta \tag{3.23}$$

- Componente aleatorio. La variabilidad de Y no explicada por η sigue una distribución de Poisson.

$$\varepsilon \quad \text{Poisson } (\mu) \tag{3.24}$$

- Función de enlace. La función que relaciona η con μ es:

$$g(\mu_i) = \log(\mu_i) \tag{3.25}$$

En el MRP, el recuento de eventos y_i sigue una distribución de Poisson con una media μ_i condicional que depende de los predictores de acuerdo con el siguiente modelo:

$$\mu_i = E(y_i \mid x_i) = \exp(x_i\beta) \tag{3.26}$$

Al aplicar la función exponencial, se consigue que $\mu_i \geq 0$, siendo esta una de las propiedades de la escala de medida de las variables de recuento (Long, 1997; Lunneborg, 1994).

La función $\mu_i = \exp(x_i\beta)$, que es la inversa de $\log(\mu_i) = x_i\beta$, se considera una expresión multiplicativa puesto que (Ato et al., 2000b):

$$\mu = \exp(x_i\beta) = \exp(\beta_0 + \beta_j X_j) = \exp(\beta_0) \times \exp(\beta_j X_j) \tag{3.27}$$

Esta propiedad es importante dado que las relaciones que pretende modelar el MRP son de tipo no lineal, de forma que posibilita la interpretación de los parámetros β_j directamente sobre la escala de la variable de respuesta.

3.3.2 La variable de exposición

En aquellos casos en que los recuentos de observaciones se basen en períodos de tiempo, tamaños poblacionales o espaciales no homogéneos entre los valores de las variables explicativas, es recomendable incluir en el modelo un término adicional: la *variable de exposición*, también denominada «*offset*», que se simboliza por *t* (Kleinbaum et al., 1988; Lunneborg, 1994; Winkelmann, 2000).

En realidad, la función de la distribución de Poisson que incluye la variable de exposición (Cameron y Trivedi, 1998)

$$\Pr(Y = y_i \mid \lambda_i) = \frac{\exp(-\lambda_i t_i)(\lambda_i t_i)^{y_i}}{y_i!} \tag{3.28}$$

puede considerarse como la expresión genérica de la distribución de Poisson, de forma que para el caso particular $t = 1$, la función de densidad es (3.14), esto es:

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

La ecuación del MRP que permite obtener los valores de recuento esperados, incorporando la variable de exposición es:

$$\mu_i = t_i \times \exp(x_i\beta) \tag{3.29}$$

donde t_i, es un vector columna que contiene los valores de la variable de exposición para cada unidad de observación y designa:

- la cantidad de tiempo durante el cual se ha llevado a cabo el recuento, o
- el tamaño del espacio en el cual se han obtenido los recuentos, o bien
- el tamaño de la población que proporciona los recuentos

Tomando los logaritmos de la expresión anterior, se obtiene el valor del predictor lineal:

$$\eta_i = \log(\mu_i) = \log(t_i) + x_i\beta \tag{3.30}$$

La variable de exposición modela únicamente los tamaños relativos de población, espacio o tiempo, de forma que afecta únicamente al valor de la constante del modelo (Lunneborg, 1994). En efecto, desarrollando y agrupando la expresión anterior obtenemos:

$$\log(\mu_i) = \log(t_i) + x_i\beta = \log(t_i) + \beta_0 + \beta_1 X_1 = [\log(t) + \beta_0] + \beta_1 X_1 \tag{3.31}$$

3.3.3 Estimación de parámetros

El método más utilizado para estimar el vector de parámetros β_j de un modelo con datos de recuento es el de máxima verosimilitud. El principio de máxima verosimilitud se basa en que los valores de los parámetros deben ser aquellos que maximicen la probabilidad de que el modelo especificado haya generado la muestra observada (Winkelmann, 2000). Otros métodos de estimación adecuados en presencia de sobredispersión, como son los semiparamétricos, serán expuestos en apartados posteriores.

La función de verosimilitud para el MRP es (Long, 1997):

$$L(\beta; Y, X) = \prod_{i=1}^{N} \Pr(y_i \mid \mu_i) = \prod_{i=1}^{N} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{3.32}$$

donde $\mu_i = \exp(x_i\beta)$.

Puesto que la transformación logarítmica es monotónica, la maximización de la función de verosimilitud es equivalente a la maximización de la función

logarítmica o función log-verosímil _ = ln$L$. En general, esta transformación simplifica los cálculos, al reemplazar los productos por sumas. Es más, permite el uso del teorema central del límite para estudiar las propiedades del estimador máximo-verosímil. La función log-verosímil para el MRP toma la forma (Cameron y Trivedi, 1998; Winkelmann, 2000):

$$\_(\beta;\, Y, X) = \sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_i'\beta - \ln(y_i!) \tag{3.33}$$

Por su parte $\hat{\beta}$, que contiene el valor que maximiza β, se obtiene a partir de:

$$\frac{\partial \ell(\beta; y, x)}{\partial \beta} = \sum_{i=1}^{n} \left[y_i - \exp(x_i'\beta)\right] x_i = 0 \tag{3.34}$$

La ecuación anterior es condición necesaria para obtener un máximo. Si la matriz de segundas derivadas, o matriz hessiana, evaluada en $\hat{\beta}$ es negativa, (3.34) es condición necesaria y suficiente para un máximo en la función log-verosímil. Finalmente, si la matriz de segundas derivadas es negativa para todos los valores de β, el máximo es global. La matriz hessiana de la función log-verosímil de Poisson viene dada por (Winkelmann, 2000):

$$H(\beta;\, y,\, x) = \frac{\partial^2 \ell(\beta; y, x)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i' \tag{3.35}$$

donde H es negativa, y la función log-verosímil del MRP es globalmente cóncava.

Puesto que (3.35) es no lineal en β, el sistema de k ecuaciones debe ser resuelto utilizando un algoritmo iterativo. Es habitual que se aplique el método de Newton-Raphson puesto que funciona bien en funciones cóncavas (Winkelmann, 2000).

El procedimiento iterativo finaliza cuando se satisface un determinado criterio de convergencia definido a priori. Los criterios de convergencia pueden ser (Winkelmann, 2000):

- el cambio en el valor de la estimación: $\hat{\beta}^{t+1} - \hat{\beta}^{t}$.
- el cambio en la función log-verosímil: $\_\hat{\beta}^{t+1} - \_\hat{\beta}^{t}$.
- el valor del gradiente en la estimación $g(\hat{\beta}^{t})$

La convergencia ocurre cuando cualquiera de estos valores, o alguna combinación de los mismos, es aproximadamente 0.

Cabe recordar que cualquier modelo lineal generalizado que utilice la función de enlace canónica, lo cual, como se ha indicado anteriormente, es el caso más frecuente, se encuentra en su forma canónica. En este caso, el algoritmo de Newton-Raphson es idéntico al de mínimos cuadrados iterativamente ponderados (Winkelmann, 2000).

3.3.4 Ajuste y selección del modelo

En cuanto al estudio de la significación estadística de los coeficientes del modelo, el contraste de la hipótesis H_0: $\beta_j = 0$ se puede llevar a cabo situando el valor resultante de la prueba (Ato et al., 2000b):

$$z = \frac{b_1}{EE(b_j)} \tag{3.36}$$

en la distribución normal (0,1).

Por otra parte, los límites del intervalo de confianza (*IC*) se hallan mediante la expresión:

$$IC(1-\alpha)\beta_j: b_j \pm z_{\alpha/2} EE(b_j) \tag{3.37}$$

Para obtener los límites del *IC* en su expresión multiplicativa, se exponencian los valores de los límites:

$$IC(1-\alpha)\beta_j: \exp[b_j \pm z_{\alpha/2} EE(b_j)] \tag{3.38}$$

Kleinbaum et al. (1988) proponen iniciar el proceso de ajuste y selección de un modelo de regresión múltiple con interacciones en base a un principio jerárquico según el cual, se evalúa en primer lugar las interacciones de mayor orden mediante pruebas de significación estadística, y se eliminan aquellas no significativas. Esta evaluación se puede llevar a cabo examinando la significación de cada una de las interacciones o por agrupaciones en función del orden al que pertenezcan.

El estudio del ajuste de un modelo se lleva a cabo mediante la prueba de la diferencia de discrepancias, cuya formulación es equivalente a la prueba de razón de verosimilitud.

En el MRP, la discrepancia viene dada por:

$$D = -2\sum_{i=1}^{n}\left(y_i \log\left(\frac{m_i}{y_i}\right) + (y_i - m_i)\right) \tag{3.39}$$

También es posible estudiar la significación de los parámetros a través de la comparación de modelos anidados. Esta comparación se lleva a cabo mediante la diferencia de discrepancias entre el modelo ampliado (MA) y el modelo reducido (MR) que, en el caso de la regresión simple será equivalente a la diferencia de discrepancias entre MA y MN (Ato et al., 2000b):

$$\Delta D = D_{MN} - D_{MA} \tag{3.40}$$

ΔD se distribuye según una distribución χ^2 con $gl = gl_{MN} - gl_{MA}$.

Puesto que en el contexto del MLG no existe, a excepción del modelo de regresión lineal, una definición ampliamente aceptada del índice de bondad de ajuste R^2, existen diversas propuestas denominadas genéricamente *pseudo*-R^2. Cameron y Trivedi (1998) proponen utilizar para el MRP, un índice que se puede interpretar como la proporción de reducción de la discrepancia del *MN* debida a la inclusión de las variables explicativas en el *MA*:

$$\text{pseudo-}R^2 = \frac{D_{MN} - D_{MA}}{D_{MN}} \tag{3.41}$$

3.3.5 Interpretación

Tal como indica Long (1997) la información que podemos extraer de la estimación de parámetros se puede clasificar en dos grandes tipos:

- El cálculo de los recuentos esperados m_i dados unos valores determinados en las variables explicativas X_j, es decir, el cálculo de los recuentos condicionales esperados, $\mu_i \mid x = x_i$.
- El cálculo de la probabilidad de un recuento y_i en función de los valores que tomen las variables explicativas, esto es, Pr $(y = y_i \mid x = x_i)$.

