UNIVERSITAT DE BARCELONA

Modelos para el análisis de supervivencia en tiempos discretos: aplicación en el área de veterinaria

Carolina Barroeta Rojo

Modelos para el Análisis de Supervivencia en Tiempo Discreto. Aplicación en el área de Veterinaria

MEMORIA PRESENTADA POR:

Carolina Barroeta Rojo

PARA OPTAR AL TÍTULO DE DOCTOR POR LA
UNIVERSIDAD DE BARCELONA

DOCTORANDO:

Carolina Barroeta Rojo

DIRECTORAS:

Dra. Olga Julià de Ferran **Dra. Anna Espinal Berenguer**

Universidad de Barcelona
Facultad de Matemáticas
Programa de Doctorado en Estadística
Departamento de Probabilidad, Lógica y Estadística
Barcelona, Enero 2016

A mi Dios, mi Roca fuerte en la tempestad.

A mi madre, que duerme en espera de la eternidad.

A Manuel Felipe mi esposo, a quien amo.

A mis hijos Manuel João y Andrés Felipe,

porque son el motivo de mi vida.

Agradecimientos

Todo camino tiene su inicio y final, en el transcurrir de este, hasta la meta que hoy he alcanzado, tuve el placer y la bendición de tener a mi alrededor personas muy valiosas, unas aún me acompañan, otras viven lejos y otras han marchado en espera de la eternidad.
Agradezco a quien cada mañana me presta el aliento de vida y me sostiene en tiempos pocos fáciles, a Dios mi amigo que siempre me acompaña.

Deseo agradecer de manera especial a mis directoras de tesis Dra. Olga Julià de Ferran y Dra. Anna Espinal Berenguer, por compartir sus conocimientos, experiencia profesional y valiosas enseñanzas muy relevantes en el desarrollo de esta tesis. Gracias a su disponibilidad de tiempo, consejos y ánimos para llegar al final de este camino.

Le estoy agradecida a los miembros que forman el grupo de investigación GRASS, por sus orientaciones y mejoras en este documento, con sus enriquecedores seminarios. Especialmente a la Dra. Guadalupe Gómez por invitarme desde el principio a los seminarios y al Dr. Klaus Langohr por sus asesorías. Gracias a todos por esa energía y entusiasmo que les caracteriza.
Igualmente agradezco el trato cordial que recibí durante mi estancia de los profesores del Departamento Probabilidad, Logíca y Estadística de la Facultad de Matemática, especialmente por parte de Carme Florit y M. Dolors Rovira. Agradezco los primeros momentos que pase en la Facultad de Biología, gracias a los profesores que integran el Departamento de Estadística, también a Roser Maldonado, por su ayuda y cordialidad durante este proceso.

Mis agradecimientos a la Universidad Centroccidental Lisandro Alvarado en Venezuela, por la oportunidad y apoyo al permitir y facilitar mi viaje a Cataluña, para realizar mis estudios de doctorado.

Agradezco a mis amigos: Abelardo Morales, por su apoyo y animo en todo momento y por compartir su experiencia profesional. A Walter Díaz, mi compañero de doctorado, por su apoyo incondicional y valioso tiempo compartido. A Noèlia Viles, mi amiga y compañera del despacho durante los últimos años del doctorado, por su apoyo y enriquecedoras conversaciones, las guardare por siempre.

Le doy las gracias a mi familia. A Manuel F. Martins, mi esposo y amigo quien me motivo a iniciar el doctorado y me ha acompañado, animado y apoyado desde el inicio hasta el final de este camino. A mis dos grandes motores, mis hijos Manuel João y Andrés Felipe, no tengo como agradecerles su apoyo y estimulo para seguir hasta el final de este camino, cada día aprendo mas de ellos. Les amo profundamente.
A mis padres Sergio Barroeta y Maria Rosa de Barroeta, gracias por la vida que me regalaron, sus constantes oraciones, dedicación y consejos lograron llevarme donde hoy estoy. En especial a ti mamá te fuisteis cuando menos lo esperaba, te extrañare por siempre, pero nos volveremos a encontrar en nuestra verdadera patria celestial.
A mis hermanas Dora J.M. y Maria Elena y a mis cuñados-hermanos de corazón Felipe Uttaro y M. Tolentino DeFreitas (difícil es tu ausencia) por el amor y cuidados que me han brindado. A mis sobrinos Salvatore, Sergio Giovanni, Samuel, David y Sussana, por su cariño y apoyo. Les quiero a todos de corazón.

Agradezco a mis amigos por su apoyo y palabras oportunas: Elizabeth Fushan, Eliasib Sánchez, Alba Araujo, Ruth C. León, Antonietta Ventrone, Margarita Cobo, Martha Muñoz, Maria C. Gil, Trina Escobar, Nelson Orellana, Giovanni Rojas.
A todos, gracias y otra vez gracias, les recordaré hasta el final de mis días.

Carolina Barroeta Rojo.
Barcelona, Enero del 2016.

Resumen

En esta tesis se han estudiado y comparado métodos para abordar tiempos discretos en el análisis de supervivencia, con especial aplicación en datos reales en el ámbito de la veterinaria.

En primer lugar, se introduce el modelo de Cox con tratamientos de empates (Efron, Breslow, Exact y Average), así como modelos para una variable respuesta binaria (logit y clog-log), con la finalidad de abordar un tiempo discreto en análisis de supervivencia. Estas metodologías han sido aplicadas a un estudio con de datos de caballos de carreras pura sangre y permitieron identificar los factores de riesgo asociados a un evento de interés, la lesión musculoesquelética catastrófica (CMI). Las covariables estadísticamente significativas fueron: si había una lesión anterior, el número de carreras donde participó el caballo y la longitud de la carrera. El género y la época, dado su interés desde el punto de vista veterinario, también fueron incluídas en los análisis. En este estudio se observó similitud entre las estimaciones obtenidas en el modelo de Cox con los diferentes tratamientos de empates y los modelos discretos logit y clog-log. Se establecieron tres grupos de resultados: (1) estimaciones proporcionadas por el método exact y el modelo discreto logit; (2) modelos proporcionados por el modelo de Cox con los métodos de tratamiento de empates Efron y Average, y el modelo discreto clog-log; (3) estimaciones proporcionadas por el modelo de Cox con tratamiento de empates Breslow.

En la segunda parte de esta tesis se introdujeron y usaron los métodos para abordar modelos en tiempo discreto en presencia de heterogeneidad no observada, incluyendo uno o dos términos de frailty. Estos modelos se aplicaron a un conjunto de datos reales donde el objetivo fue caracterizar el tiempo (en número de lactancias) hasta el primer diagnóstico de mastitis en vacas de producción de leche. Se pudo constatar que entre las variables fijas el tipo de ordeño fue siempre estadísticamente significativa. Además, también se

obtuvo un efecto rebaño, resumido en el término de frailty. Al considerar un segundo término de frailty correspondiente a la zona geográfica, también resultó estadísticamente significativo.

En la tercera parte de esta tesis se ha realizado una comparación de tres software disponibles (R, Stata y SAS) para abordar datos de análisis de supervivencia para tiempo discreto. Esta comparación se ha realizado con el estudio del diagnóstico de mastitis en vacas lecheras, y para modelos con uno o dos términos de frailty. Se establecieron tres grupos de resultados: (1) formado por el modelo de Cox con método de empates Average y el modelo discreto clog-log; (2) formado por el modelo de Cox con método de empates Exact y el modelo logit; (3) formado por el modelo de Cox con método de tratamiento de empates Breslow.

Finalmente, cabe destacar que en esta tesis se pone de relieve la importancia de considerar la naturaleza discreta del tiempo, en estudios de análisis de la supervivencia. Además, se considera también la ventaja de recoger la influencia, mediante uno o más términos de frailty, de la heterogeneidad no observada cuando ésta es relevante.

Resum

En aquesta tesi s'han estudiat i comparat mètodes per abordar temps discrets en l'anàlisi de supervivència, amb especial aplicació en dades reals en l'àmbit de la veterinària.
En primer lloc, s'introdueix el model de Cox amb tractaments d'empats (Efron, Breslow, Exact i Average), així com models per a una variable resposta binària (logit i clog-log), amb la finalitat d'abordar un temps discret en anàlisis de supervivència. Aquestes metodologies han estat aplicades a un estudi amb dades de cavalls de carreres pura sang i van permetre identificar els factors de risc associats a un esdeveniment d'interès, la lesió musculoesquelètica catastròfica (CMI). Les covariables estadísticament significatives van ser: si hi havia una lesió anterior, el nombre de carreres on va participar el cavall i la longitud de la carrera. El gènere i l'època, donat el seu interès des del punt de vista veterinari, també van ser incloses en les anàlisis. En aquest estudi es va observar similitud entre les estimacions obtingudes en el model de Cox amb els diferents tractaments d'empats i els models discrets logit i clog-log. Es van establir tres grups de resultats: (1) estimacions proporcionats pel mètode exact i el model discret logit; (2) models proporcionats pel model de Cox amb els mètodes de tractament d'empats Efron i Average, i el model discret clog-log; (3) estimacions proporcionades pel model de Cox amb tractament d'empats Breslow.
A la segona part d'aquesta tesi es van introduir i usar els mètodes per abordar models en temps discret en presència d'heterogeneïtat no observada, incloent un o dos termes de frailty. Aquests models es van aplicar a un conjunt de dades reals on l'objectiu va ser caracteritzar el temps (en nombre de lactàncies) fins al primer diagnòstic de mastitis en vaques de producció de llet. Es va poder constatar que entre les variables fixes, el tipus de munyiment va ser sempre estadísticament significativa. A més, també es va obtenir l'efecte ramat, resumit en el terme de frailty. En considerar un segon terme de frailty

corresponent a la zona geogràfica, també va resultar estadísticament significatiu.
A la tercera part d'aquesta tesi s'ha realitzat una comparació de tres programaris disponibles (R, Stata i SAS) per abordar dades d'anàlisi de supervivència per temps discret. Aquesta comparació s'ha realitzat amb l'estudi del diagnòstic de mastitis en vaques lleteres, i per models amb un o dos termes de frailty. Es van establir tres grups de resultats: (1) format pel model de Cox amb mètode d'empats Average i el model discret clog-log; (2) format pel model de Cox amb mètode d'empats Exact i el model logit; (3) format pel model de Cox amb mètode de tractament d'empats Breslow.
Finalment, cal destacar que en aquesta tesi es posa en relleu la importància de considerar la naturalesa discreta del temps, en estudis d'anàlisi de la supervivència. A més, es considera també l'avantatge de recollir la influència mitjançant un o més termes de frailty, de l'heterogeneïtat no observada.

Summary

In this work we have studied and compared methods to deal discrete times in survival analysis. The main focus has been in the analysis of real data from veterinary medicine. First, in order to deal with discrete time data, have been introduced the Cox model handling for ties (Efron, Breslow, Exact and Average) and models for a binary response variable (logit and clog-log). These methodologies have been applied for analysing a dataset of Thoroughbred racehorses. The results allowed identifying a set of risk factors associated with the event of interest, a catastrophic musculoskeletal injury (CMI). Statistically significant covariates were: a dummy variable of previous injury, the number of races participated by the horse and the length of the race. The analysis were also controlled by gender and season, even though they didn't be statistically significant. In this study we got very close results between the Cox model with different methods for handling ties and the discrete-time models with links logit and clog-log. Three groups were established: (1) results obtained by the Cox model with the method Exact and the discrete logit model; (2) results obtained with methods Efron and Average, and clog-log discrete model; (3) results provided by the Cox model with method Breslow. In the second part of this thesis have been introduced methods to deal with discrete-time models in presence of unobserved heterogeneity, including one or two terms of frailty. These models were applied to a real data set where the main goal was to characterize in dairy cows the time (in number of lactations) until the first diagnosis of mastitis. The milking type was always statistically significant, among the fixed covariates. In addition, a herd effect (summarized in the term of frailty) was also obtained. When considering a second term of frailty corresponding to the geographical area, was also statistically significant. In the third part of this thesis, were compared three available software (R, Stata and SAS), to deal with discrete time survival data. This comparison has been carried out

for the data of the diagnosis of mastitis in dairy cows. Models with one or two terms of frailty have been considered. Three groups of results were established: (1) results from the Cox model with method Average and the discrete clog-log model; (2) results from the Cox model with method Exact and the logit model; (3) results from the Cox model with treatment of ties using the Breslow method. Finally, we emphasize that this thesis highlights how important is consider the discrete nature of time. Moreover, the use of specific models with frailty for taking into account possible unobserved heterogeneity.

Producción

Este trabajo ha dado lugar a la publicación de las comunicaciones orales, póster y ponencias a Congresos:

Barroeta Rojo C, Julià O., Espinal A, Morales A. Discrete-time models in Survival Analysis. An application for the time to an exposed fracture in racehorses. XIII Conferencia Española y III Encuentro Iberoamericano de Biometría. Barcelona, 7-9 Septiembre 2011.

Barroeta Rojo C, Julià O., Espinal A, Morales A. Frailty Discrete-time models of Survival Analysis: Fracture in Thoroughbred racehorse with a sire random effect. XIV Conferencia Española de Biometría. Ciudad Real, 22-24 Mayo 2013.

Vazquez, A., Espinal A, Julià O, Barroeta Rojo Carolina, Morales A. Comparación de modelos para tiempos de supervivencia discretos. XXXIV Congreso Nacional de Estadística e Investigación Operativa. Castellón, 10-13 Septiembre 2013.

Barroeta Rojo C, Julià O., Espinal A. Different estimation methods for discrete survival time data with frailty. XV Conferencia Española y V Encuentro Iberoamericano de Biometría. Bilbao, 22-25 Septiembre 2015.

Lista de Abreviaturas

AIC: criterio de información Akaike.
BIC: criterio de información bayesiano.
CMI: lesiones musculoesqueléticas catastroficas.
Cov: Covariables.
Dev Est: Desviación estándar.
EE: error estándar.
EM: esperanza-maximización.
HL: verosimilitud jerárquica.
HR: hazard ratio.
IC: intervalo de confianza.
ICC: correlación intraclase.
IT: indicadores de tiempo.
np: número de parámetros.
OFS: otros sitios anatómicos como metacarpo, carpo y radio-cubito.
PH: riesgos proporcionales (proportional hazards).
PSB: hueso sesamoideo proximal.
SubM: submodelos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

PRESENTACIÓN

Las metodologías de análisis de supervivencia a menudo se citan bajo diferentes nombres: modelos de duración, análisis del tiempo hasta un evento de interés, técnicas de fiabilidad o análisis de la historia de un evento. En la práctica éstas técnicas se aplican a un conjunto de datos, donde la variable de interés es un tiempo hasta un evento concreto, por lo que los valores de la variable serán positivos, lo cual descarta la asunción de normalidad. Algunas distribuciones alternativas serán la exponencial, Weibull o Gamma (Cox, 1972).

Es relevante hacer notar que cuando se registra el tiempo hasta un evento se puede tener información incompleta, por ejemplo puede ocurrir que algún individuo abandone antes que ocurra el evento de interés, lo que conlleva a observaciones censuradas de la variable tiempo. Las técnicas del análisis de supervivencia permiten incluir también estas observaciones en los análisis, lo que representa una fortaleza frente a considerar estos datos como perdidos.

Frecuentemente características del individuo pueden ser de interés, donde las aplicaciones de estos métodos van más allá de evaluar la supervivencia global de los individuos. Por ejemplo, caracterizar el efecto de covariables tales como edad, raza, peso, presencia de enfermedad, sobre el riesgo de experimentar una enfermedad fatal en determinados individuos.

En esta línea las técnicas de análisis de supervivencia, son utilizadas en estudios en diversas áreas como en el ámbito clínico, social, fiabilidad industrial o tecnología de

alimentos. Ejemplos de su aplicación lo encontramos en estudios de: tiempo que demora un trabajador expuesto en desarrollar una enfermedad profesional, tiempo que demora un tratamiento en ser efectivo, tiempo de adquirir un aprendizaje determinado, tiempo de rechazo de un alimento por parte de los consumidores, tiempo de demora en culminar un trabajo o duración de la lactancia materna.

En áreas de veterinaria, se han aplicado técnicas de análisis de supervivencia en el ámbito clínico, reproducción y producción animal. En estos casos el evento de interés puede definirse según el enfoque: recurrencia de una enfermedad, eficacia de una intervención clínica, longevidad de animales de producción (carne o leche) o caracterización del periodo de lactación. Así es posible, estimar el tiempo de supervivencia hasta una respuesta, muerte, recaída o desarrollo de un determinado evento, incluyendo todos los animales de la muestra, tanto si han desarrollado o no el evento de interés .

Hay muchas situaciones en la práctica donde la información de un determinado suceso o proceso no se han medido con un tiempo continuo. (como normalmente los métodos tradicionales lo requieren). Así, en algunos casos se registran tiempos agrupados (días, semanas, meses, años) y en otros casos se dispone de tiempos propiamente discretos (número de partos, número de lactancias, grado escolar cursado, número de hijos). El análisis de este tipo de datos, puede abordarse utilizando el modelo de Cox con tratamientos de empates (Therneau & Grambsch, 2000). Pero esta situación también puede ser abordada utilizando los modelos para variables de respuesta binaria con links logit y clog-log. La aplicación de estos últimos modelos se puede llevar a cabo mediante una conversión de los datos originales (Singer & Willett, 2003).

Adicionalmente, es necesario recordar que los métodos tradicionales en análisis de supervivencia asumen que las poblaciones son homogéneas y que los individuos tienen el mismo riesgo de experimentar el evento, condicionado al valor de las covariables. Pero en algunas situaciones de experimentación y análisis, es razonable asumir que no se pueden observar todas las covariables, ya sea porque son difíciles de medir o porque no están disponibles: heterogeneidad no observada, que se puede pensar como la variabilidad entre individuos debida a características no medibles. Ésta se aprecia por ejemplo cuando hay grupos que comparten las mismas características, es decir, comparten áreas geográficas, instituciones educativas o de salud, familias, granjas, países, o en los casos donde los individuos presentan eventos repetidos. Esta heterogeneidad no observada se incluye como un efecto aleatorio y se llama frailty.

El presente capítulo presenta en la sección 1.1. los objetivos de la tesis y la sección 1.2 describe la estructura como ha sido desarrollada la tesis y se mencionan las bases de datos utilizadas en esta tesis.

1.1. Objetivos

De acuerdo a lo anteriormente expuesto, el objetivo principal de esta tesis es estudiar, analizar y comparar las metodologías de análisis de supervivencia que permiten abordar tiempos discretos, donde se consideró modelos convencionales y modelos que incluyen heterogeneidad no observada (frailty). Los modelos presentados en esta tesis, han sido aplicados a dos bases de datos: una proveniente del hipódromo la Rinconada de Venezuela y otra relacionada a datos de diagnósticos de mastitis, en vacas lecheras de los estados Mérida, Trujillo y Zulia de Venezuela.
El desarrollo de esta tesis contribuirá en el diseño y desarrollo de futuras investigaciones en el área de la veterinaria para predecir y determinar factores involucrados con el riesgo de que ocurra el evento objeto de estudio cuando el tiempo es discreto. Los objetivos específicos planteados en esta tesis doctoral son los siguientes:

1. Describir y aplicar el modelo de regresión logística utilizando la transformación o expansión del conjunto de datos originales.
2. Describir y comparar los modelos para el análisis de supervivencia de tiempos discretos (modelo logit y modelo clog-log), con el modelo de Cox con tratamiento de empates.
3. Aplicación de los modelos mencionados al estudio del tiempo hasta una lesión musculoesquelética catastrófica (CMI) en caballos de carreras de la raza pura sangre. Se trata no sólo de valorar si determinados factores afectan o no el riesgo del evento de interés, sino también, en que sentido y proporción afectan.
4. Análisis del número de lactancias hasta el primer diagnóstico de mastitis en vacas lecheras mediante un modelo discreto con frailty. Estimar el efecto de las covariables fijas así como el o los términos de heterogeneidad no observada (frailty).

5. Comparar tres software disponibles R, Stata y SAS para analizar modelos con y sin término de frailty en análisis de supervivencia para tiempo discreto.

1.2. Estructura y Desarrollo

Para el desarrollo y alcance de los objetivos planteados anteriormente se tienen las siguientes etapas: revisión de antecedentes y presentación de los modelos; base de datos utilizadas en esta tesis; aplicación de los modelos considerados en este estudio y comparación de software; discusión, conclusión y trabajos futuros, que son presentados a continuación.

1.2.1. Revisión de antecedentes y presentación de los modelos

La primera etapa contempló la recolección de la información teórica necesaria, para sustentar el desarrollo de esta investigación y revisó el uso del análisis de supervivencia en medicina veterinaria, especialmente estudios en análisis de supervivencia con datos en tiempos discretos.

Métodos utilizados para analizar tiempos discretos en análisis de supervivencia

Prentice & Gloeckler (1978) desarrollaron un modelo discreto de supervivencia, basándose en el modelo de riesgo proporcional de Cox (1972), donde asumen que los tiempos de duración están registrados y agrupados en intervalos. Allison (2010); Therneau & Grambsch (2000) describen los diferentes métodos para el tratamiento de empates. Más tarde Singer & Willett (1993), describe detalladamente una vía alternativa, para obtener estimadores de máxima verosimilitud de los parámetros en modelos discretos usando modelos logit y clog-log, por medio de métodos de regresión logística utilizando una conversión de la base de datos originales. A partir de sus estudios una cantidad relevante de trabajos han sido desarrollados en áreas relacionadas a las ciencias sociales.

Métodos utilizados para analizar tiempos discretos en análisis de supervivencia con efectos aleatorios (frailty)

En la práctica, puede ocurrir que los individuos se disponen formando grupos de manera natural, ya sea formando comunidades, familias, camadas de animales, agrupación por zonas geográficas.

En recientes años muchos estudios han sido publicados, donde se han presentado extensiones de modelos clásicos de análisis de supervivencia como modelos que son adecuados para el análisis de efectos aleatorios. En estos estudios una amplia variedad de modelos de frailty con diversas técnicas numéricas para ajustar estos modelos han sido presentadas. Sin embargo, sólo un número limitado de libros han detallado aspectos relacionados a los modelos de frailty (Aalen et al., 2008; Duchateau & Janssen, 2008; Hougaard, 2000; Therneau & Grambsch, 2000).

Adicionalmente, otros estudios han utilizado la combinación de modelos de regresión logística con un conjunto de datos expandidos lo que ha permito ajustar modelos de supervivencia en tiempos discretos con efecto aleatorio. Así, Hedeker et al. (2000) discutió éstos modelos con tiempos agrupados usando las funciones link logit y clog-log; por su parte (Barber et al., 2000) demostró que el modelo de análisis de supervivencia con tiempo discreto y multinivel pueden ser estimados usando funciones de software de regresión logística multinivel estándar.