3.3.5.1 Cálculo de los recuentos condicionales esperados

El cambio en los valores de recuento *m* para unos valores determinados en las variables explicativas se puede evaluar de diversas formas (Long, 1997):

· *Factor de cambio*

En la expresión aditiva del MRP

$$\log(m_i) = x_i b \tag{3.42}$$

el valor del coeficiente b representa el cambio en el logaritmo del recuento esperado por cada unidad de cambio en la variable explicativa X manteniendo constantes el resto de variables explicativas, y este cambio es lineal, es decir, constante entre valores de X equidistantes.

La interpretación de los coeficientes de un modelo de regresión de Poisson a partir de la ecuación del modelo en su forma multiplicativa

$$m_i = \exp(x_i b) \tag{3.43}$$

conlleva una serie de ventajas (Ato et al., 2000b; Long, 1997):

- es más sencilla: al exponenciar el coeficiente de regresión asociado a X, se obtiene el factor de cambio, esto es, el valor por el cual se multiplica el recuento esperado por un incremento unitario en X, manteniendo constante todas las demás variables explicativas.

  Más allá del incremento unitario en los valores de X, el factor de cambio de m al pasar de x_k a $x_k+\delta$, viene dado por (Long, 1997):

$$\frac{E(y \mid x, x_j + \delta)}{E(y \mid x, x_j)} = \frac{\exp(b_0)\exp(b_1 x_1)...\exp(b_j x_j)\exp(b_j \delta)...\exp(b_J x_J)}{\exp(b_0)\exp(b_1 x_1)...\exp(b_j x_j)...\exp(b_J x_J)}$$
$$= \exp(b_j \delta) \tag{3.44}$$

  Así, los parámetros pueden ser interpretados de la siguiente forma:

  Para un cambio de δ unidades en x_j, el recuento esperado se incrementa en un factor de $\exp(b_j\delta)$ manteniendo el resto de variables constantes.

- obtenemos relaciones no lineales –las más frecuentes entre los estudios en los que intervienen variables explicativas de tipo recuento-, es decir, no se trata de un cambio constante entre valores X equidistantes.
- los valores esperados de recuento se encuentran en la misma escala de medida que la variable de respuesta.

· *Porcentaje de cambio*

También puede aportar una información valiosa el poder interpretar el factor de cambio como porcentaje de cambio, valor que se obtiene mediante la expresión:

$$[\exp(b_j\delta)-1]\times 100 \tag{3.45}$$

· *Cambio discreto*

El efecto de una variable puede ser evaluado calculando el cambio discreto en el valor esperado de y para un cambio en x_j, desde x_P hasta x_F:

$$\frac{\Delta E(y \mid x)}{\Delta x_j} = E(y \mid x, x_j = x_F) - E(y \mid x, x_j = x_P) \tag{3.46}$$

El cambio discreto puede ser interpretado como el cambio en el valor esperado de recuento al pasar x_j de x_P a x_F.

3.3.5.2 Cálculo de la probabilidad de un recuento

Los parámetros también pueden ser utilizados para calcular la distribución de probabilidad de los recuentos y_i para unos valores determinados de las variables explicativas. De acuerdo con (3.43) y una vez obtenidos los parámetros β_j a través de sus estimaciones b_j, podemos obtener para cada valor de las variables explicativas un valor de recuento esperado m_i. Se trata ahora de calcular la probabilidad de obtener un recuento y_i si $m = m_i$, o lo que es lo mismo, puesto que m_i depende por los valores de la variables explicativas X_j, queremos obtener la probabilidad $\Pr(y = y_i \mid x = x_i)$. Dicha probabilidad viene determinada por:

$$\Pr(y = y_i \mid x = x_i) = \frac{\exp(-m)m^{y_i}}{y_i!} \quad \text{para } y = 0,1,2,3\ldots\infty;\ m_i \geq 0 \tag{3.47}$$

Siendo $m_i = \exp(x_i b)$.

Por otra parte, el promedio de la probabilidad predicha para cada valor de recuento y_i puede ser utilizada para resumir las predicciones del modelo a partir de la expresión (Long, 1997):

$$\overline{\Pr}(y = y_i \mid x = x_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{\exp(-m)m^{y_i}}{y_i!} \tag{3.48}$$

El intervalo de confianza de la predicción de un evento «*proporciona al investigador una estimación de la precisión para la predicción, de forma que la interpretación de los resultados de un MLG son más informativos*» (Liao, 2000, p. 149).

Utilizando los límites del *IC* del predictor lineal en la función de probabilidad de la distribución de Poisson, se puede obtener el *IC* de la probabilidad predicha de un recuento y_i en función de unos valores determinados en las variables explicativas (Liao, 2000). Siguiendo la propuesta de Liao (2000), el *IC* de la probabilidad predicha de un valor de recuento y_i dado un conjunto específico X_0 de valores de las variables explicativas se obtiene a partir de la expresión:

$$IC[\Pr(y = y_i \mid x = x_i)] = \frac{\exp\left[-\exp\{\eta_{0i} \pm z_{\alpha/2} EE(e_{0i})\}\right] \times \left[\exp\{\eta_{0i} \pm z_{\alpha/2} EE(e_{0i})\}\right]^{y_i}}{y_i!} \quad (3.49)$$

donde y_i es el valor de recuento para el que se ha obtenido la probabilidad predicha, η_{0i} es el valor del predictor lineal para el conjunto específico X_0 de valores de las variables explicativas, y $EE(e_{oi})$ es el error estándar del error de predicción, que coincide con el *EE* del número esperado de recuentos $EE(m_i)$. Con el objetivo de evitar cálculos analíticos complejos, el $EE(m_i)$ se obtiene, de la misma forma que en el cálculo del *IC*, a partir de $EE(b_0)$ en el modelo «desviado» (Ato et al., 2000b).

3.3.6 Equidispersión

Tal como indica Krzanowski (1998), la distribución normal está completamente determinada por sus dos primeros momentos, la media μ y la variancia σ^2. Así, los modelos con variable de respuesta normal incluyen un parámetro variancia, que puede ser estimado a partir de los datos, de tal forma que el modelo puede acomodar cualquier grado de variabilidad en relación a la media.

Sin embargo, en la distribución de Poisson, como en otras distribuciones no normales, no existe un parámetro variancia independiente, sino que la media de la distribución determina a la variancia, a través de algún tipo de relación entre ambas. En la distribución de Poisson la relación entre media y variancia se caracteriza por la equidispersión, situación ya definida en (3.22).

Cuando la distribución de recuentos no presenta esta relación de identidad media-variancia, la igualdad (3.22) pasa a ser, tal como se ha apuntado anteriormente, una de las dos desigualdades siguientes:

- $\mathrm{Var}(y_i \mid x_i) > E(y_i \mid x_i)$, situación que define a la sobredispersión.
- $\mathrm{Var}(y_i \mid x_i) < E(y_i \mid x_i)$, es decir, infradispersión.

McCullagh y Nelder (1989) advierten que, en la práctica, el cumplimiento del supuesto de equidispersión, esto es, presencia de variancia nominal, es más la excepción que la norma. Tal como señalan Krzanowski (1998) y Winkelmann (2000), en ausencia de equidispersión, es mucho más frecuente una situación de sobredispersión que de infradispersión. De hecho, las pruebas para evaluar equidispersión son denominadas habitualmente pruebas de sobredispersión.

Cuando el modelo usado para analizar datos de recuento no es capaz de dar cuenta del exceso de variancia en los datos, las estimaciones de los errores estándar pueden resultar sesgadas, pudiendo comportar errores en las inferencias realizadas a partir de los parámetros del modelo de regresión (Krzanowski, 1998).

En esta situación, resulta imprescindible el poder disponer tanto de pruebas para el diagnóstico de la sobredispersión, como de procedimientos para poder abordarla.

En cuanto al diagnóstico, Lindsey (1995b) propone aplicar el coeficiente de variación como indicador como para evaluar el supuesto de equidispersión. El coeficiente de variación, un índice no específico de la evaluación de la sobredispersión, se define como la razón entre la variancia estimada y la media estimada:

$$\frac{\mathrm{Var}(\mu_i)}{\mu_i} \tag{3.50}$$

Puesto que teóricamente, $\mathrm{Var}(\mu_i)= \mu$, el coeficiente de dispersión debería ser igual a 1. Así, desviaciones respecto a 1 indican que posiblemente la distribución no sea Poisson (Lindsey, 1995b). Concretamente, si el coeficiente de dispersión es mayor que 1 posiblemente exista sobredispersión. Por otro lado, un coeficiente menor que 1 es indicador de infradispersión.

Lógicamente, la presencia tanto de sobredispersión como de infradispersión dependerá de la magnitud del valor del coeficiente de dispersión. En este sentido, Cameron y Trivedi (1998, p. 77) afirman que «*si la variancia muestral es más de dos veces la media muestral, entonces posiblemente los datos presenten sobredispersión después de la inclusión de los regresores. Esto es particularmente cierto para datos transversales, para los que los regresores explican habitualmente menos de la mitad de la variación en los datos*»

Este sencillo índice, constituye sólo una simple aproximación a la detección de sobredispersión. Es necesario el uso de otras pruebas que proporcionan un criterio estadístico acerca de la presencia de sobredispersión

En cuanto al tratamiento de la sobredispersión, existen tres grandes estrategias, basadas en:

- Modelos basados en distribuciones mixtas. De entre ellos, el más habitual es el MRBN (véase el apartado 4.6.1), cuyas caracterizaciones principales son:

- La función variancia Negbin I: $\mathrm{Var}(y_i \mid x_i) = \mu_i + \alpha\mu_i$
- La función variancia Negbin II: $\mathrm{Var}(y_i \mid x_i) = \mu_i + \alpha\mu_i^2$

· Modelos específicos, conocidos como extensiones, que modifican el modelo de datos de recuento de referencia para incorporar la sobredispersión debida a un error de especificación determinado.

· Modelos con variancia generalizada con alternativas paramétricas y semiparamétricas. Esta última es, de las dos alternativas, la más utilizada y se caracteriza por asumir una función variancia del tipo $\mathrm{Var}(y_i \mid x_i) = \varphi\mu_i$, donde el parámetro de escala $\varphi = \sigma^2$ debe ser estimado a través de pseudo máxima-verosimilitud.