Modelos de análisis de supervivencia con tiempo discreto en ciencias veterinarias

En diferentes áreas de veterinaria se han encontrado un conjunto de investigaciones que utilizan diversos métodos de análisis de supervivencia y algunas de sus extensiones. Dentro del campo de la producción animal es en el área de la genética, específicamente en estudios de longevidad donde con mayor frecuencia se ha utilizado el análisis de supervivencia, el estudio de Famula (1981) fue el primero en desarrollarse en esta área. Más tarde Ducrocq & Casella (1996) uso el análisis de supervivencia en el área de la me-

jora genética, quien propuso bajo un enfoque bayesiano. Además desarrollo el software estadístico Survival Kit (Sölkner & Ducrocq, 1996) muy usado en estudios de datos de longevidad, donde en los modelos de frailty utiliza la estimación bayesiana, con distribuciones: gamma, normal y normal-multivariante, en esta última se puede incluir matrices de relaciones de parentescos. Posteriormente otros trabajos en diferentes áreas de veterinaria han asumido algunas propuestas desarrolladas en Ducrocq (Casellas et al., 2007; Legrand et al., 2005; Tarrés et al., 2005).
Muchos trabajos de análisis de la supervivencia en el área de la veterinaria se han desarrollado utilizando modelos en tiempo continuo; por el contrario pocos son los estudios donde han considerado tiempos discretos, por ejemplo, en el campo de la salud animal, un estudio fue desarrollado por Evans & Sayers (2000) para determinar los factores de riesgos de infección por campylobacter en granjas avícolas. Lo mismo ocurre con los modelos para tiempo discreto con terminos de frailty. Por ejemplo, dentro del área de epidemiología hemos encontrado un estudio (Southey et al., 2003) que se refiere a la mortalidad de ovejas que utilizan análisis de supervivencia en tiempo discreto con frailty. Otro estudio (Hudson et al., 2014) relacionado con fertilidad utilizó un modelo de análisis de supervivencia con tiempo discreto multinivel, con una estructura jerárquica a tres niveles.

1.2.2. Desarrollo de la tesis

En este capítulo, se presenta un esbozo del planteamiento a desarrollar, la formulación de los objetivos a dar respuesta, así como también la estructura y contenidos a considerar en la presente tesis. Los capítulos 2, 3 y 4, recopilan la revisión teórica y antecedentes relacionados con los modelos estadísticos utilizados. Así, el capítulo 2, menciona los aspectos básicos de análisis de supervivencia y los requerimientos mínimos para su aplicación; en el capítulo 3 se ha realizado una revisión y comparación del modelo de Cox con las diferentes técnicas de abordar los empates, los modelos de tiempo discreto de análisis de supervivencia logit y clog-log; y el capítulo 4, explica los modelos discretos de análisis de supervivencia y modelos de análisis de supervivencia con términos de frailty (heterogeneidad no observada). Al final del presente capítulo se presenta un esquema (Figura 1.1), que corresponde a la estructura de las etapas desarrolladas en la presente tesis.

1.2.3. Bases de datos utilizadas en esta tesis

Los modelos discretos logit y clog-log utilizados en el desarrollo de esta tesis, requieren que los datos sean convertidos o transformados a un conjunto de datos expandidos. En la base de datos donde se ha realizado esta transformación, cada individuo se replicó, según el tiempo transcurrido hasta presentarse el evento o quedar censurado; esto será explicado en detalle en el capítulo 3. Las bases de datos presentadas en esta tesis, fueron estructuradas en dos formas, una como base de datos originales y otra como una base de datos expandidos, según fuese el caso del modelo abordado.

Datos de caballos de carreras de la raza pura sangre

Los datos para la realización de este estudio fueron suministrados por el Departamento de patología y hospital veterinario del hipódromo la Rinconada de Caracas-Venezuela. Estos datos fueron recopilados y registrados en la división de Sanidad animal del Instituto Nacional de hipódromos de Venezuela. Esta base de datos contempla las lesiones musculoesqueléticas que ocurrieron entre los años 2000 a 2011, en caballos de carreras pura sangre; el estudio se caracteriza porque el tiempo (edad del caballo) es considerado discreto y presenta censuras.

Datos de vacas con mastitis

Estos datos fueron recolectados entre las zonas Mérida, Zulia y Trujillo como parte de un proyecto a nivel regional. En este estudio se definió el tiempo como el número de lactancias hasta el primer diagnóstico de la infección con mastitis; por lo tanto, es un tiempo discreto. Los datos incluyeron además de características tanto del propio animal como de manejo, las variables rebaños y zonas geográficas, las cuales fueron utilizadas para el estudio como heterogeneidad no observada.

1.2.4. Aplicación de los modelos considerados en este estudio y comparación de funciones de software

Aplicación de los modelos de análisis de supervivencia en tiempos discretos

Con la finalidad de dar respuestas a los objetivos de la presente tesis, en el capítulo 5, se utilizó el conjunto de datos referidos a las lesiones musculoesqueléticas catastróficas (CMI) en caballos de carreras pura sangre. Este conjunto de datos fue analizado con técnicas para tiempo discreto, considerando los siguientes modelos: (a) modelo de Cox, el cuál se utilizó con las diferentes técnicas de abordar los empates (Efron, Breslow, Exact y Average), (b) modelos en tiempos discretos logit y clog-log.

Comparación de distintos software estadísticos para tiempos discretos y con términos de frailty

En esta parte de la tesis se analizaron los datos de vacas lecheras. Se desarrolló una aplicación con métodos que permiten incluir un término de frailty. En los capítulos 6 y 7 se presentan los resultados. Primero se presentan los modelos incluyendo términos de frailty y caracterizando los factores para el diagnóstico de mastitis. En el capítulo 7 se realiza una comparación de los software R, Stata y SAS, para abordar un análisis de supervivencia con tiempos discretos con y sin términos de frailty.

1.2.5. Discusión, conclusiones y trabajos futuros

Para finalizar en el capítulo 8 se desarrolló una tercera parte, donde se plantea una discusión general y se formulan las conclusiones finales. Finalmente en el capítulo 9, se plantean nuevos trabajos para continuar en esta línea de investigación, en algunas áreas de las ciencias veterinarias.

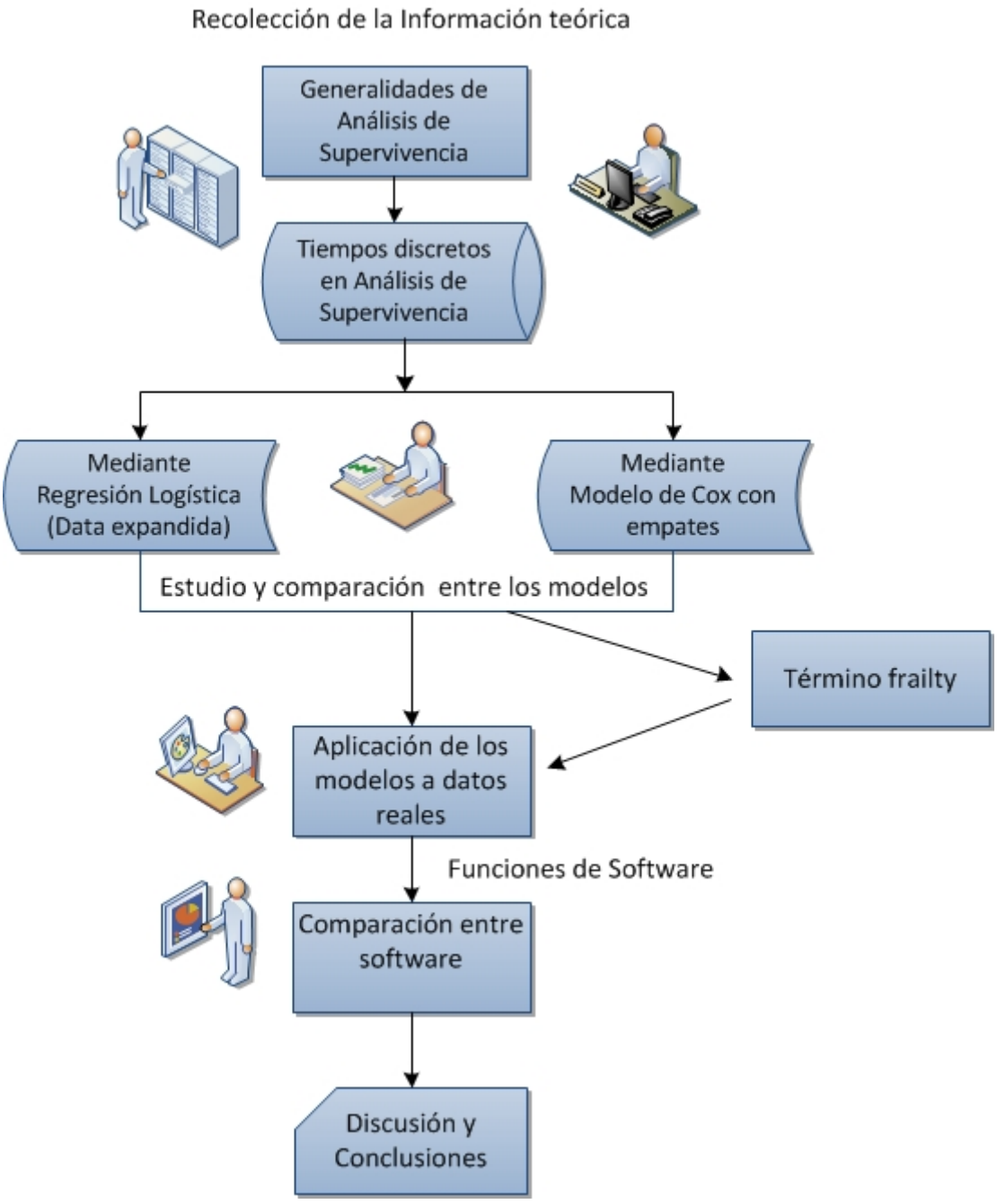

Figura 1.1: Etapas de la estructura del desarrollo de la Tesis

Capítulo 2

GENERALIDADES DE ANÁLISIS DE SUPERVIVENCIA

2.1. Introducción

El análisis de supervivencia es la parte de la estadística que estudia variables aleatorias que representan tiempos hasta un evento de interés. Un evento frecuentemente estudiado es la muerte, sin embargo no es el único evento que puede ser de interés, cualquier situación en la que se mide el tiempo hasta un evento especifico, será una posible variable para este tipo de análisis; como puede ser el tiempo hasta un aprendizaje determinado o tiempo de recurrencia de una enfermedad.

Una característica en éstos análisis es que la información sobre el tiempo hasta el evento de algunos individuos es incompleta, registrándose solo información parcial sobre el tiempo de interés. Por ejemplo, si el evento de interés es la primera infección de neumonía en becerros, todos aquellos animales sanos al finalizar el estudio contribuyen como información parcial sobre la ocurrencia del evento.

Cuando se está interesado en determinar factores que están influyendo en la ocurrencia de un evento es importante distinguir entre los métodos de regresión logística y modelos de supervivencia que consideran variables explicativas como por ejemplo el modelo de Cox. Así, la regresión logística estudia una variable respuesta binaria que registra la ocurrencia o no de un evento en particular en un periodo de tiempo fijo, es decir sin tener en

cuenta el momento exacto en que ha sucedido el evento, mientras que en el análisis de la supervivencia se tiene en cuenta no sólo si el evento ha sucedido o no, sino también cuando ha sucedido. El presente capítulo es una breve revisión de algunos conceptos básicos y terminología del análisis de supervivencia relacionados con el contenido de esta tesis. En la sección 2.2 se tratan las características relevantes del análisis de supervivencia, en la sección 2.3 se presentan las funciones relevantes, en la sección 2.4 se describen los estimadores de Kaplan-Meier (para la función de supervivencia) y Nelson-Aalen (para la función de riesgo acumulado), se presenta el modelo de Cox y se mencionan los métodos diagnósticos.

2.2. Características relevantes del análisis de supervivencia

2.2.1. Tiempo

La característica relevante del análisis de supervivencia es que la variable de estudio es el tiempo hasta un evento de interés; por tanto, al inicio del estudio, este debe estar inequívocamente definido y también hay que determinar de forma clara cual es el momento inicial.