Una exposición detallada de las pruebas para la detección de la sobredispersión, así como su tratamiento a través de las diferentes alternativas de modelado, se hallan en el siguiente capítulo.

4
La sobredispersión en los modelos de recuentos

4.1 Introducción

El MRP presenta, como se ha señalado en otras ocasiones, un ámbito de aplicación restringido debido a la propia restrictividad del modelo. Tal restrictividad es el resultado de los supuestos de la distribución de Poisson, a saber:

- *Independencia*: los eventos ocurren independientemente a través del tiempo. Es decir, la ocurrencia de un evento no afecta la ocurrencia futura de ese mismo evento.
- *Estabilidad*: los eventos son estables o estacionarios, es decir, la probabilidad de un evento es constante en el espacio o el tiempo.

La violación de cualquiera de estos dos supuestos distribucionales implica, habitualmente, una violación del supuesto de equidispersión. Sin embargo, la ausencia de equidispersión no está originada únicamente por el incumplimiento de las asunciones anteriores, sino que puede ser el resultado de otras situaciones o procesos que describiremos más adelante en este capítulo. En cualquier caso, cuando los eventos objeto de estudio presentan algún tipo de dependencia o bien no son estacionarios, el mecanismo generador de datos no es poissoniano.

4.2 Procesos no poissonianos

A continuación se exponen los diversos tipos de procesos no poissonianos a través de un ejemplo adaptado de Cameron y Trivedi (1998) y Winkelmann (2000).

Supóngase que realizamos n extracciones aleatorias de m urnas, cada una de las cuales contiene a bolas rojas y b bolas azules. Las bolas rojas simbolizan la

ocurrencia de un evento (éxito), mientras que las bolas azules representan la no ocurrencia.

Si la composición de las urnas no cambia en el tiempo (extracción con reposición), los ensayos son independientes, es decir, los resultados de los experimentos en diferentes puntos del tiempo son independientes, y la probabilidad marginal de un suceso es constante a través del tiempo e igual a la proporción de bolas rojas. De esta forma, la probabilidad total de ocurrencia de un evento viene dada por $\Pr(Y = a) = a/(a+b)$.

4.2.1 Ausencia de independencia

Si la composición de las urnas es inicialmente homogénea pero es modificada como consecuencia de la presencia o ausencia de eventos previos entre los momentos t y $t+\Delta$, entonces $\Pr(Y = a)_t \neq \Pr(Y = a)_{t+\Delta}$. Dos son las causas que desembocan en una situación de dependencia:

· **Dependencia de ocurrencia**

La composición cambia como consecuencia de éxitos previos. Esta situación se denomina *dependencia de ocurrencia*. Esta «*dependencia dinámica entre la ocurrencia de eventos sucesivos*», tal como la describen Cameron y Trivedi (1986, p. 31), también se denomina *contagio*, así como *dependencia de estado*. En el ámbito de las Ciencias de la Salud, se habla de modelo de contagio cuando la presencia de un episodio (evento) en un individuo modifica la probabilidad de ocurrencia de futuros episodios, esto es, episodios dependientes. En este caso existe una «dependencia intrapaciente» Navarro et al. (2001).

La dependencia de ocurrencia o contagio puede ser:

- positivo, si la ocurrencia de un evento incrementa la probabilidad de eventos posteriores, o bien,
- negativo, si la ocurrencia un de evento disminuye la probabilidad de eventos posteriores.

Retomando el ejemplo anterior, inicialmente hay a bolas rojas y b bolas azules en las urnas. Se extrae una bola aleatoriamente. Si la bola es roja (éxito), ésta es reemplazada juntamente con s bolas rojas. Si la bola es azul, la proporción $a/(a+b)$ permanece inalterada puesto que la bola azul es reemplazada. Si este procedimiento es repetido n veces y X representa el número total de veces que se saca una bola roja, entonces X sigue una distribución Pòlya-Eggenberger. Si el número de bolas rojas se incrementa después de un éxito ($s > 0$), entonces una ocurrencia incrementa la posibilidad de futuras ocurrencias y el modelo de la urna refleja un contagio (verdadero) positivo y se ajusta a una distribución binomial negativa (Cameron y Trivedi, 1998). Para $s = 0$ el modelo se reduce al modelo binomial con ensayos independientes. Para $s = -1$, el modelo de la urna se

corresponde con la extracción sin reemplazamiento, y la distribución resultante es la distribución hipergeométrica. Así, la distribución hipergeométrica es una distribución para el contagio negativo (Winkelmann, 2000).

Xekalaki (1983) identifica otra forma de dependencia de ocurrencia que es inherente a la concepción de que los eventos ocurren en «rachas» o períodos («*spells*») y que la ocurrencia de rachas obedece a alguna ley de probabilidad, mientras que los eventos dentro de una racha, que ocurren de acuerdo con una ley de probabilidad diferente, pueden ser dependientes.

- **Dependencia de duración**

  La composición cambia como consecuencia de no-éxitos previos, de forma que el resultado de un experimento depende del tiempo (nº de extracciones) transcurrido desde el último suceso. Esta situación se denomina *dependencia de duración* (Gourieroux y Magnac, 1997; Gourieroux y Visser, 1997; Winkelmann, 1995). El tratamiento de este tipo de violación del supuesto de independencia pasa por relacionar recuentos y tiempo transcurrido entre eventos, a través de los denominados *modelos de tiempos de espera* y *modelos de duración* (Winkelmann, 2000). Estos modelos escapan del objetivo de este trabajo, por lo que no serán tratados.

4.2.2 Ausencia de estabilidad

La ausencia de estabilidad se caracteriza por el hecho de que la probabilidad de un evento es independiente de la ocurrencia de eventos previos pero no es constante en el espacio o el tiempo. Según la estabilidad sea producto de una evolución temporal o, por el contrario, sea una situación de partida, diferenciamos entre:

- **No-estacionariedad**

  Si la composición de las urnas cambia a través de ensayos consecutivos debido a efectos externos o exógenos mientras que no existe una dependencia de ensayos previos, entonces $\Pr(Y = a)_t \neq \Pr(Y = a)_{t+\Delta}$. Esto ocurre cuando el mecanismo generador de datos obedece a una situación de *inestabilidad* o *no-estacionariedad* («*non-stationarity*»).

  La no-estacionariedad no invalida necesariamente la distribución de Poisson (Winkelmann, 2000).

- **Heterogeneidad no observada**

  Si la composición de las urnas es ya inicialmente heterogénea, es decir, si la proporción inicial de bolas rojas varía entre urnas, entonces $\Pr(Y = a)_t \neq a/(a+b)_t$. Esta situación es el resultado de una heterogeneidad no recogida por el modelo, esto es, *heterogeneidad no observada*.

En el ámbito de las Ciencias de la Salud, especialmente en Medicina y Biología, la no-estacionariedad es conocida como *modelo de fragilidad* o *modelo de propensión* (Navarro et al., 2001, p. 448). Estos modelos se dan cuando existen individuos más propensos a presentar un evento determinado, pero el hecho de haberlo presentado no modifica la probabilidad de volver a tenerlos (eventos independientes) (Navarro et al., 2001). Tal como indican Navarro et al.,(2001), la propensión es el resultado de la no inclusión de variables explicativas, esto es, heterogeneidad no observada.

Cuando la población es heterogénea, se da una ambigüedad en la relación entre el proceso estocástico subyacente y la distribución de recuentos. En presencia de heterogeneidad no observada, la probabilidad de ocurrencia de un evento deja de ser una constante para convertirse en una variable aleatoria (Winkelmann, 2000).

Cuando existe heterogeneidad no observada la probabilidad de ocurrencia de un evento varía entre grupos o «*clusters*» de observaciones, de forma que aunque la probabilidad de ocurrencia de un evento no se ve modificada por la ocurrencia de eventos pretéritos, las propias características diferenciales de esas subpoblaciones producen unas probabilidades de ocurrencia de (futuros) eventos que difieren entre tales subpoblaciones (Winkelmann, 2000). Por este motivo, esta situación es denominada *contagio aparente* o *espurio*. Sin embargo, no siempre es posible identificar el contagio aparente. Teniendo en cuenta que la diferencia entre el contagio real (dependencia de ocurrencia) y el aparente, es que en el primero se parte de una muestra homogénea mientras que en el segundo no, si el modelo de datos de recuento no recoge la evolución de los recuentos, sino que sólo tiene en cuenta los recuentos acumulados tras el período de estudio, difícilmente se podrá diferenciar entre ambos tipos de contagio. Concretamente, tal como indican Long (1997) y Winkelmann (2000), haciendo referencia al «teorema de la imposibilidad» de Bates y Neyman (1951), en presencia de datos de recuento transversales, resulta imposible distinguir entre contagio real y aparente.

4.3 Fuentes de especificación errónea en el MRP

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

1. $\mu_i \mid x_i \sim \text{Poisson}(\mu)$
2. $\mu_i = \exp(x_i\beta)$
3. $\mu_i \mid x_i$, $i = 1, 2, \ldots n$ están distribuidas independientemente.

Una violación de cualquiera de las tres asunciones implica un error de especificación («*misspecification*»). Por otro lado, algunas de las situaciones

generadoras de especificación errónea que se describen a continuación pueden afectar a más de una asunción.

Una de las características del MRP es la íntima relación entre regresión, función variancia y distribución, de tal forma que la violación de la relación media-variancia implica siempre una violación del supuesto distribucional (Winkelmann, 2000). En este sentido, tal como señalan entre otros, Cameron y Trivedi (1986), la imposición inapropiada de la igualdad entre media condicional y variancia condicional puede sesgar a la baja la estimación de los errores estándar de $\hat{\beta}_j$, produciendo de forma espuria valores *z* grandes, valores de significación inferiores a los reales, así como intervalos de confianza espúriamente estrechos (Heo, 2000; Krzanowski, 1998; Lawless, 1987)

Tal como indican diversos autores (Cameron y Trivedi, 1986; King, 1989b; Long, 1997), es frecuente que el MRP presente problemas de ajuste debidos al hecho de que los valores de la variancia condicional no coinciden con los de la media condicional. Concretamente, el problema más común es, tal como se ha expuesto anteriormente, que los valores de la variancia condicional sean superiores a los de la media condicional, esto es, sobredispersión. La sobredispersión es el resultado de la presencia de algunas fuentes de especificación errónea que presentaremos a continuación.

4.3.1 Función media incorrecta

Tal como se ha indicado anteriormente la función media del MRP es:

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

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

$$E_0(y_i \mid x_i) = \mu_0 = f(x_i, \beta_0) \qquad (4.1)$$

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

Los errores de especificación de la función media pueden ser debidos a (Winkelmann, 2000):

- 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. Por ejemplo, la función de enlace es la identidad en lugar de la log-lineal.

Si la función media está correctamente especificada, pero existe sobredispersión, las estimaciones del MRP son consistentes, pero ineficientes (Gourieroux et al., 1984a).

Por otro lado, Winkelmann (2000) afirma que en la práctica difícilmente la función media estará correctamente especificada en todos los aspectos, por lo que aconseja (op. cit) partir de la hipótesis de especificación correcta del modelo para ir probando, de uno en uno, errores de especificación.

4.3.2 Heterogeneidad no observada

En un MRP con heterogeneidad observada, la distribución de $(y_i \mid x_i)$ depende de los valores de las variables explicativas x_i. La situación de *heterogeneidad no observada* surge cuando estas últimas 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).

En el modelo lineal general, el sesgo por variables omitidas surge cuando las variables explicativas x_i incluidas en el modelo y las variables z_i no incluidas están correlacionadas. Sin embargo, en ausencia de correlación las variables omitidas no son problemáticas. De hecho, las variables omitidas son, además del error de medida en y, el argumento estándar para la introducción de una relación estocástica con término de error aditivo.

Sin embargo, el MRP posee una naturaleza estocástica distinta puesto que el modelo no tiene término de error aditivo. De esta forma, en el MRP la heterogeneidad de Y es modelada a través de una función determinística de las variables explicativas. Esto implica que Y es una variable aleatoria porque el proceso de recuento es intrínsecamente estocástico, dada una intensidad μ que se mide sin error. Así pues, un modelo con heterogeneidad no observada no puede seguir una distribución de Poisson (Crepon y Duguet, 1997; Winkelmann, 2000).

Si especificamos el modelo verdadero como:

$$\tilde{\mu}_i = \exp(x_i\beta + z_i\gamma) \tag{4.2}$$

donde z_i es un vector de variables que no han sido incluidas en el MRP usado, podemos derivar la ecuación (Winkelmann, 2000)

$$\tilde{\mu}_i = \mu_i u_i \,,\; u_i > 0, \tag{4.3}$$

donde $\mu_i = \exp(x_i\beta)$ y $u_i = \exp(z_i\gamma)$. Se asume que el término estocástico u_i es independiente de los regresores (Cameron y Trivedi, 1998).

Partiendo del supuesto de que u_i es no observada, debe ser tratada como una variable aleatoria. De esta forma, el parámetro $\tilde{\mu}_i$ se convierte en una variable aleatoria. Las dos fuentes de variación en el parámetro de Poisson $\tilde{\mu}_i$ interactúan de forma multiplicativa: la primera fuente de variación es sistemática y depende de las variables explicativas x_i, mientras que la segunda fuente está causada por un efecto aleatorio individual u_i independiente de x_i. Sea $\ln \varepsilon_i = \ln u_i$; entonces, $\tilde{\mu}_i = \exp(x_i\beta+\varepsilon_i)$ de forma que el error es aditivo en la escala logarítmica (Winkelmann, 2000).

Asúmase que u_i es i.i.d., $E(u_i) = 1$ y $\mathrm{Var}(u_i) = \sigma_u^2$. Los momentos de y_i son (Cameron y Trivedi, 1998):

$$E(y_i \mid x_i) = \mu_i \tag{4.4}$$

$$\mathrm{Var}(y_i \mid x_i) = \mu_i + \mu_i^2\sigma_u^2 \tag{4.5}$$

De esta forma, $\mathrm{Var}(y_i \mid x_i) = \mu_i + \mu_i^2\sigma_u^2 > E(y_i \mid x_i) = \mu_i$, esto es, sobredispersión.

Según se ha expuesto, los modelos de heterogeneidad no observada asumen que las observaciones varían aleatoriamente de una forma no explicada a partir de los predictores del modelo. Esta variación no aleatoria es incorporada en el modelo a través de un término estocástico u_i que se relaciona con μ_i de forma multiplicativa.

Tal como indican Cameron y Trivedi (1998) y Winkelmann (1995), la heterogeneidad no observada implica, generalmente, sobredispersión. Más concretamente, Mullahy (1997, p. 337) afirma que en presencia de heterogeneidad no observada el MRP «*infrapredice la frecuencia real de ceros, sobrepredice la frecuencia real de otros valores pequeños, e infrapredice la frecuencia real de recuentos elevados*».

4.3.3 Proceso dependiente

El MRP presenta errores de especificación si el proceso que genera el evento tiene «memoria», esto es, 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 principales aproximaciones para modelar los procesos dependientes han consistido en:

- la derivación de un MRBN, el cual será discutido en el apartado 4.6.1.
- un modelo de dependencia de duración (Winkelmann, 1995) en procesos de recuento, en el cual se muestra que la distribución de datos de recuento puede ser derivada a partir de una distribución más general que es la distribución gamma. Cabe señalar que según Winkelmann (2000), la ventaja de este modelo con respecto al MRBN es qu la distribución gamma permite la existencia tanto de dependencia de duración negativa, la cual genera sobredispersión, como de dependencia de duración positiva, generadora de infradispersión.

  Teniendo en cuenta que el modelo de dependencia de duración aventaja al MRBN únicamente en el hecho de que admite infradispersión, y dado que tal violación del supuesto de equidispersión no es objeto del presente trabajo, no se detalla dicho modelo.

4.3.4 Selectividad

Si bien la heterogeneidad no observada hace referencia a una situación en la que se puede considerar que existe una pérdida de observabilidad en las variables explicativas, se presenta ahora una situación en que la limitación se sitúa en la observabilidad de la variable de respuesta: la *selectividad*.

La selectividad hace referencia a aquella situación en que 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 o subconjunto y de los mismos. Esta fuente de error en la especificación está presente si los datos están truncados, censurados, así como en situaciones en las que la información está parcializada (véase el apartado 4.3.5), y suele ser debida a la selectividad en el muestreo o bien en el registro de las observaciones (Gurmu, 1991; Gurmu y Trivedi, 1992).

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 la segunda, ciertas observaciones son omitidas completamente de la muestra. Es decir, la censura implica que las observaciones (y_i, x_i) están presentes para un rango restringido de y_i, mientras que todos los valores x_i están disponibles. En el truncamiento se excluyen de la muestra todas las observaciones asociadas a un rango de valores de y_i. De esta forma, la censura conlleva una pérdida de información mucho menos importante que el truncamiento (Cameron y Trivedi, 1998; Winkelmann, 2000). De hecho, la censura ha recibido mucha menos atención en la literatura que el truncamiento. En el presente trabajo, y en relación a la selectividad, nos centraremos en el truncamiento (véase el apartado 4.6.3).

Si se ignora el efecto de la selectividad, las estimaciones del MRP resultan, en general, inconsistentes. Es más, en caso de truncamiento, las estimaciones resultan, además, ineficientes (op. cit).

Se pueden distinguir dos patrones de selectividad:

- Las observaciones pueden ser censuradas o truncadas dependiendo del resultado de y^*. P.ej. algunos cuestionarios presentan la categoría de respuesta «x o más». En este caso, los datos están censurados por la derecha.
- Las observaciones pueden ser censuradas o truncadas dependiendo del resultado de otra variable c, que puede ser independiente o dependiente de y^*. A esta situación se la denomina *censura* o *truncamiento incidental*, así como *selectividad endógena* (Greene, 2000).

4.3.5 Información parcializada

Tal como señala Ribeiro (1999), en muchas aplicaciones los recuentos observados en una muestra constituyen sólo una fracción de todos los eventos, de forma que la información que obtiene está parcializada. Es decir, los eventos ocurren aleatoriamente siguiendo un proceso de Poisson, aunque sólo es reportada una parte de dichos eventos. Este tipo de situaciones reciben el nombre de *información parcializada* (*«underreporting»* o *«selective reporting»*) (Fader y Hardie, 2000).

Aunque el problema de la información parcializada puede ser interpretado en términos de selectividad, conlleva modelos de diferentes tipos; de ahí, la necesidad de distinguir entre ambas fuentes de especificación errónea.

Es habitual que esta fuente de errores de especificación provenga de datos obtenidos de respuestas a encuestas o cuestionarios. Considérese, por ejemplo, un cuestionario de cribado que contiene preguntas sobre antecedentes familiares de psicopatología. Es probable que el encuestado desconozca algunos de los antecedentes de trastornos psicológicos. En consecuencia, el recuento de antecedentes será inferior al real.

Sea y^* el número total (real) de eventos e y el número de recuentos reportados; en una situación de información parcializada $y \leq y^*$. Alternativamente, y^* puede interpretarse como el número de eventos potenciales e y el número de eventos reales. Por ejemplo, y^* podría ser el número de empleos ofertados durante un período de tiempo determinado e y el número de empleos aceptados. De esta forma, $y^* - y$ sería el número de empleos rechazados. Ambas interpretaciones tienen la misma estructura formal (Winkelmann, 2000):

$$y = \sum_{i=1}^{y^*} B_i \tag{4.6}$$

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 de información parcializada, cada uno de los cuales generan un modelo diferente (Winkelmann, 2000):

- 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_i = 1) es una función logística.
- Modelo de recuento total: Los eventos son registrados sólo si se sobrepasa un umbral determinado.

Tal como muestra Ribeiro (1999) la información parcializada puede producir sobredispersión.

Cabe destacar que, quizás debido a la relativa baja frecuencia del problema de información parcializada, no existen modelos específicos para su tratamiento, sino estrategias aisladas para abordarla, cuyas bondades están aún por consensuar. Sin embargo, parece prometedora la reciente aplicación de métodos de simulación para la estimación de parámetros, concretamente la estimación a través de MCMC («*Markov chain Monte Carlo*») (Ribeiro, 1999). De hecho, en el ámbito de las variables de recuento en general, los desarrollos de métodos MCMC están apareciendo con fuerza del mismo modo que el enfoque con el que está estrechamente relacionado: el enfoque bayesiano (Fahrmeir y Lang, 2001; Hauer, 2001; Scollnik, 1995).