2.2.2. Censura

En el análisis de supervivencia, el tiempo exacto hasta que se produce el suceso en algunos individuos no puede llegar a observarse, teniendo entonces información incompleta. Esto se presenta por ejemplo cuando el evento de interés ocurre antes de que un individuo entre al estudio o cuando el estudio termine antes que el evento sea observado en el individuo. Esta característica del análisis de supervivencia es conocida como censura (Klein & Moeschberger, 2003). Se asume que los individuos censurados quedan representados por los no censurados; a este tipo de censura se la conoce como censura no informativa. En particular esto se cumple cuando el tiempo y la censura son independientes, lo cual se conoce como censura aleatoria.

Censura por la derecha

Tenemos censura por la derecha cuando para un individuo el tiempo (T) hasta el evento de interés no es observado pero se sabe que es mayor a un tiempo C (último tiempo de seguimiento). En este caso, los datos vienen representados por un par de variables aleatorias (Y, δ) definidas como:

$$Y = \text{mín}\{T, C\}, \qquad \delta = \begin{cases} 1 & \text{si } T \leq C\text{: dato no censurado} \\ 0 & \text{si } T > C\text{: dato censurado.} \end{cases} \tag{2.1}$$

La variable aleatoria δ es el indicador de no-censura, aunque usualmente se le conoce como el indicador de censura.
En general, para una muestra de n individuos, observaremos $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$ donde

$$Y = \text{mín}\{T_i, C_i\}, \qquad \delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i\text{: el individuo } i \text{ no está censurado} \\ 0 & \text{si } T_i > C_i\text{: el individuo } i \text{ está censurado.} \end{cases} \tag{2.2}$$

La censura por la derecha puede ser censura tipo I (fija, progresiva o generalizada) y censura tipo II. Tenemos censura tipo I cuando el evento se observa sólo si ocurre antes de un tiempo preespecificado (censura). Diremos que es fija si la censura es la misma para todos los individuos; progresiva si tenemos diferentes censuras para distintos grupos de individuos y generalizada si la censura es distinta para cada individuo.
Tenemos censura de tipo II cuando se decide finalizar el estudio después de que hayan ocurrido un número predeterminado de (r) eventos, es decir, el estudio continúa hasta que se ha observado el r-ésimo evento.

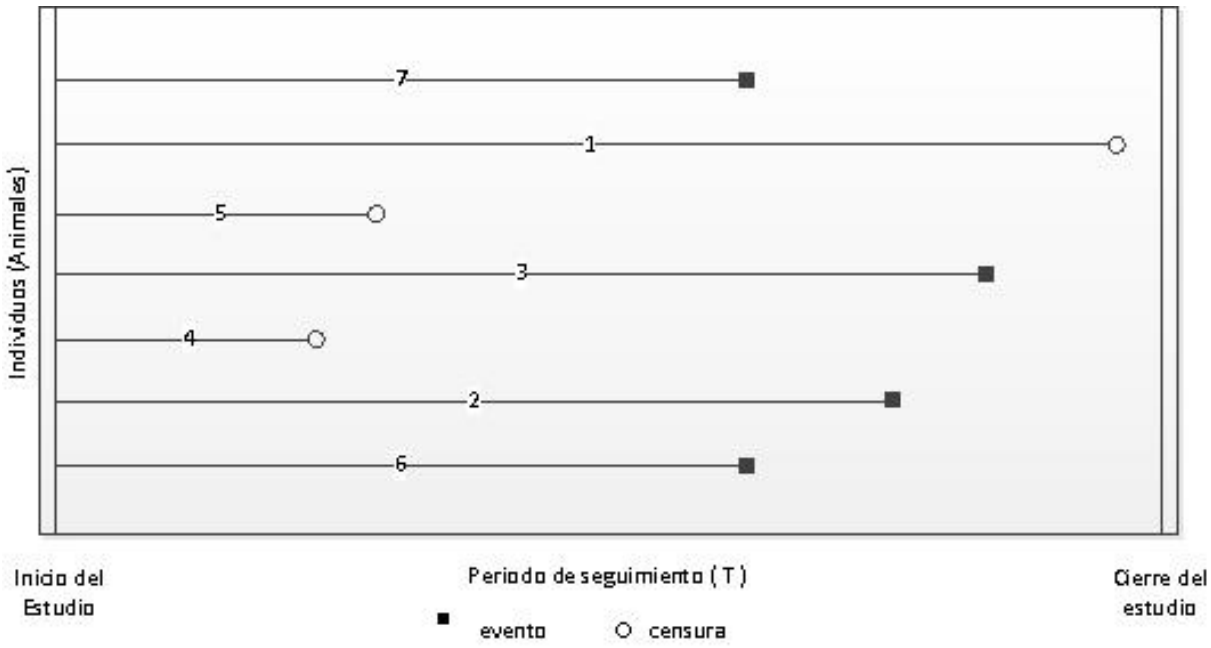

Figura 2.1: Esquema de los datos censurados por la derecha

Un ejemplo de lo que ocurre cuando tenemos censura por la derecha, lo encontramos en la figura 2.1, donde se han representado los tiempos de 7 individuos. Así, los que han sido identificados como 2, 3, 6 y 7 experimentaron el evento de interés mientras que los individuos 1, 4 y 5 sus tiempos esta censurado (información incompleta).

Censura en intervalo

Tenemos censura en un intervalo cuando desconocemos el valor del tiempo hasta el evento de interés pero sabemos que se encuentra en un intervalo. Por ejemplo cuando un evento se produce entre dos visitas médicas, sabemos que se ha producido entre dos tiempos, pero no sabemos el valor exacto.

Censuras por la izquierda

Una observación se dice censurada por la izquierda si se desconoce el valor exacto del tiempo hasta el evento de interés, pero se sabe que ha ocurrido antes del inicio del

seguimiento del individuo.

2.3. Funciones relevantes

2.3.1. Función de supervivencia

La función de supervivencia se define como la probabilidad de que un individuo sobreviva (no le haya ocurrido el evento de interés) al momento t. Una definición más formal puede darse de la siguiente manera: sea T una variable aleatoria positiva (o no negativa), que representa el tiempo hasta el evento de interés de un individuo. Se define la función de supervivencia como:

$$S(t) = P(T > t). \tag{2.3}$$

La función de distribución se define como $F(t)$=P(T $\leq t$) para $t \geq 0$, y representa la probabilidad de que un individuo le ocurra el evento antes del tiempo t. Se relaciona con la función de supervivencia de la forma:

$$F(t) = 1 - S(t). \tag{2.4}$$

1 Si T es continua, con densidad $f(t)$, la función de supervivencia y la función de distribución son obtenidas por la integración de la función de densidad:

$$S(t) = P(T > t) = \int_t^{\infty} f(s)ds, \quad F(t) = P(T \leq t) = \int_0^t f(s)ds.$$

2 Si T es discreta y toma los valores $t_1 \leq t_2 \leq \cdots$, la función de supervivencia y la función de distribución son calculadas por adición de la función de masa de probabilidad $p(t_j)$= $P(T = t_j)$:

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j), \quad F(t) = P(T \leq t) = \sum_{t_j \leq t} p(t_j)$$

2.3.2. Función de Riesgo

La función de riesgo es ampliamente usada en el análisis de supervivencia, pues es una formalización de la idea intuitiva de riesgo y es fácilmente interpretable.

1 Si T es una variable continua con densidad $f(t)$, la función de riesgo se define como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t \leq T < t + \Delta t | T \geq t]. \tag{2.5}$$

La función λ se interpreta como la probabilidad instantánea de que para un individuo se produzca el evento de interés en t, sabiendo que antes de t no se había producido. Además, λ(t) puede expresarse como:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t)). \tag{2.6}$$

Por tanto la función de supervivencia tiene la expresión:

$$S(t) = \exp\left\{ -\int_0^t \lambda(s) ds \right\}. \tag{2.7}$$

2 Si T es una variable discreta y toma los valores $t_1 < t_2 < \cdots$, con función de masa de probabilidad $p(t_j)$, entonces la función de riesgo se define como:

$$h(t_j) = P(T = t_j | T \geq t_j) = P(T = t_j | T > t_{j-1}).$$

Tenemos las siguientes expresiones:

$$h(t_j) = \frac{p(t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}, \tag{2.8}$$

$$S(t) = \prod_{j: t_j \leq t} (1 - h(t_j)). \tag{2.9}$$

Es importante resaltar que la función de riesgo $h(t)$ para el caso de tiempo discreto es una probabilidad mientras que $\lambda(t)$ no lo es, por ello utilizamos notaciones distintas.

2.3.3. Función de Riesgo Acumulado

La función de riesgo acumulado es muy útil técnicamente aunque no tiene interpretación intuitiva clara.

1. Si T es una variable continua con función de riesgo $\lambda(t)$, se define la función de riesgo acumulado como:

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.10}$$

   Utilizando la expresión 2.7 tenemos:

$$S(t) = \exp\{-\Lambda(t)\}, \quad \Lambda(t) = -\ln S(t) \tag{2.11}$$

2. Cuando T es discreta y toma los valores $t_1 < t_2 < \cdots$, con función de riesgo $h(t_j)$, la función de riesgo acumulado se define como:

$$\Lambda(t) = \sum_{j:t_j \leq t} h(t_j). \tag{2.12}$$

   Aquí ya no tenemos la relación (2.11) entre la función de riesgo acumulado y la función de supervivencia.

2.4. Estimaciones

2.4.1. Kaplan-Meier: Estimador de la función de supervivencia

El estimador de Kaplan-Meier es conocido como estimador del limite del producto, el cual está basado en la descomposición de la función de supervivencia en un producto de probabilidades condicionadas (ecuación 2.9). Como método no paramétrico no asume

ninguna distribución particular para la variable tiempo.

$$\hat{S}(t) = \begin{cases} 1 & \text{si} \quad t < Y_{(1)} \\ \prod_{j:Y_j \leq t}(1 - \frac{d_i}{n_i}) & \text{si} \quad t \geq Y_{(1)} \end{cases} \tag{2.13}$$

donde: d_j representa el número de individuos que registraron el evento de interés en el momento t_j; n_j representa el número total de individuos en riesgo en el momento t_j, es decir, que su tiempo es mayor o igual a t_j. Al conjunto de individuos en riesgo se le indica por $R(t)$.

En esta sección se ha utilizado nuestro conjunto de datos de caballos de carreras pura sangre presentados en el Capítulo 1, para estimar la función de supervivencia por Kaplan-Meier. En este conjunto de datos la variable T fue definida como el tiempo hasta que ocurrió el evento de interés (CMI). Dentro de los 214 caballos de carreras que conforman el conjunto de datos, 112 caballos, que representan 52.3 %, experimentaron el evento de intéres, es decir, CMI y 102 (47.7 %) caballos no presentaron CMI y sobrevivieron a este evento, por lo tanto fueron considerados datos censurados.
Para estos tiempos la supervivencia estimada por Kaplan-Meier se muestra en la Tabla 2.1, donde para cada tiempo registrado también se observa: el número de caballos a riesgo en padecer el evento (n_i), el número de caballos que padecieron el evento (d_i), así como la estimación de la supervivencia. En la figura 2.2 está representado el estimador de Kaplan-Meier en este caso.

Tabla 2.1: Estimador de Kaplan-Meier para los datos de los caballos.

Tiempo	d_j	n_j	$\hat{S}_j$	**EE**	**IC Li.**	**IC Ls.**
2	214	42	0.804	0.027	0.752	0.859
3	125	27	0.630	0.036	0.563	0.706
4	76	15	0.506	0.041	0.431	0.593
5	42	15	0.325	0.046	0.247	0.428
6	18	7	0.199	0.047	0.125	0.315
7	9	6	0.066	0.035	0.024	0.186

Nota: donde d_j representa el número de caballos que registraron el evento de interés (CMI) en el momento j; n_j representa el número total de caballos a riesgo en el tiempo j. $\hat{S}_j$ corresponde al estimador de Kaplan-Meier para la supervivencia.