4.3.6 Exceso de ceros

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

Tal como indica Mullahy (1997), aunque la heterogeneidad no observada es tratada habitualmente como una fuente de sobredispersión distinta del exceso de ceros, en realidad, puede considerarse como una forma de heterogeneidad no observada, puesto que, es una consecuencia directa de esta última.

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 $1 - \omega$, el resultado es una variable aleatoria de Poisson (λ). Tanto ω como λ pueden depender de las variables explicativas.

En apartados posteriores se amplía el concepto de exceso de ceros así como los modelos de recuentos asociados.

4.3.7 Función variancia incorrecta

Tal como se ha indicado en diversas ocasiones, la ausencia de variancia nominal de Poisson implica una violación del supuesto distribucional.

La función variancia del MRP de referencia es $\mathrm{Var}(y_i \mid x_i) = E(y_i \mid x_i) = \exp(x_i\beta)$, situación que, como es sabido, se denomina equidispersión, mientras que cuando no se da la relación de identidad anterior, se produce sobredispersión, caracterizada por $\mathrm{Var}(y_i \mid x_i) > E(y_i \mid x_i)$, o bien infradispersión, definida a través de la desigualdad $\mathrm{Var}(y_i \mid x_i) < E(y_i \mid x_i)$

Cualquiera de las situaciones generadoras de errores de especificación expuestas anteriormente pueden violar el supuesto de equidispersión, tal como se ha ido indicando en los apartados correspondientes.

La sobredispersión y la infradispersión aparecen cuando en el mecanismo generador de datos subyacente, la función que relaciona la media condicional con la variancia condicional no es la función identidad. 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):

$$\mathrm{Var}(y_i \mid x_i, u_i) = f[\exp(x_i\beta), u_i] = \exp(x_i\beta)u_i \tag{4.7}$$

Esta modificación en la función 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.):

$$\mathrm{Var}(y_i \mid x_i) = \mu_i + \mu_i\alpha^{2-k},\ \alpha \quad ^{+},\ k \tag{4.8}$$

En el caso particular que $k = 0$, la función variancia es equivalente a la del MRBN estándar:

$$\mathrm{Var}(y_i \mid x_i) = \mathrm{Var}_{Negbin\ II}(y_i \mid x_i) = \mu_i + \alpha\mu_i^2 \tag{4.9}$$

Por otro lado, si $k = 1$

$$\mathrm{Var}(y_i \mid x_i) = \mathrm{Var}_{Negbin\ I}(y_i \mid x_i) = \mu_i + \alpha\mu_i \tag{4.10}$$

4.4 Diagnóstico de la sobredispersión

La mayor parte de los errores de especificación conllevan una violación de la asunción de la relación de igualdad entre media y variancia condicionales. Como ya se ha señalado, es la sobredispersión la consecuencia más habitual de la falta de equidispersión (Krzanowski, 1998; Winkelmann, 2000). Este es el principal motivo por el que el presente trabajo se centra, tanto en el apartado que se presenta a continuación como en su parte aplicada, en el problema de la sobredispersión.

Cabe resaltar, de todas formas, que la mayor parte de las pruebas diagnósticas de sobredispersión que se presentan en este apartado pueden considerarse genéricas, en el sentido de que en realidad detectan la ausencia de equidispersión y no únicamente la presencia de sobredispersión.

Las pruebas para la detección de sobredispersión que se presentan a continuación 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 cual 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 un modelo 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.

  A modo de ejemplo, 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 (véase el apartado 4.6.4), que son extensiones de los modelos de referencia como pueden ser el MRP o el MRBN. Dichas extensiones son, con respecto a sus modelos de referencia, ejemplos de modelos no anidados.
- Pruebas basadas en la regresión: De la misma forma que un análisis de residuales en el modelo lineal general con errores distribuidos normalmente puede revelar heterocedasticidad, los residuales de Poisson pueden indicar una violación del supuesto de equidispersión (Winkelmann, 2000). El

análisis de residuales puede llevarse a cabo gráficamente o mediante regresiones auxiliares.

Antes de exponer las pruebas específicas para la evaluación de la ausencia de equidispersión, cabe destacar que es frecuente (Ato et al., 2000b) la aplicación de pruebas algo más genéricas para la evaluación de la sobredispersión que implican evaluar la relación entre χ^2 o la discrepancia y grados de libertad:

- $\frac{\chi^2}{gl}$
- $\frac{D}{gl}$

4.4.1 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 más 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 haciendo efectiva la igualdad anterior tanto en (4.8), en (4.9) como en (4.10), se obtiene:

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

Las que se listan a continuación son las pruebas clásicas de evaluación de sobredispersión:

- Prueba de razón de verosimilitud (*Likelihood ratio, LR*)
- Prueba de Wald
- Prueba multiplicador de Lagrange (*Lagrange Multiplier, 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. Por su parte, Fahrmeir y Tutz (2001, p. 48) señalan que la prueba de Wald y la prueba LM son aproximaciones «*computacionalmente atractivas*» de la prueba LR. Por otro lado, cabe destacar que es poco conocido el grado en que estas pruebas se aproximan a una distribución χ^2 en muestras pequeñas (Long, 1997).

No parece haber un consenso acerca de cual es la prueba más adecuada. Así, mientras que Rothenberg (1984) concluye, al comparar las LR y Wald, que ninguna de las dos pruebas es uniformemente superior, otros autores, como Tu y Zhou (1999) sugieren una cierta superioridad en la potencia de la prueba LR, sobretodo en muestras pequeñas. Fahrmeir y Tutz (2001, p. 48), señalan que la «*ventaja de las pruebas de Wald y LM es que están adecuadamente definidas para modelos con sobredispersión puesto que sólo están involucrados el primer y segundo momento*».

Long (1997) sugiere escoger entre la prueba LR y la prueba de Wald en función del software disponible y de su complejidad de cálculo. En este sentido, mientras que la prueba LR requiere la estimación de dos modelos, el cálculo de la prueba implica simples operaciones aritméticas como la sustracción. En el caso de la prueba de Wald, se requiere sólo la estimación de un modelo pero su cálculo requiere manipulaciones matriciales.

4.4.1.1 Prueba de razón de verosimilitud (LR)

Sea $\hat{\ell}_r$ el valor de la función log-verosímil evaluada en las estimaciones de la máxima verosimilitud restringida (p. ej. el MRP), y $\hat{\ell}_{nr}$ 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):

$$\text{LR} = -2(\hat{\ell}_r - \hat{\ell}_{nr}) \sim \chi^2_{(k)} \tag{4.11}$$

donde $\chi^2_{(k)}$ es una distribución χ^2 con k grados de libertad. Para obtener el nivel de significación, debe usarse un valor crítico $\chi^2_{\alpha,k}$.

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 al nivel δ, donde $\delta > 0.5$, se rechaza H_0 si la prueba estadística supera $\chi^2_{1-2\delta}$ en lugar de $\chi^2_{1-\delta}$.

La prueba LR requiere el uso de la misma muestra para todos los modelos empleados. Puesto que la estimación máximo-verosímil excluye 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.

4.4.1.2 Prueba de Wald

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

Mientras que en su forma más genérica, la prueba de Wald puede ser usada para evaluar restricciones no lineales, consideramos aquí sólo restricciones lineales de la forma (Long, 1997):

$$R\theta = q \tag{4.12}$$

donde R es una matriz de constantes y q es un vector de constantes. $\hat{\theta}$ sigue asintóticamente una distribución normal (Winkelmann, 2000)

$$\hat{\theta} \overset{a}{\sim} N(\theta_0, \hat{Var}(\hat{\theta})) \tag{4.13}$$

y representa los coeficientes de regresión estimados más cualquier parámetro adicional como $\hat{\delta}^2_{nr}$.

Al especificar R y q, pueden definirse una gran variedad de restricciones lineales. Por ejemplo, en un modelo con dos variables explicativas, se podría establecer la restricción $\beta_1 = \beta_2 = 0$. En este caso, R, θ y q quedarían definidos como:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

La hipótesis H_0: $Q\beta=r$ puede ser evaluada mediante el estadístico de Wald (Breslow, 1996; Winkelmann, 2000)

$$W = [R\hat{\theta} - q]' \, [R \, \hat{Var}(\hat{\theta}) \, R')]^{-1} \, [R\hat{\theta} - q] \tag{4.14}$$

W sigue una distribución χ^2 con grados de libertad igual al número de restricciones –el número de filas de R, si la hipótesis nula es correcta. Si el número de restricciones es 1, el estadístico queda reducido al cuadrado del estadístico *t*. Dividiendo W en (4.14) por sus grados de libertad produce un estadístico F.

Tal como se observa en (4.14), la prueba de Wald consiste en dos componentes (Long, 1997):

- $[R\hat{\theta} - q]$ al prinicipio y 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.

A modo de ejemplo, asúmase 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*:

$$\frac{(\hat{\alpha} - 0)}{\left[\sqrt{\hat{Var}(\hat{\alpha})}\right]} \tag{4.15}$$

De hecho, tal como indican Cameron y Trivedi (1998), el test de 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 $z_{1-\alpha}$.

En cuanto a la implementación de la prueba de Wald, para evitar la acumulación de errores de redondeo, Long (1997) recomienda el uso de una alta precisión en las estimaciones de los coeficientes y la matriz de covariancias.

4.4.1.3 Prueba multiplicador de Lagrange (LM)

La prueba multiplicador de Lagrange, conocida también 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

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_r} \tag{4.16}$$

debe estar próxima a 0.

Así, la restricción es rechazada si la puntuación (4.16) 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). Así, la prueba se puede basar en (Cameron y Trivedi, 1998):

$$\mathrm{LM} = \left[\frac{\partial \ell}{\partial \theta}\Big|_{\theta_r}\right]' \left[I(\hat{\theta}_r)\right]^{-1} \left[\frac{\partial \ell}{\partial \theta}\Big|_{\theta_r}\right] \tag{4.17}$$

que sigue una distribución χ^2 con grados de libertad igual al número de restricciones bajo la hipótesis nula.