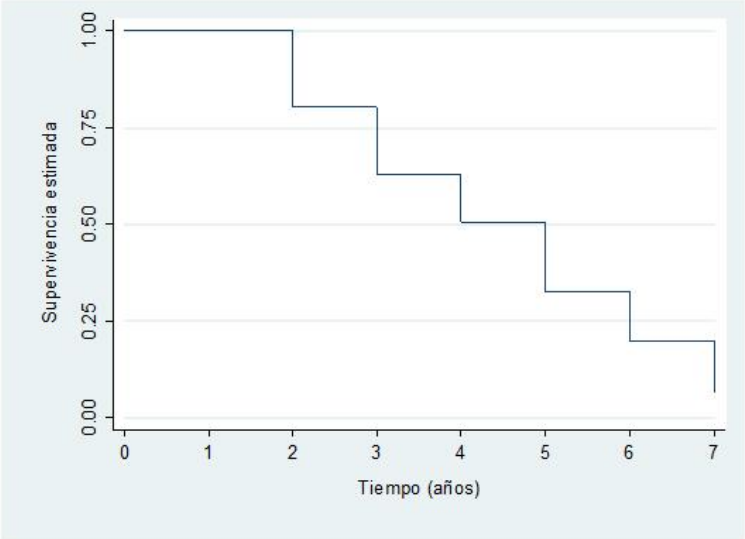

Figura 2.2: Estimación de la función de supervivencia mediante Kaplan-Meier para los datos de caballos de carreras

2.4.2. Nelson-Aalen: Estimador de la función de riesgo acumulado

El estimador de Nelson-Aalen para la función de riesgo acumulado se define de la siguiente forma:

$$\hat{\Lambda}_{NA}(t) = \begin{cases} 0 & \text{si} \quad t < Y_{(1)} \\ \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i} & \text{si} \quad t \geq Y_{(1)}. \end{cases} \tag{2.14}$$

donde igual que antes d_i representa el número de fallos ocurridos en el instante t_i y n_i el número de individuos en riesgo en t_i. Una estimación de la función de riesgo en t_i, a partir del estimador de Nelson-Aalen, se puede obtener como $\frac{d_i}{n_i}$.

2.4.3. Modelo de Riesgos Proporcionales o modelo de Cox

En algunas situaciones de la práctica interesa determinar la influencia de las variables pronósticas o covariables sobre el tiempo de supervivencia de los individuos. El modelo de Cox permite introducir en el modelo la influencia de las covariables. Este modelo fue propuesto por primera vez por Cox (1972), es llamado de riesgos proporcionales, debido a que el cociente entre el riesgo para dos sujetos es constante en el tiempo. Durante años ha sido uno de los modelos más utilizados en análisis de supervivencia, especialmente en la investigación médica, donde ha tenido aplicabilidad en una variedad de estudios clínicos. Por ejemplo, en el campo de la veterinaria se podría estar interesado en modelar la función de riesgo de la mortalidad en becerros en función de covariables tales como: peso al nacimiento, sexo, tipo de manejo, raza, época de nacimiento, género; en este caso, el modelo de Cox con estas covariables podría ser considerado.
El modelo consiste en expresar la función de riesgo en de dos componentes, uno no paramétrico que depende sólo del tiempo y otro paramétrico que depende sólo de las

covariables, de la siguiente forma:

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}} \quad \forall t \geq 0$$

donde λ_0 es la función de riesgo basal (correspondiente a un individuo cuando todas las covariables son cero), $\mathbf{Z}$ es el vector de covariables y $\boldsymbol{\beta}$ sus coeficientes. Al conjunto de covariables se le llama perfil del individuo. El término $\exp \boldsymbol{\beta}'\mathbf{Z}$ representa el cociente de riesgos entre un individuo con covariables $\mathbf{Z}$, respecto a un individuo con covariables $\mathbf{Z} = 0$, no depende del tiempo pero sí de los valores de las covariables, las cuales entran linealmente en el modelo vía los coeficientes de regresión.
Así, la función de riesgo puede depender de variables y de factores. La variable toma valores numéricos normalmente en escala continua, y en el modelo de riesgos proporcionales se incorporan asignando a cada una de ellas un coeficiente β. Por su parte los factores toman un número limitado de valores (niveles del factor) y para incorporar un factor que tiene a niveles se incluirán a -1 coeficientes en el modelo, donde el primer nivel actúa como nivel basal o de referencia.

Métodos diagnósticos para el modelo de Cox

Una parte importante en los análisis consiste en ratificar o evaluar la adecuación del modelo de Cox, para ello disponemos de diferentes procedimientos diagnósticos. Esta verificación del modelo se realiza por medio de los residuos (ver Collett (2003); Gómez (2005); Therneau & Grambsch (2000)).
Se entiende como residuo una función de la diferencia entre el valor observado y el valor estimado, es decir la porción que queda sin explicar para cada individuo. Se han propuesto diferentes formulaciones para los residuos lo cual representa una ventaja al permitir examinar diferentes aspectos relacionados con el modelo de riesgos proporcionales. Frecuentemente estos residuos se representan gráficamente y se observan los patrones establecidos para decidir, junto a las pruebas de hipótesis establecidas, la validez del modelo de Cox, ver Collett (2003); Klein & Moeschberger (2003).
La representación gráfica de los distintos residuos permiten para una covariable especifica, encontrar la mejor forma funcional que explique la influencia de la covariable sobre

la supervivencia; verificar la asunción de riesgos proporcionales; comprobar la precisión de predecir la supervivencia en un sujeto determinado (aquí estamos interesados en los individuos que presentaron el evento demasiado pronto o demasiado tarde en comparación con lo que predice el modelo ajustado), así como examinar la influencia que tiene cada individuo sobre el ajuste del modelo.

Capítulo 3

TIEMPOS DISCRETOS EN ANÁLISIS DE SUPERVIVENCIA

Como se describió en el capítulo 2, cuando la variable tiempo es continua, entre las técnicas de análisis de supervivencia se tiene el modelo de Cox. Sin embargo, en la práctica pueden presentarse situaciones en que la variable tiempo es discreta (ya sea por agrupación de tiempos continuos o por que la variable tiempo por naturaleza sea propiamente discreta), situaciones que pueden ser abordadas mediante ciertas modificaciones al modelo de Cox, donde serán necesarias técnicas para tratar los empates, y obtener así los estimadores adecuados para estos tiempos discretos. Por otra parte, estos datos pueden ser abordados utilizando modelos para una variable respuesta binaria, ambas situaciones se detallan en este capítulo.

3.1. Introducción de los modelos para tiempos discretos en análisis de supervivencia

En análisis de supervivencia, la variable respuesta T puede ser definida como continua o discreta dependiendo como ésta ha sido medida (Singer & Willett, 2003). Si hay muchos empates entre los tiempos de supervivencia, los métodos para datos de tiempo continuo,

tales como el modelo de regresión de Cox, necesitan algunas modificaciones. Es decir, cuando la información esta disponible en tiempos discretos (como mediciones en años, meses), el modelo de Cox puede no ser adecuado. Para una variable aleatoria T discreta se pueden distinguir dos situaciones:

1. Cuando T es estrictamente discreta, esto es, que el evento ocurre solo en tiempos específicos $(t_0, t_1, \dots)$ y el tiempo de supervivencia es tratado como resultados ordenados, por ejemplo, número de nacimientos, número de carreras o número de lactancias, los cuales pueden ser o no censurados. Estos enfoques han sido investigados por Han & Hausman (1990); McCullagh (1980).

2. Cuando T es una variable continua pero los tiempos están agrupados en intervalos $(0, t_1], (t_1, t_2], \dots, (t_{j-1}, t_j]$. En este caso la variable respuesta tiempo, está definida como t_j, si el tiempo observado ocurre dentro del intervalo $(t_{j-1}, t_j]$. Este enfoque ha sido utilizado en psicología y educación (Singer & Willett, 1993)

Ambas situaciones han sido consideradas en esta tesis. La vía más común para caracterizar la distribución de una variable discreta es la función de masa de probabilidad y la función de distribución. En análisis de supervivencia, sin embargo, comúnmente se utiliza la función de riesgo $h(t_j)$, que es una probabilidad condicional, definida como la probabilidad de que el evento de interés ocurra en t_j $(T = t_j)$ sabiendo que antes de t_j el evento no había ocurrido $(T \geq t_j)$.

Varios modelos han sido propuestos en la literatura para abordar el análisis de tiempos discretos. Kalbfleisch & Prentice (1973) propusieron una versión del modelo de riesgos proporcionales para datos agrupados. Un modelo de regresión logística ordinal fue desarrollada por McCullagh (1980). Una comparación entre ambas metodologías ha sido presentada en Grilli (2005). Singer & Willett (1993, 2003), consideraron un enfoque basado en un modelo de regresión logística binaria para estimar los efectos de las covariables en la función de riesgo. En el presente capítulo fueron combinadas las ideas de Singer & Willett (2003) y Grilli (2005), y se compararon los métodos para abordar tiempos discretos en análisis de supervivencia. El presente capítulo esta estructurado en las siguientes secciones: la sección 3.1 presenta una introducción; el modelo de Cox con tratamientos de empates es desarrollado en la sección 3.2; también la Verosimilitud para tiempos discretos de supervivencia se presenta en la sección 3.3; la transformación del conjunto

de datos a un conjunto individuo-periodo se muestra en la sección 3.4; posteriormente los modelos en tiempos discretos para análisis de supervivencia son presentados en la sección 3.5; la alternativas para la comparación de modelos se muestra en la sección 3.6; finalmente en la sección 3.7 se presentan los software estadísticos utilizados.

3.2. Modelo de Riesgos Proporcionales (modelo de Cox)

Sea T un tiempo de supervivencia continuo. Si este se asume que tiene una ley que cumple con los supuestos del modelo de Cox, la función de riesgo presenta la siguiente forma:

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}} \tag{3.1}$$

donde $\mathbf{Z}' = (Z_1, \ldots, Z_p)$ es el vector de covariables y $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ el vector de los coeficientes (lo cual implica que el riesgo cambia proporcionalmente cuando cambian las covariables $\mathbf{Z}$). En los modelos de riesgos proporcionales, dados dos individuos, la proporción entre sus riesgos se asume constante en cualquier punto del tiempo, asumiendo que no hay interacción entre las covariables y el tiempo. Además, el modelo de Cox no requiere especificar la distribución de la función de riesgo basal, lo cual hace que este sea más práctico que otros enfoques, por lo que es usado ampliamente.
La estimación de los parámetros en el modelo de Cox se obtiene maximizando la función de verosimilitud parcial, que se presenta a continuación siguiendo el texto de Kalbfleisch & Prentice (2002).

Notación

Sea $\{(t_i, \delta_i, \mathbf{Z}_i),\ i = 1, \ldots, n\}$ nuestra muestra; donde como siempre t_i son los tiempos, δ_i el indicador de censura y $\mathbf{Z}_i$ las covariables del individuo i.

Supongamos primero que no hay empates ni censuras y sean $\mathbf{O}(\mathbf{t}) = [t_{(1)}, \ldots t_{(n)}]$ el estadístico de orden y $\mathbf{r}(\mathbf{t}) = [(1), (2), \ldots, (n)]$ el estadístico de los rangos. Observemos

que la información contenida en $(t_1, \ldots, t_n)$ es equivalente a $\mathbf{O(t)}$ i $\mathbf{r(t)}$. Definimos $\mathbf{Z}_{(i)}$ y $\delta_{(i)}$ las covariables y el indicador de censura del individuo que corresponde al tiempo $t_{(i)}$. Recordemos que $R(t_{(i)})$ es el conjunto de individuos a riesgo en el momento $t_{(i)}$.

Ejemplo: Si $n = 4$ y $(t_i, \mathbf{Z}_i) = (5, tr), (17, ct), (12, ct), (15, tr)$, tendremos:

$$\begin{aligned}
&\mathbf{O(t)} = [5, 12, 15, 17], \\
&\mathbf{r(t)} = [1, 3, 4, 2], \\
&\mathbf{Z}_{(1)} = \mathbf{Z}_{(3)} = tr, \quad \mathbf{Z}_{(2)} = \mathbf{Z}_{(4)} = ct.
\end{aligned}$$

Verosimilitud parcial sin empates ni censuras

La estimación de los parámetros en el modelo de Cox se hace mediante la función de la verosimilitud parcial, que supone que para la estimación de $\boldsymbol{\beta}$, sólo es necesario conocer el orden de los fallos y no los valores de los tiempos de fallo.

El argumento es el siguiente: supongamos que tenemos el modelo de Cox, es decir, la función de riesgo se expresa como la ecuación 5.1 Consideremos el grupo de las transformaciones estrictamente crecientes y diferenciables $G = \{g : (0, \infty) \longrightarrow (0, \infty);\ g \in \mathcal{C}^1\ g \nearrow\}$. Sean $g \in G$ y $v = g^{-1}(t)$. La distribución condicional de v dado $\mathbf{Z}$ tiene por función de riesgo:

$$\lambda_0(g(v))g'(v)e^{\boldsymbol{\beta}'\mathbf{Z}} = \lambda_1(v)e^{\boldsymbol{\beta}'\mathbf{Z}},$$

donde $\lambda_1(v) = \lambda_0(g(v))g'(v)$. Por tanto, si los datos se hubiesen presentado de la forma $\{(v_i, \mathbf{Z}_i),\ i = 1, \ldots, n\}$ con $g(v_i) = t_i$, el problema de inferencia sobre $\boldsymbol{\beta}$ seria el mismo, dado que λ_0 es totalmente no especificada. Por tanto, la inferencia sobre $\boldsymbol{\beta}$ es invariante bajo el grupo G de transformaciones del tiempo de supervivencia t.

Por otra parte, si consideramos la acción de G sobre el espacio muestral, todo estadístico $\mathbf{O(t)}$ puede ser transformado en cualquier otro por una $g \in G$ conveniente, mientras que el estadístico de rangos permanece invariante. Dado que la estimación de $\boldsymbol{\beta}$ es invariante bajo el grupo G de transformaciones del tiempo, sólo el estadístico de los rangos puede contener información sobre $\boldsymbol{\beta}$. Para estimar $\boldsymbol{\beta}$ utilizamos la distribución marginal de los rangos. Se define la verosimilitud parcial como la probabilidad de haber observado

nuestro vector de rangos, dada por

$$\begin{aligned} P(\mathbf{r};\boldsymbol{\beta}) &= P(\mathbf{r}=[(1),(2),\ldots,(n)];\boldsymbol{\beta}) = \\ &= \int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \prod_1^n f(t_{(i)};\mathbf{Z}_{(i)}) dt_{(n)} dt_{(n-1)} \ldots dt_{(1)} \\ &= \frac{\exp(\sum_1^n \boldsymbol{\beta}'\mathbf{Z}_{(i)})}{\prod_{j=1}^n [\sum_{l\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{Z}_l)]}. \end{aligned} \tag{3.2}$$

Verosimilitud parcial con censuras y sin empates

Para tratar con las censures hay que modificar el argumento anterior, ya que el modelo con censura no es invariante respecto el grupo de transformaciones. Cuando se obtiene una muestra censurada sólo se tiene información parcial sobre el vector de rangos. Por ejemplo si tenemos los tiempos 114, 90^+, 63, 108^+, donde el signo + indica censura, el estadístico de los rangos serán uno de estos:

$$[3,2,4,1] \quad [3,4,2,1] \quad [3,2,1,4]$$

$$[3,4,1,2] \quad [3,1,2,4] \quad [3,1,4,2]$$

De cara a la inferencia sobre $\boldsymbol{\beta}$ se utiliza la probabilidad marginal de que el estadístico de rangos sea uno de los posibles. La verosimilitud se genera con la parte observada del estadístico $\mathbf{r}(\mathbf{t})$. Aunque los valores exactos de los tiempos censurados se ignoren, la invariancia respecto el modelo sin censura sugiere que la distancia entre sucesivos tiempos no censurados es irrelevante. En consecuencia parece razonable suponer que los tiempos reales de las censures relativas a los tiempos no censurados adyacentes no contribuirán en la inferencia sobre $\boldsymbol{\beta}$.

Supongamos que se han observado k individuos con las etiquetas $(1),\ldots(k)$ fallando en los tiempos $t_{(1)} < \ldots < t_{(k)}$ con sus correspondientes covariables $\mathbf{Z}_{(1)},\ldots,\mathbf{Z}_{(k)}$. Supongamos también que m_i individuos con covariables $\mathbf{Z}_{i1},\ldots,\mathbf{Z}_{im_i}$ están censurados en el intervalo $[t_{(i)}, t_{(i+1)})$, $i=0,\ldots,k$ donde $t_{(0)}=0$ y $t_{(k+1)}=\infty$. El conjunto de los posibles rangos se caracterizará por las condiciones:

$$t_{(1)} < \ldots < t_{(k)}, \qquad t_{(i)} < t_{i1}, \ldots t_{im_i}, \quad i = 0, 1, \ldots, k$$

Donde $t_{i1}, \ldots t_{im_i}$ son los tiempos no observados de los individuos censurados en $[t_{(i)}, t_{(i+1)})$. Dado $t_{(i)}$, el suceso $t_{(i)} < t_{i1}, \ldots t_{im_i}$ (los individuos con covariables $\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{im_i}$ sobreviven a $t_{(i)}$) tiene la probabilidad condicionada:

$$p(t_{(i)}) = \exp\left[-\sum_{j=1}^{m_i} \exp(\boldsymbol{\beta}'\mathbf{Z}_{ij}) \int_0^{t(i)} \lambda_0(u)du\right] \qquad i = 0, \ldots, k.$$

La verosimilitud parcial es ahora proporcional a:

$$\begin{aligned} &\int_0^\infty \int_{t_{(1)}}^\infty \ldots \int_{t_{(k-1)}}^\infty \prod_{i=1}^k f(t_{(i)}; \mathbf{Z}_{(i)}) p(t_{(i)}) dt_{(k)} dt_{(k-1)} \ldots dt_{(1)} \\ &= \frac{\exp(\sum_1^k \boldsymbol{\beta}'\mathbf{Z}_{(i)})}{\prod_{i=1}^k [\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{Z}_l)]}. \end{aligned} \tag{3.3}$$

Observemos que esta no es la probabilidad de observar las censuras tal como las hemos visto, ya que esta probabilidad depende del mecanismo de la censura y seguramente también de $\lambda_0(t)$; sino que es la probabilidad de que, bajo la versión del experimento sin censuras, el suceso observado se produzca. Es decir, es la probabilidad de todos los vectores de rangos posibles según nuestra muestra.

3.2.1. Modelo de Cox con empates

Supongamos que observamos $t_{(1)} < \ldots < t_{(k)}$ tiempos diferentes de muerte, y en cada $t_{(i)}$ tenemos d_i muertes. Sea $\mathbf{S}_i$ el vector que resulta de la suma de las covariables de los d_i individuos que mueren en $t_{(i)}$. Si observáramos el orden en que se han producido las muertes, $t_{(i1)} < \ldots, < t_{(id_i)}$, en la verosimilitud tendríamos el producto:

$$\prod_{i=1}^k \frac{\exp(\sum_{j=1}^{d_i} \boldsymbol{\beta}'\mathbf{Z}_{i,j})}{\prod_{j=1}^{d_i} \left[\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{Z}_l)\right]} = \prod_{i=1}^k \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\prod_{j=1}^{d_i} \left[\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{Z}_l)\right]} \tag{3.4}$$

Se trata de dar diferentes alternativas para los denominadores de (3.4), siguiendo la denominación empleada en Therneau & Grambsch (2000).

1. **Exact**

   El mismo razonamiento utilizado para incorporar censuras se puede aplicar en el caso de empates. Aquí otra vez tenemos solo información parcial del estadístico de rangos. Sabemos cuáles son los rangos que les corresponden al conjunto de individuos que han fallado en $t_{(i)}$ pero la repartición de los d_i rangos se desconoce. La probabilidad de que un vector de rangos sea uno de estos posibles dada la muestra, es una suma de $\prod_1^k d_i!$ términos del tipo (3.2).

   Este cálculo queda simplificado haciendo que la asignación de los rangos de los d_i individuos que fallan en $t_{(i)}$ no quede afectada por la asignación de los rangos de los d_j individuos que fallan en $t_{(j)}$. La suma queda entonces reducida al producto de k sumas, una para cada tiempo de fallo. Sea $\mathbf{Q}_i$ el conjunto de permutaciones de los símbolos $i_1, \ldots, i_{d_i}$ y $\mathbf{P}= (p_1, \ldots, p_{d_i})$ un elemento de $\mathbf{Q}_i$. Como siempre $R(t_{(i)})$ es el conjunto de riesgo y definimos por $R(t_{(i)}, p_r)$ al conjunto que resulta de la diferencia $R(t_{(i)}) - \{p_1, \ldots, p_{r-1}\}$. Entonces, la verosimilitud parcial para $\boldsymbol{\beta}$ es:

$$\prod_{i=1}^{k} \left\{ \exp(\mathbf{S}_i\boldsymbol{\beta}) \sum_{\mathbf{P}\in\mathbf{Q}} \prod_{r=1}^{d_i} \left[\sum_{l\in R(t_{(i)},p_r)} \exp(\mathbf{Z}_l\boldsymbol{\beta}) \right]^{-1} \right\} \tag{3.5}$$

   La notación en (3.5) es lo suficientemente general como para admitir datos censurados.

2. **Breslow**

   El cálculo de (3.5) puede ser muy engorroso cuando el número de empates es grande en cada uno de los tiempos. Si el número de individuos que fallan, d_i, es pequeño comparado con el número de individuos en riesgo, la expresión (3.5) se puede aproximar bien por:

$$\prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\left[\sum_{l\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{Z}_l)\right]^{d_i}} \tag{3.6}$$

Fijémonos que en (3.6) para cada tiempo $t_{(i)}$ todos los denominadores son iguales, coincidiendo con el primero de (3.4).

3. **Efron**

   Efron sugiere una alternativa para la aproximación de Breslow que consiste en disminuir el denominador proporcionalmente:

$$\prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\prod_{k=1}^{d_i}\left[\sum_{l\in R(t_{(i)})}\exp(\boldsymbol{\beta}'\mathbf{Z}_l) - \frac{k-1}{d_i}\sum_{j\in D(t_{(i)})}\exp(\boldsymbol{\beta}'\mathbf{Z}_j)\right]} \tag{3.7}$$

   donde $D(t_{(i)}) = \{j;\ t_j = t_{(i)}\}$

   Simulaciones numéricas indican que cuando la fracción d_i/n_i de empates en cada tiempo es grande, (3.5) aún da buenas aproximaciones mientras que (3.6) muestra un sesgo. La aproximación de Efron (3.7) se comporta un poco mejor que (3.6).

4. **Average**

   Cox(1972) propuso tratar los empates utilizando la siguiente modificación de la verosimilitud parcial:

$$\prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}'\mathbf{S}_i)}{\sum_{l\in R_{d_i}(t_{(i)})}\exp(\boldsymbol{\beta}'\mathbf{S}_l)} \tag{3.8}$$

   donde $R_{d_i}(t_{(i)})$ son todos los subconjuntos de d_i individuos que se pueden hacer con los individuos en riesgo $R(t_{(i)})$.