Un estimador consistente de la variancia puede ser obtenido mediante (Winkelmann, 2000):

$$\sum_{i=1}^{n} \frac{1}{2\hat{\mu}_i^2} \tag{4.18}$$

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

$$LM = \sqrt{\left[\sum_{i=1}^{n} \frac{1}{2\hat{\mu}_i^2}\right]} \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 - y_i \tag{4.19}$$

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.

La prueba LM o prueba de puntuación se aplica a modelos de datos de recuento para diversos tipos de hipótesis nula y alternativa. Así, Gurmu y Trivedi (1992) y Gurmu (1991) derivan pruebas LM para la detección de sobredispersión en modelos de Poisson truncados en 0. Por otro lado, van den Broek (1995) desarrolla una prueba LM para el exceso de ceros.

4.4.2 Pruebas para modelos no anidados

Considérense dos modelos condicionales F_α y G_β, donde $F_\alpha = \{f(y \mid x; \alpha), \alpha \quad A\}$ y $G_\beta = \{g(y \mid x; \beta), \beta \quad A\}$. F_α esta anidado en G_β si $F_\alpha \quad G_\beta$.

Es decir, F_α es un caso particular de G_β, o dicho de otra forma, G_β puede ser equivalente bajo ciertas condiciones a F_α. Por ejemplo, el MRBN y el MRP son modelos anidados puesto que, tal como se ha indicado anteriormente, el MRBN se reduce a un MRP si $\alpha = 0$.

Por el contrario F_α y G_β no son modelos anidados, si $F_\alpha \cap G_\beta \neq \quad$.

El MRBN y el modelo ZINB (modelo de binomial negativa de ceros aumentados), son ejemplos de modelos no anidados. Asimismo, el MRP y modelo ZIP (modelo de Poisson de ceros aumentados) son también modelos no anidados.

4.4.2.1 Prueba de Vuong

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

$$LR_{NA} = \frac{\frac{1}{\sqrt{n}}\left[\left(l_f(\hat{\alpha})\right) - \left(l_g(\hat{\beta})\right)\right]}{\omega} \tag{4.20}$$

donde

$$\omega^2 = \frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i \mid x_i, \hat{\alpha}) - l_g(y_i \mid x_i, \hat{\beta})\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i \mid x_i, \hat{\alpha}) - l_g(y_i \mid x_i, \hat{\beta})\right]\right]^2 \tag{4.21}$$

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:

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

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

En cuanto a la selección del modelo, sea c el valor crítico para el nivel de significación. Si el estadístico $LR_{NA} > c$ (p. ej., $c = 1.96$), se rechaza la hipótesis nula de igualdad entre modelos, y se selecciona el modelo f en lugar del modelo g. Por el contrario, un valor $LR_{NA} < (\tau)$ es indicador de que el modelo g es mejor que el modelo f. Finalmente, 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).

4.4.2.2 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 una restricción en el hiperparámetro (Winkelmann, 2000). Un ejemplo es el modelo Negbin k, ya presentado en apartado anteriores (Cameron y Trivedi, 1986; Winkelmann y

Zimmermann, 1991; Winkelmann y Zimmermann, 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*,

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

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 $\text{Var}(y \mid x) = \mu + \alpha\mu = \text{Var}_{\text{Negbin I}}$
- Si $k = 0$, entonces $\text{Var}(y \mid x) = \mu + \alpha\mu^2 = \text{Var}_{\text{Negbin II}}$

4.4.3 Pruebas basadas en la regresión

Cameron y Trivedi (1990) presentan diversas pruebas basadas en los residuales del MRP que no requieren la estimación del MRBN, y que pueden ser útiles cuando no se dispone de un paquete estadístico que implemente el MRBN. Los autores (op. cit) proponen estimar el MRP y después llevar a cabo una regresión lineal auxiliar (sin constante):

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha\hat{\mu}_i + u_i \qquad (4.22)$$

donde u_i es un término de error. El estadístico *t* obtenido para α es asintóticamente normal bajo la hipótesis nula de equidispersión contra la alternativa de sobredispersión de la forma Negbin II. Análogamente, el test de sobredispersión para Negbin I es:

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha + u_i \qquad (4.23)$$

Estas pruebas de regresión auxiliar coinciden con la prueba LM (Lagrange Multiplier) para MRP contra MRBN (Cameron y Trivedi, 1998).

Más allá de la aceptación o el rechazo de la hipótesis nula de equidispersión, es importante conocer la magnitud de la desviación en relación a la equidispersión. Las estimaciones de α para la función variancia Negbin I $(1+\alpha)\mu_i$ se interpretan como (Cameron y Trivedi, 1998):

- Infradispersión: si $\alpha < 0$

- Sobredispersión:
  - moderada si $0 < \alpha < 1$
  - elevada si $\alpha > 1$.

Para la función variancia Negbin II $\mu_i + \alpha \mu_i^2 = (1 + \alpha\mu_i)\ \mu_i$:

- Infradispersión: si $\alpha < 0$. Cabe indicar que la función variancia Negbin II puede ser inapropiada para datos con infradispersión puesto que la variancia estimada es negativa para las observaciones con $\alpha < -1/\mu_i$.
- Sobredispersión: $\alpha\mu_i > 1$ indica sobredispersión elevada porque el multiplicador $1 + \alpha\mu_i > 2$. Es decir, un valor de, por ejemplo, $\alpha = 0.5$ puede indicar sobredispersión moderada si los valores más frecuentes de la variable dependiente están entre 0 y 2. En cambio si los valores más frecuentes son superiores a 9, existirá una acusada sobredispersión.

4.5 Corrección de la estimación del error estándar de los coeficientes del MRP

Antes de pasar a detallar los principales modelos que pretenden tomar cuenta de la sobredispersión, es importante señalar que también existe la posibilidad, como señalan, entre otros Palmer et al. (2001), de corregir directamente el error estándar de los coeficientes del MRP mediante su producto por diferentes índices:

- $EE\sqrt{\dfrac{\chi^2}{gl}}$

- $EE\sqrt{\varphi}$, donde φ es el parámetro de dispersión

- $EE\sqrt{\dfrac{D}{gl}}$, donde D es la discrepancia

También es posible la corrección del error estándar a través de técnicas relacionadas con la simulación como (Cameron y Trivedi, 1998):

- sándwich
- jacknife

- bootstrap

En el último experimento de simulación que se presenta en este trabajo se verifica la adecuación real de estos procedimientos bajo diferentes grados de sobredispersión y distintos tamaños muestrales, al tiempo que se comparan entre sí para establecer reglas de aplicación práctica.

En cualquier caso, es importante destacar que estos procedimientos sólo pretenden corregir la infraestimación de los errores estándar de los coeficientes en el MRP en presencia de sobredispersión, mientras que los modelos que se presentan a continuación pretenden modelar directamente la causa de la sobredispersión.

4.6 Modelos para datos de recuento en presencia de sobredispersión

Si en apartados previos se han expuesto las causas de especificación errónea que son generadoras de sobredispersión, se presentan ahora modelos de datos de recuento que permiten dar cuenta de esa sobredispersión. Puesto que ya se han presentado las fuentes de especificación errónea, partiremos de esta información para presentar una clasificación de los modelos que constituye, en realidad, una reagrupación de la clasificación presentada en el apartado 3.3.6.

Los modelos para datos de recuento con presencia de sobredispersión pueden clasificarse en tres grupos, en función de su especificidad en relación a la situación generadora de sobredispersión, es decir, según si la estrategia que aplican para abordar la sobredispersión se basa en modelar el origen de la especificación errónea o bien directamente sus 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. De esta forma, la solución pasa por adoptar 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).

    El modelo de regresión binomial negativa (MRBN) es, sin duda, el miembro de esta categoría que se aplica con mayor frecuencia, motivo por el cual merece un tratamiento más extenso que el resto y se presenta en detalle en el siguiente apartado. En general, 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).

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.

4.6.1 Modelo de regresión binomial negativa (MRBN)

Como ya hemos indicado antes, 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 situación.

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:

$$\tilde{\mu}_i = \exp(x_i\,\beta+\varepsilon) \qquad (4.24)$$

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 $\tilde{\mu}$ y la μ «original» es:

$$\tilde{\mu}_i = \exp(x_i\beta)\exp(\varepsilon_i) = \mu_i \exp(\varepsilon_i) = \mu_i \delta_i \tag{4.25}$$

donde δ_i se define como equivalente a $\exp(\varepsilon_i)$. La concreción 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):

$$E(\delta_i) = 1 \tag{4.26}$$

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

$$E(\tilde{\mu}_i) = E(\mu_i \delta_i) = \mu_i E(\delta_i) = \mu_i \tag{4.27}$$

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

$$\Pr(y_i \mid x_i, \delta_i) = \frac{\exp(-\tilde{\mu}_i)\tilde{\mu}_i^{y_i}}{y_i!} = \frac{\exp(-\mu_i\delta_i)(\mu_i\delta_i)^{y_i}}{y_i!} \tag{4.28}$$

Sin embargo, puesto que δ es desconocido no podemos calcular $\Pr(y \mid x\delta)$. Para calcular $\Pr(y \mid x)$ sin tener en cuenta δ, promediamos $\Pr(y \mid x\delta)$ 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):

$$\Pr(y_i \mid x_i) = \int_0^{\infty} [\Pr(y_i \mid x_i, \delta_i)] g(\delta_i) d\delta_i = \int_0^{\infty} \frac{e^{-\exp(x_i\beta+\delta_i)} \exp(x_i\beta + \delta_i)^{yi}}{y_i!} g(\delta_i) d\delta_i \quad (4.29)$$

Esta expresión define la distribución de Poisson compuesta (Cameron y Trivedi, 1986). Tal como indican estos mismos autores (op. cit.), 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 (4.29) calcula la probabilidad de y como una mezcla de dos distribuciones de probabilidad (Long, 1997). Asimismo, la forma de (4.29) depende de la selección de $g(\delta_i)$, 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:

$$g(\delta_i) = \frac{v_i^{v_i}}{\Gamma(v_i)} \delta_i^{v_i-1} \exp(-\delta_i v_i) \text{ para } v_i > 0 \quad (4.30)$$

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 $E(\delta_1) = 1$, ecuación que coincide con la asunción del MRBN expuesta anteriormente, y $\text{Var}(\delta_i) = 1/v_i$. 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):