3.3. Verosimilitud para tiempos discretos de supervivencia

Supongamos que nuestra muestra $(t_1, \delta_1, \mathbf{Z}_1), (t_2, \delta_2, \mathbf{Z}_2), \ldots (t_n, \delta_n, \mathbf{Z}_n)$ proviene de una distribución de tiempos discretos, donde como siempre las t_i son los tiempos y las δ_i los

indicadores de censura (1 si es un tiempo observado, 0 si es un tiempo censurado) y $\mathbf{Z}_i$ las covariables del individuo i.

La verosimilitud es proporcional a:

$$L \approx \prod_{i=1}^{n} \left[P(T = t_i|\mathbf{Z}_i)^{\delta_i} P(T > t_i|\mathbf{Z}_i)^{(1-\delta_i)} \right] \tag{3.9}$$

Sean $u_1 < u_2 < \dots u_k$ los valores que puede tomar la variable discreta T.

Recordemos las fórmulas que relacionan las probabilidades anteriores con las funciones de riesgo $h(t_i|\mathbf{Z}_i)$:

$$P(T = t_i|\mathbf{Z}_i) = h(t_i|\mathbf{Z}_i) \prod_{m;u_m<t_i} (1 - h(u_m|\mathbf{Z}_i)) \quad P(T > t_i|\mathbf{Z}_i) = \prod_{m;u_m \leq t_i} (1 - h(u_m|\mathbf{Z}_i))$$

Substituyendo en (3.9):

$$L \approx \prod_{i=1}^{n} \left[h(t_i|\mathbf{Z}_i) \prod_{m;u_m<t_i} (1 - h(u_m|\mathbf{Z}_i)) \right]^{\delta_i} \left[\prod_{m;u_m \leq t_i} (1 - h(u_m|\mathbf{Z}_i)) \right]^{(1-\delta_i)} \tag{3.10}$$

Definimos $r_{im} = \delta_i \, \mathbb{1}(u_m = t_i)$, así, $r_{im} = \delta_i$ si $u_m = t_i$ y cero en el caso contrario. Entonces:

$$L \approx \prod_{i=1}^{n} \prod_{m;u_m \leq t_i} \left(\frac{h(u_m|\mathbf{Z}_i)}{1 - h(u_m|\mathbf{Z}_i)} \right)^{r_{im}} (1-h(u_m|\mathbf{Z}_i)) = \prod_{i=1}^{n} \prod_{m;u_m \leq t_i} h(u_m|\mathbf{Z}_i)^{r_{im}} (1-h(u_m|\mathbf{Z}_i))^{1-r_{im}} \tag{3.11}$$

Que es la misma verosimilitud que para una muestra Bernoullis, ya que si $r_{im} = 1$ tenemos $h(u_m|\mathbf{Z}_i)$ y si $r_{im} = 0$ tenemos $1 - h(u_m|\mathbf{Z}_i)$.

3.4. Transformación del conjunto de datos a un conjunto individuo-periodo

A partir del resultado anterior la transformación de la base da datos queda justificada ya que la función de verosimilitud para el modelo de supervivencia a tiempo discreto es equivalente a la función de verosimilitud para un determinado modelo de respuesta binaria. Esta equivalencia es útil para usar un software estándar para la estimación del modelo lineal generalizado. No obstante, el uso de análisis de supervivencia con tiempos discretos requiere que el conjunto de datos originales se transforme a un conjunto de datos (ver Singer & Willett (2003)), que en este estudio se llamará conjunto de datos individuo-periodo. Esta transformación o conversión es el paso inicial para utilizar métodos estadísticos de regresión logística estándar. En esta transformación cada individuo tiene registros múltiples que dependerán del tiempo hasta el evento. Por lo tanto cada tiempo discreto (año, mes o semana) observado para cada individuo en cada periodo, le corresponde una variable respuesta binaria (o indicador binario) que indica la ocurrencia o no del evento objeto de estudio en cierto momento del tiempo. Es decir, el indicador asignará el valor 0 a cada observación hasta la fecha que ha sido observado y en el último registro asignará el valor de 1 si el individuo experimentó el evento y el valor 0 en el caso de ser una observación censurada. Esta transformación requiere de cuidados en su construcción con respecto a las censuras que se deben reflejar en la última observación de cada individuo.

Conjunto de datos estándar individuo-nivel

Cada individuo en el estudio tiene un registro en el conjunto de datos (ver Tabla 3.1). Este conjunto de datos es a menudo conocido como un conjunto de persona-nivel, pero en esta tesis se le llama individuo-nivel, y registra la siguiente información para cada i-ésimo individuo:

1. Duración (T): el tiempo observado.
2. Censura (C): Indica si el tiempo del individuo está censurado o no. La censura ocurre cuando el individuo al final del periodo en observación no ha experimentado

el evento de interés. El valor C_i es 0, si el individuo i no ha experimentado el evento en un determinado periodo de tiempo, y 1, si este ha experimentado el evento.

3. Covariables (Z): variables explicativas de interés.

Tabla 3.1: Conjunto de datos individuo-nivel

ID	T	C	Z
1	3	1	2
2	4	1	2
3	5	0	3

Conjunto de datos individuo-periodo

Antes de abordar los modelos discretos, se requiere la transformación del conjunto de datos individuo-nivel a un conjunto de datos individuo-periodo. Así, si los valores distintos observados del tiempo discreto son $u_1 < u_2 < \ldots u_J$; un individuo con tiempo t_i tendrá tantos registros como valores $u_k \leq t_i$ haya. Cada registro tendrá J indicadores de tiempo, todos cero excepto uno. El k-ésimo registro del individuo i, si $t_i \geq u_k$, tendrá el correspondiente indicador al tiempo u_k igual a uno y los demás cero. (para más detalle ver Singer & Willett (1993)). Así, este conjunto de datos individuo-periodo, tiene la siguiente información:

1. Los indicadores de tiempo $D_1, D_2, \ldots, D_J$, que son variables binarias de los valores u_k. Las variables toman valores 1 o 0, que identifican el periodo particular del tiempo registrado al que hace referencia. Así, D_{1k} será 1 en el k-ésimo registro del individuo i, siempre que $t_i \leq u_k$.

2. Las covariables $\mathbf{Z}_i$ del individuo i-ésimo. En todos los registros del individuo, las covariables $\mathbf{Z_i}$ son constantes.

3. La variable y_{ij}, que es el indicador de que el evento del individuo i ocurrió en el tiempo u_j:

$$y_{ij} = \begin{cases} 1 & \text{si } t_i = u_j \text{ y } \delta_i = 1; \\ 0 & \text{en caso contrario.} \end{cases}$$

Tabla 3.2: Conjunto de datos individuo-periodo, con sus indicadores de tiempo

ID	T	Y	D_1	D_2	D_3	D_4	D_5	Z
1	1	0	1	0	0	0	0	2
1	2	0	0	1	0	0	0	2
1	3	1	0	0	1	0	0	2
2	1	0	1	0	0	0	0	2
2	2	0	0	1	0	0	0	2
2	3	0	0	0	1	0	0	2
2	4	1	0	0	0	1	0	2
3	1	0	1	0	0	0	0	3
3	2	0	0	1	0	0	0	3
3	3	0	0	0	1	0	0	3
3	4	0	0	0	0	1	0	3
3	5	0	0	0	0	0	1	3

Para el conjunto de datos individuo-nivel de la Tabla 3.1, construimos el conjunto de datos individuo-periodo (Tabla 3.2). Así, el individuo $i = 3$ ha sido observado en cinco momentos de tiempo desde 1 a 5, por tanto tiene 5 registros. Además experimentó el evento de interés, luego $Y_{35} = 1$ y el valor de su covariable es $Z = 2$.

3.5. Dos modelos en tiempos discretos para análisis de supervivencia

3.5.1. Modelo Logit

Debido al hecho que la función de riesgo es una probabilidad, un enfoque de odds proporcionales, puede ser usado, tomando para cada tiempo la probabilidad correspondiente a la función de riesgo. Esto no ocurre en el caso donde el tiempo es continuo, ya que la función de riesgo no es una probabilidad. Además, tomando una transformación apropiada, permite el uso de un modelo de regresión logística, definido como:

$$\begin{aligned} \text{logit}(\text{h}\,(t/\mathbf{Z})) &= (\alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_J D_J) \\ &\quad + (\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p) \end{aligned} \tag{3.12}$$

donde $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)$ es el vector de las covariables y los parámetros $\beta_1, \beta_2, \ldots, \beta_p$ los coeficientes desconocidos que serán estimados. Las variables $D_1, \ldots, D_J$ son los indicadores de tiempo y α_j son los correspondientes parámetros, los cuales determinan la función de riesgo basal (para un individuo de referencia) del modelo logit. Adicionalmente, el signo y magnitud de estos coeficientes, describen la forma de la función de riesgo. No todos los individuos contribuyen en todos los tiempos. Por ejemplo, el individuo 1 de la tabla 3.2 no contribuye a los periodos 4 y 5.

La función de riesgo para este modelo, se puede obtener aplicando la transformació inversa apropiada de la ecuación (3.12):

$$h(t_j/\mathbf{Z}) = \frac{e^{\boldsymbol{\beta}'\mathbf{Z}+\alpha_j}}{1 + e^{\boldsymbol{\beta}'\mathbf{Z}+\alpha_j}}$$

3.5.2. Modelo clog-log

Este modelo responde a pensar que los tiempos provienen de una variable continua, U, donde sus valores han sido agrupados. Es decir, la semi recta $[0, \infty)$ está dividida en intervalos disjuntos $(t_{j-1}, t_j]$, $0 = t_0 < t_1 < \ldots$; cada intervalo esta representado por el valor t_j. La variable discreta T toma el valor t_j si $t_{j-1} < U \leq t_j$. Asumiendo que U sigue un modelo Cox, se tiene:

$$\begin{aligned} P(T=t_j|\mathbf{Z}) &= P(t_{j-1}<U\leq t_j|\mathbf{Z}) = S(t_{j-1}) - S(t_j) \\ &= \exp\left\{-\int_0^{t_{j-1}}\lambda(u|\mathbf{Z})du\right\} - \exp\left\{-\int_0^{t_j}\lambda(u|\mathbf{Z})du\right\} \\ &= \exp\left\{-\int_0^{t_{j-1}}\lambda_0(u)e^{\boldsymbol{\beta}'\mathbf{Z}}du\right\} - \exp\left\{-\int_0^{t_j}\lambda_0(u)e^{\boldsymbol{\beta}'\mathbf{Z}}du\right\} \\ &= \exp\left\{-\int_0^{t_{j-1}}\lambda_0(u)e^{\boldsymbol{\beta}'\mathbf{Z}}du\right\}\left(1-\exp\left\{-\int_{t_{j-1}}^{t_j}\lambda_0(u)e^{\boldsymbol{\beta}'\mathbf{Z}}du\right\}\right) \\ &= S(t_{j-1}|\mathbf{Z})\left(1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}}\int_{t_{j-1}}^{t_j}\lambda_0(u)du\right\}\right) \end{aligned} \tag{3.13}$$

donde $S(t)$ es la función de supervivencia de U; usando (3.13) la función de riesgo en el caso discreto es dada por:

$$\begin{aligned} h(t_j|\mathbf{Z}) &= P(T=t_j|T\geq t_j,\ \mathbf{Z}) = \frac{P(T=t_j|\mathbf{Z})}{P(T\geq t_j|\mathbf{Z})} \\ &= \frac{S(t_{j-1}|\mathbf{Z})\left(1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}}\int_{t_{j-1}}^{t_j}\lambda_0(u)du\right\}\right)}{S(t_{j-1}|\mathbf{Z})} \\ &= 1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}}\int_{t_{j-1}}^{t_j}\lambda_0(t)du\right\} = 1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}+\gamma_j}\right\} \end{aligned}$$

donde $\gamma_j = \log\int_{t_{j-1}}^{t_j}\lambda_0(u)du$.