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + v_i)}{\Gamma(y+1)\Gamma(v_i)} \left(\frac{v_i}{v_i + \mu_i}\right)^{v_i} \left(\frac{\mu_i}{v_i + \mu_i}\right)^{y_i} \quad \text{para } y_i = 0, 1, 2\ldots \quad (4.31)$$

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

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

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

$$\text{Var}(y_i \mid x) = \mu_i\left(1 + \frac{\mu_i}{v_i}\right) = \exp(x_i\beta)\left(1 + \frac{\exp(x_i\beta)}{v_i}\right) \tag{4.32}$$

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 $\text{Var}(y \mid x) \rightarrow \mu$. 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 (4.32) 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):

$$v_i = \alpha^{-1} \text{ para } \alpha > 0 \tag{4.33}$$

De esta forma, la densidad (4.31) queda reexpresada como:

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \tag{4.34}$$

Por otro lado, la asunción (4.33) 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:

$$\text{Var}(y_i \mid x) = \mu_i\left(1 + \frac{\mu_i}{\alpha^{-1}}\right) = \mu_i(1+\alpha\mu_i) = \mu_i + \alpha\mu_i^2 \qquad (4.35)$$

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

La densidad (4.34) y la variancia (4.35) 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,

$$\Pr(y_i \mid x_i) = \frac{\Gamma(y_i + \alpha^{-1}\mu_i)}{\Gamma(y+1)\Gamma(\alpha^{-1}\mu_i)}\left(\frac{\alpha^{-1}\mu_i}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \qquad (4.36)$$

y su variancia condicional es:

$$\text{Var}(y_i \mid x) = \mu_i + \alpha\mu_i = (1+\alpha)\mu_i \qquad (4.37)$$

Además de los modelos Negbin I i Negbin II considerados anteriormente, algunos autores como Cameron y Trivedi (1986) o Winkelmann y Zimmermann (1991) proponen el denominado hipermodelo Negbin *k*, en el cual $\text{Var}(y \mid x) = \mu + \alpha\mu^{2-k}$ (Véase apartado 4.4.2.2).

4.6.1.1 Estimación

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

$$L(\beta; Y, X) = \prod_{i=1}^{n} \Pr(y_i \mid x_i)$$

$$= \prod_{i=1}^{n} \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \quad y=0, 1, 2, \ldots \quad (4.38)$$

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):

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

$$= \sum_{i=1}^{n} \left[\sum_{j=0}^{y_i - 1} \ln(j + \alpha^{-1}) - \ln y_i! - (y_i + \alpha^{-1}) \ln(1 + \alpha\mu_i) + y_i \ln \alpha + y_i x_i' \beta \right] \qquad (4.39)$$

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

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

4.6.1.2 Interpretación

Los métodos de interpretación basados en el recuento esperado $E(y \mid x)$ 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 basadas en

$$\hat{\Pr}(y \mid x) = \frac{\Gamma(y_i + \hat{\alpha}^{-1})}{\Gamma(y+1)\Gamma(\hat{\alpha}^{-1})} \left(\frac{\hat{\alpha}^{-1}}{\hat{\alpha}^{-1} + \hat{\mu}} \right)^{\hat{\alpha}^{-1}} \left(\frac{\hat{\mu}}{\hat{\alpha}^{-1} + \hat{\mu}} \right)^{y_i} \qquad (4.40)$$

para el modelo Negbin II, y en

$$\hat{\Pr}(y \mid x) = \frac{\Gamma(y_i + \hat{\alpha}^{-1}\hat{\mu})}{\Gamma(y+1)\Gamma(\hat{\alpha}^{-1}\hat{\mu})} \left(\frac{\hat{\alpha}^{-1}\hat{\mu}}{\hat{\alpha}^{-1} + \hat{\mu}} \right)^{\alpha^{-1}} \left(\frac{\hat{\mu}}{\hat{\alpha}^{-1}\hat{\mu} + \hat{\mu}} \right)^{y_i} \qquad (4.41)$$

para el modelo Negbin I.

4.6.2 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 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, GEC*$_k$) introducido por King (1989b). 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.

4.6.2.1 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.

Su función de densidad es:

$$f(y_i) = \left(\frac{\mu_i}{1+a\mu_i}\right)^{y_i} \frac{(1+a\mu_i)^{y_i-1}}{y_i!} \exp\left(-\frac{\mu_i(1+ay_i)}{1+a\mu_i}\right) \tag{4.42}$$

La media condicional de la RPG es

$$\mathrm{E}(y_i \mid x_i) = \mu_i \tag{4.43}$$

y su variancia condicional

$$\mathrm{Var}(y_i \mid x_i) = \mu_i (1+a\mu_i)^2 \tag{4.44}$$

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 (4.42) queda reducida al MRP.

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

$$\ell = (a, \beta, y_i) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\mu_i}{1 + a\mu_i} \right) + (y_i - 1)\ln(1 + ay_i) - \frac{\mu_i (1 + ay_i)}{1 + a\mu_i} - \ln(y_i!) \right\} \tag{4.45}$$

4.6.2.2 Regresión de Poisson robusta

Una de las asunciones básicas del MRP, al igual que para el resto de modelos lineales generalizados, es que la densidad verdadera pertenece a una distribución determinada de la familia exponencial. Es decir, el MRP es un ejemplo de modelado paramétrico en el que se especifica un mecanismo generador de datos dependiente de algunos parámetros. En este apartado se revisa la robustez del MRP cuando el mecanismo generador de datos no es Poisson (Fahrmeir y Tutz, 2001; Winkelmann y Zimmermann, 1995).

En el MRP la estructura de la media está relacionada con la estructura de la variancia. Como ya se ha expuesto anteriormente, 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 $(y_i \mid x_i) = \varphi\mu$. 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 máximo-verosímil («*pseudo-maximum likelihood*», *PML*). Winkelmann (2000, p. 84) advierte que la estimación QML es «*en general inconsistente e ineficiente*». Sin embargo, Gourieroux, Monfort y Trognon (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

$$\sum_{i=1}^{n} [y_i - \exp(x_i'\beta)]x_i = 0 \tag{4.46}$$

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

$$\hat{\beta} \overset{a}{\sim} N[\beta, \mathrm{Var}_{PML}(\hat{\beta})] \tag{4.47}$$

donde

$$\mathrm{Var}_{PML}(\hat{\beta}) = \left(\sum_{i=1}^{n} \mu_i x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} \omega_i x_i x_i'\right) \left(\sum_{i=1}^{n} \mu_i x_i x_i'\right)^{-1} \tag{4.48}$$

y $\omega_i = \hat{Var}(y_i \mid x_i)$.

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

- Ausencia de asunción (Breslow, 1990):

  En este caso, la función variancia estimada es:

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

· Función variancia lineal (McCullagh y Nelder, 1989)

En esta situación, la función variancia estimada viene dada por:

$$\hat{Var}(y_i \mid x_i) = \hat{\delta}^2 \hat{\mu}_i$$

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

· Función variancia cuadrática (Gourieroux et al., 1984a)

La función variancia estimada es:

$$\hat{Var}(y_i \mid x_i) = \hat{\mu}_i + \hat{\delta}^2 \hat{\mu}_i^2$$

A través de regresión auxiliar (véase apartado 4.4.3) puede obtenerse un estimador de δ^2:

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

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 quasi-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).

4.6.3 Modelos de datos de recuento truncados

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 $c = 1$.

Existen dos grandes tipos de truncamiento (Winkelmann, 2000):

1. *Y* es un conjunto de enteros positivos con exclusión del 0 {1,...,∞}, esto es, distribución de recuentos con *ceros truncados*.
2. *Y* es un conjunto de {0, …, *a*} donde *a* es algún entero positivo, esto es, distribución 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).

En las distribuciones con ceros truncados, las observaciones pasan a formar parte de la muestra una vez ha ocurrido el primer recuento. Un ejemplo lo constituye una investigación acerca del número de episodios de insomnio en muestra clínica que presenta al menos 1 episodio de insomnio (Lichstein et al., 2001).

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.

4.6.3.1 MRP de ceros truncados

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

$$\Pr(y_i = 0 \mid x_i) = \exp(-\mu_i) \tag{4.49}$$

$$\Pr(y_i > 0 \mid x_i) = 1 - \exp(-\mu_i) \tag{4.50}$$

La distribución de probabilidad para el MRP de ceros truncados se define como (Long, 1997; Winkelmann, 2000):

$$\Pr(y \mid y_i > 0, x_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!(1-\exp(-\mu_i))}, \quad y = 1, 2, \ldots \tag{4.51}$$

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):

$$\mathrm{E}(y \mid y_i > 0, x) = \frac{\mu_i}{1-\exp(-\mu_i)} \tag{4.52}$$

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

$$\text{Var}(y \mid y_i > 0, x) = \text{E}(y \mid y_i > 0, x_i)\,[1 - \Pr(y_i = 0 \mid x_i)\,\text{E}(y \mid y_i > 0, x_i)]$$

$$= \frac{\mu_i}{1 - \exp(-\mu_i)} \left(1 - \frac{\mu_i}{\exp(\mu_i) - 1} \right) \tag{4.53}$$

Puesto que μ (la media de la distribución no truncada) es mayor que cero, $0 < \exp(-\mu) < 1$, de forma que el modelo truncado se desplaza a la derecha.

Winkelmann y Zimmermann (1995) afirman que el MRP de ceros truncados tiende a presentar infradispersión en relación al MRP no truncado puesto que $0 < 1 - \mu(\exp(\mu) - 1) < 1$. 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.

4.6.3.2 MRBN de ceros truncados

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

$$\Pr(y_i = 0 \mid x_i) = (1 + \alpha\mu_i)^{-\alpha^{-1}} \tag{4.54}$$

$$\Pr(y_i > 0 \mid x_i) = 1 - (1 + \alpha\mu_i)^{-\alpha^{-1}} \tag{4.55}$$

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

$$\Pr(y \mid y_i > 0, x_i) = \frac{\dfrac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\dfrac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \left(\dfrac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}}, \; y = 1, 2, \ldots \tag{4.56}$$

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

$$\text{E}(y \mid y_i > 0, x_i) = \frac{\mu_i}{1 - (1 + \alpha\mu_i)^{-\frac{1}{\alpha}}} \tag{4.57}$$

$$\mathrm{Var}(y \mid y_i > 0, x_i) = \frac{\mu_i}{1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}}} \times \left[1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}} \frac{\mu_i}{1-(1+\alpha\mu_i)^{-\frac{1}{\alpha}}}\right] \quad (4.58)$$

4.6.4 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. Puesto que es el exceso de ceros el fenómeno más frecuente y el que produce sobredispersión –la escasez de ceros produce el efecto contrario, es decir, infradispersión (Winkelmann y Zimmermann, 1995)-, nos centraremos en este apartado en los modelos que tratan el exceso de ceros.

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).

Por ejemplo, en el ámbito de la seguridad vial, los recuentos de accidentes pueden concebirse como un proceso biestado:

- Un estado es el de cero-accidentes que estaría compuesto por aquellas secciones de carretera consideradas inherentemente seguras, en las cuales se espera que no haya accidentes.
- El otro estado es el «estado-de-accidente» donde los siniestros siguen una distribución conocida, como la distribución de Poisson o la binomial negativa.

Si un proceso biestático como el descrito, es modelado como un proceso simple o monoestado en el cual se asume que todas las secciones de la carretera forman parte del segundo estado –tal como se asume en el MRP o en el MRBN, las estimaciones de los modelos resultan sesgadas debido a una sobrerrepresentación en los datos de las observaciones de cero accidentes. Tales estimaciones pueden, además, sugerir la existencia de sobredispersión en los datos, indicando la idoneidad de aplicar un MRBN cuando el MRP puede ser perfectamente válido, puesto que la sobredispersión es el resultado de un modelo incorrectamente especificado (Shankar et al., 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.

4.6.4.1 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 si el valor de tales observaciones supera o no cierto umbral. Esta bifurcación se consigue combinando un modelo dicotómico para el resultado binario de estar por encima o por debajo del umbral, con un modelo truncado para resultados superiores al 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 WT parece tener en la literatura dos acepciones. En ambos casos se considera que el modelo WT 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*.

4.6.4.1.1 Acepción A

El modelo WT 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). De esta forma, este podría ser un modelo que se ajustase a una situación habitual en la investigación en Psicología clínica: al pasar una prueba para valorar la existencia de un trastorno a una muestra subclínica (supongamos que es obtenida al azar), tendremos un exceso de ceros si dicha prueba realiza un recuento del número de

síntomas presentes. En este caso podemos considerar que las observaciones provienen, en realidad, de dos poblaciones: una «patológica» y la otra «no-patológica». En esta situación, la alternativa seria utilizar previamente una prueba de cribado de la cual se obtuviera el resultado binario {presencia/ausencia de patología} y posteriormente se realizara la prueba para valorar el número de síntomas presentes.

Para una formulación del modelo de umbral, asúmase que f_1 y f_2 son funciones de distribución de probabilidad para enteros no negativos. Si f_1 gobierna la parte de umbral (valor 0) y f_2 el proceso una vez sobrepasado el umbral (>0) la función de probabilidad del modelo de umbral viene dada por (Winkelmann, 2000):

$$\begin{aligned} &\Pr(y = 0) = f_1(0) \\ &\Pr(y = k) = \Phi f_1(k) \text{ para } k = 1, 2, \ldots \end{aligned} \qquad (4.59)$$

siendo $\Phi = \dfrac{1 - f_1(0)}{1 - f_1(0)}$.

De lo cual sigue que si $f_1 = f_2$, $\Phi = 1$ y el modelo de umbral se convierte en el modelo de referencia (p.ej. la distribución de Poisson).

El valor esperado viene dado por:

$$\mathrm{E}_h(y) = \sum_{k=1}^{\infty} k f_2(k) \Phi \qquad (4.60)$$

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):

$$\mathrm{Var}_h(y) = \sum_{k=1}^{\infty} k^2 f_2(k) \Phi - \left[\Phi \sum_{k=1}^{\infty} k f_2(k) \right]^2 \qquad (4.61)$$

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

$$\frac{Var_h(y)}{E_h(y)} = \frac{\sum_{k=1}^{\infty} k^2 f_2(k) - \Phi\left[\sum_{k=1}^{\infty} k f_2(k)\right]^2}{\sum_{k=1}^{\infty} k f_2(k)} \qquad (4.62)$$

Tal como se ha indicado, si $\Phi = 1$ 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 $\Phi \neq 1$, (4.59) define el modelo de umbral: Si $0 < \Phi < 1$ existe sobredispersión.

La función de log-verosimilitud es:

$$\ell = \sum_{y=0} \ln f_1(0) + \sum_{y>0} \ln\left[1 - f_1(0)\right] + \sum_{y>0} \{\ln f_2(y) - \ln\left[1 - f_2(0)\right]\} \qquad (4.63)$$

4.6.4.1.2 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 $1 - \psi$. 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 (3.14), es decir:

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

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):

$$\Pr(y_i = 0 \mid x_i) = \psi + (1-\psi)\exp(-\mu_i) \tag{4.64}$$

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):

$$\Pr(y_i \mid x_i) = (1-\psi)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, \text{ para } y > 0 \tag{4.65}$$

4.6.4.2 Modelos de datos de recuento con ceros aumentados

Existen dos modelos de datos de recuento con ceros aumentados:

- El *modelo de Poisson de ceros aumentados* (*zero-inflated Poisson, ZIP*), cuya distribución de probabilidad deriva de la distribución de Poisson.
- El *modelo de binomial negativa de ceros aumentados* (*zero-inflated negative binomial, ZINB*), cuya distribución de probabilidad deriva de la distribución binomial negativa.

El concepto que sustenta los modelos de recuentos con ceros aumentados es que una probabilidad binomial gobierna el resultado binario de que el recuento sea 0 o bien 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:

$$y_i = \begin{cases} 0 & \text{si } c_i = 1 \\ y_i^* & \text{si } c_i = 0 \end{cases} \tag{4.66}$$

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

$$f(y_i) = \omega_i(1-d_i) + (1-\omega_i)g(y_i), \quad y_i = 0,1,2 \ldots \tag{4.67}$$

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.

Winklemann (2000) indica, en relación a la primera de las acepciones del modelo WT expuesta anteriormente, la diferencia entre el modelo de datos de recuento

con ceros aumentados y el modelo WT 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 $y_i^* = 0$.

4.6.4.2.1 Modelo de Poisson de ceros aumentados (ZIP)

Lambert, (1992) introdujo el modelo de Poisson de ceros aumentados (*ZIP*) para poder capturar la influencia de las variables explicativas en la probabilidad de ceros adicionales (Winkelmann, 2000):

$$\omega_i = F(z_i, \gamma) = \frac{\exp(z_i\gamma)}{1 + \exp(z_i\gamma)} \tag{4.68}$$

El resto de valores de recuento, es decir, los generados a partir de un segundo proceso en el cual la probabilidad de un valor de recuento viene dada por la distribución de referencia, pueden ser generados por una distribución de Poisson o por una binomial negativa. Así, en el caso de Poisson la probabilidad de recuentos 0 o bien estrictamente positivos, viene dada por (3.14).

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

$$\Pr(y_i = 0 \mid x_i) = \omega_i + (1-\omega_i)\exp(-\mu_i) \tag{4.69}$$

$$\Pr(y_i \mid x_i) = (1-\omega_i)\ \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \quad \text{para } y_i > 0 \tag{4.70}$$

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

$$\ell(\beta,\gamma) = \sum_{d_i=1} \ln(\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))$$

$$+ \sum_{d_i=0} y_i x_i'\beta - \exp(x_i'\beta) - \ln(y!) - \sum_{i=1}^{n} \ln(1 + \exp(z_i'\gamma)) \tag{4.71}$$

En cuanto a la interpretación de las probabilidades predichas del modelo ZIP, ésta se diferencia en función del valor de recuento (Long, 1997):

· La probabilidad predicha de un recuento igual a 0 es:

$$\hat{\Pr}(y = 0 \mid x) = \hat{\omega} + (1 - \hat{\omega})\exp(-\hat{\mu}) \tag{4.72}$$

donde $\hat{\mu} = \exp(x_i\hat{\beta})$ y $\hat{\omega} = F(z_i\hat{\gamma})$.

· La probabilidad predicha de un recuento positivo es:

$$\hat{\Pr}(y \mid x) = (1-\hat{\omega})\,\frac{\exp(-\hat{\mu})\hat{\mu}^{y_i}}{y!} \quad \text{para } y_i > 0 \tag{4.73}$$

4.6.4.2.2 Modelo de binomial negativa de ceros aumentados (ZINB)

Por su parte, en el modelo ZINB, las probabilidades de diferentes valores de recuento se obtienen a partir de

$$\Pr(y_i = 0 \mid x_i) = \omega_i + (1-\omega_i)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \tag{4.74}$$

$$\Pr(y_i \mid x_i) = (1-\omega)\frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \tag{4.75}$$

De la misma forma que en el modelo ZIP, las probabilidades predichas para el modelo ZINB, se obtienen de forma diferente según el valor de recuento (Long, 1997):

· La probabilidad predicha de un recuento igual a 0:

$$\hat{\Pr}(y = 0 \mid x) = \hat{\omega} + (1-\hat{\omega})\left(\frac{\hat{\alpha}^{-1}}{\hat{\alpha}^{-1} + \hat{\mu}}\right)^{\hat{\alpha}^{-1}} \tag{4.76}$$

· la probabilidad predicha de un recuento positivo es:

$$\hat{\Pr}\ (y \mid x) = (1-\hat{\omega})\frac{\Gamma(y_i+\hat{\alpha}^{-1})}{y_i!\Gamma(\hat{\alpha}^{-1})}\left(\frac{\hat{\alpha}^{-1}}{\hat{\alpha}^{-1}+\hat{\mu}}\right)^{\hat{\alpha}^{-1}}\left(\frac{\hat{\mu}}{\hat{\alpha}^{-1}+\hat{\mu}}\right)^{y_i} \qquad (4.77)$$

donde $\hat{\mu} = \exp(x_i\hat{\beta})$ y $\hat{\omega} = F(z_i\hat{\gamma})$.