El modelo que surge de forma natural en este contexto, fue presentado en Prentice & Gloeckler (1978), y se conoce como modelo clog-log, que es una versión del modelo de riesgos proporcionales cuando el tiempo T es una variable discreta y viene dado por:

$$\begin{aligned} \text{clog-log}(h\,(t/\mathbf{Z})) &= \log(-\log[1-h(t|\mathbf{Z})]) \\ &= (\gamma_1 D_1+\cdots+\gamma_J D_J)+(\beta_1 Z_1+\cdots+\beta_p Z_p) \end{aligned} \tag{3.14}$$

Notemos que el vector $\boldsymbol{\beta}$ es el mismo que en el modelo de Cox (PH); lo que permite una interpretación de sus estimaciones en término de cociente de riesgos. Como en el modelo logit, γ_j determina el valor de la función de riesgo basal.

A pesar de que el modelo clog-log surge a partir de un modelo de Cox, no es un modelo de riesgos proporcionales, ya que:

$$\frac{h(t_j \mid \mathbf{Z})}{h(t_j \mid \mathbf{Z}=0)} = \frac{1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}}\int_{t_{j-1}}^{t_j}\lambda_0(u)du\right\}}{1-\exp\left\{-\int_{t_{j-1}}^{t_j}\lambda_0(u)du\right\}}.$$

Sin embargo se observa que los logaritmos del complementario del riesgo, son proporcionales:

$$\frac{\log(1-h(t_j|\mathbf{Z}))}{\log(1-h(t_j|\mathbf{Z}=0))} = \frac{e^{\boldsymbol{\beta}'\mathbf{Z}}\int_{t_{j-1}}^{t_j}\lambda_0(u)du}{\int_{t_{j-1}}^{t_j}\lambda_0(u)du} = e^{\boldsymbol{\beta}'\mathbf{Z}}$$

Para valores pequeños de los riesgos, las funciones clog-log y logit están cercanas, de hecho ambas son cercanas al logaritmo del riesgo. Es de resaltar que el modelo clog-log no puede ser interpretado en términos de odds proporcionales del riesgo, como el modelo logit.

La función de riesgo para el modelo clog-log se puede obtener aplicando la transformació inversa apropiada de la ecuación (3.14):

$$h(t_j/\mathbf{Z}) = 1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}+\gamma_j}\right\}$$

Transformación clog-log para la función de distribución

Usando la misma idea anterior pero aplicada a la función de distribución (ver Prentice & Gloeckler (1978)) obtenemos un modelo diferente para tiempo discreto. Siguiendo la notación anterior se puede expresar la función de distribución de U como:

$$\begin{aligned} F(t_j|\mathbf{Z}) &= 1-S(t_j|\mathbf{Z}) = 1-P(T>t_j|\mathbf{Z}) \\ &= 1-P(U>t_j|\mathbf{Z}) = 1-S(t_j|\mathbf{Z}) \\ S(t_j|\mathbf{Z}) &= \exp\left\{-\int_0^{t_j}\lambda(u|\mathbf{Z})du\right\} \\ &= \exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}}\int_0^{t_j}\lambda_0(u)du\right\} \end{aligned}$$

Por lo tanto, usando la transformación clog-log, esto puede ser escrito como:

$$\begin{aligned}\text{clog-log}(F+(t_j|\mathbf{Z})) &= \log\left[-\log\left(1-F(t_j|\mathbf{Z})\right)\right] = \log\left[e^{\boldsymbol{\beta}\mathbf{Z}'}\int_0^{t_j}\lambda_0(u)du\right] \\ &= \boldsymbol{\beta}'\mathbf{Z}+\log\int_0^{t_j}\lambda_0(u)du,\end{aligned}$$

$$\text{clog-log}(F(t_j|\mathbf{Z})) = \log\left[-\log(S(t_j|\mathbf{Z}))\right] = \boldsymbol{\beta}'\mathbf{Z}+\log\int_0^{t_j}\lambda_0(u)du = \boldsymbol{\beta}'\mathbf{Z}+\gamma_j',$$

donde ahora $\gamma_j' = \log\int_0^{t_j}\lambda_0(u)du$.

La función de distribución y la de la supervivencia se pueden obtener por medio de la transformación apropiada:

$$F(t|\mathbf{Z}) = 1-\exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}+\gamma_j'}\right\}, \qquad S(t|\mathbf{Z}) = \exp\left\{-e^{\boldsymbol{\beta}'\mathbf{Z}+\gamma_j'}\right\}.$$

La estimación de los parámetros se realiza utilizando la misma técnica que para los modelos con variable respuesta binaria.

Relación entre los dos modelos de clog-log

Los dos modelo de clog-log están relacionados. En concreto:

1. para los parámetros γ_j y γ_j':

$$e^{\gamma_j'} = \int_0^{t_j}\lambda_0(u)du = \sum_{k=1}^{j}\int_{t_{k-1}}^{t_k}\lambda_0(u)du = \sum_{k=1}^{j}e^{\gamma_k}. \tag{3.15}$$

2. Los coeficientes β son los mismos en ambos modelos, pero sus estimaciones puede diferir.

3. Ninguno de estos modelos cumple la hipótesis de riesgos proporcionales.

Entre estos dos modelos, el modelo clog-log para la función de riesgos es el más utilizado. Principalmente nos interesa por su conexión con el modelo de Cox. En consecuencia,

en esta tesis, nos hemos centrado en este modelo, para el cual las diferentes técnicas estadísticas para la estimación de los parámetros y la interpretación de los mismos se han desarrollado.

3.6. Comparación de modelos

Para comparar distintos modelos utilizamos los estadísticos deviance, el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). Como el estadístico de deviance, el AIC y el BIC están basados en el logaritmo de la verosimilitud, pero con penalización. Así, en AIC la penalización se basa en el número de parámetros estimados, mientras que BIC va más allá, ya que su penalización se basa no sólo en el número de parámetros, sino también en el tamaño de la muestra, aunque en caso de datos de supervivencia se recomienda utilizar el número de eventos observados (Raftery, 1995).

3.7. Software estadísticos

Diferentes procedimientos o funciones de los software estadísticos R, Stata y SAS, fueron utilizadas en esta tesis, para obtener los resultados correspondientes a los estudiados realizados tanto para el modelo de Cox como para los modelos con tiempos discretos. En el capítulo 7 se hace una comparación de estos software.

Software R

coxph Esta función ajusta modelos de Cox y permite el tratamientos de empates, con la opción `ties` los métodos: Efron, Breslow y Exact. Mientras el método Average no está disponible. Por defecto utiliza la aproximación de Efron.

glm Esta función a partir de la matriz de datos expandida y especificando la opción `link=` logit o clog-log, y `family=` binomial, proporciona las estimaciones de los parámetros, así como los estadísticos de bondad de ajuste de los modelos para variable de respuesta binaria.

Software Stata

stcox Esta función permite ajustar un modelo de Cox, e incluye cuatro métodos para abordar empates: Efron, Breslow, Exact (`exactp`) y Average (`exactm`).

logit /cloglog Estas funciones sirven para ajustar modelos de respuesta binaria a partir de la base de datos expandida. Se obtiene las estimaciones de los parámetros y los correspondientes estadísticos de bondad de ajuste.

Software SAS

phreg Este procedimiento permite ajustar un modelo de Cox, utilizando los cuatro métodos para abordar empates (opción `ties`). El método Exact corresponde a la opción (`discrete`) y el método Average corresponde a la opción (`exact`). Por defecto utiliza la aproximación de Breslow.

glimmix Este procedimiento sirve para ajustar modelos lineales generalizados, por lo que se utilizará para obtener los resultados de los modelos logit y clog-log a partir de la matriz expandida. Las opciones especificas son `distrib=binary` y `link` igual al correspondiente enlace.

Capítulo 4

FRAILTY EN TIEMPOS DISCRETOS DE ANÁLISIS DE SUPERVIVENCIA

Por lo general, en el análisis de la supervivencia, se asume que la población es heterogénea, donde cada individuo del estudio está sujeto a un riesgo distinto dependiendo del valor de sus covariables. No siempre es posible recoger todas las covariables relacionadas con el tiempo hasta el evento, tanto sea por razones económicas como porque se desconoce su influencia. Otras veces los individuos pueden encontrarse formando grupos que comparten unas mismas características (zona geográfica, instituciones de salud o educativas, granjas, países, etc.). Añadir un efecto aleatorio en el modelo es una forma de tener en cuenta la heterogeneidad no observada causada por las covariables no medidas.

Esta heterogeneidad no observada, en análisis de supervivencia es conocida con el término frailty, introducido por Vaupel et al. (1979) en el contexto de tasas de mortalidad. Previamente (Clayton, 1978) introdujo el mismo concepto sin utilizar la notación frailty para el estudio de la incidencia de las enfermedades crónicas en la misma familia. Paralelamente Lancaster (1979) introdujo, en la literatura econométrica, el mismo concepto para tener en cuenta el efecto de las covariables omitidas en un modelo. En el campo de la bioestadística y la demografía, el término frailty usualmente es interpretado para un determinado individuo como un efecto aleatorio constante a lo largo del tiempo (Aalen

et al., 2008).

Algunos estudios han evaluado las consecuencias de omitir la heterogeneidad no observada (frailty). En donde ignorar este término puede causar sesgo en la estimación de los coeficientes de las variables explicativas en el modelo de riesgos, así como una subestimación de la función de riesgo (ver Nicoletti & Rondinelli (2010); Popkowski & Bass (1998); Scheike & Jensen (1997)).
En recientes décadas, una gran cantidad de estudios sobre modelos de frailty se han desarrollado en diferentes áreas de la ciencia. Algunos estudios han utilizado el modelo de Cox con término de frailty (Ha et al., 2012), mientras otros estudios han abordado la heterogeneidad no observada utilizando la regresión logística con factores aleatorios (Rabe-Hesketh et al., 2001). También, el enfoque bayesiano ha proporcionado otra vía para tomar en cuenta la heterogeneidad no observada y estimar los coeficientes de los parámetros en los modelos de frailty multivariante Clayton (1991); Korsgaard et al. (1998).

Este estudio no pretende abarcar todas las extensiones de los modelos de frailty, sólo se consideran modelos de análisis de supervivencia para tiempos discretos considerando uno y dos términos de frailty. El capítulo esta estructurado en las siguientes secciones: la sección 4.1 describe el modelo de Cox con efecto aleatorio; la sección 4.2 corresponde a Modelos de análisis de supervivencia en tiempos discretos con frailty; también la sección 4.3 muestra las distribuciones del término de frailty; finalmente la sección 4.4 presenta lo referido a Software estadísticos.

4.1. Modelo de Cox con efectos aleatorios

Cuando hay una característica compartida por un grupo de individuos, para tener en cuenta su efecto, se puede introducir en el modelo mediante uno o más términos de frailty. Este término aleatorio se asume con media cero y varianza desconocida, la cual debe ser estimada a partir de los datos.

4.1.1. Modelo de Cox con un término de frailty

El efecto de heterogeneidad no observada puede ser introducida en el modelo de Cox mediante un término multiplicativo en la función de riesgo que contiene tres componentes: un término frailty (efecto aleatorio), una función de riesgo basal y un término que modela la influencia de las covariables observadas.

Sea K grupos de individuos con n_k individuos por grupo, $k = 1, \ldots, K$. Entonces la función de riesgo en el tiempo t para el k-ésimo grupo dadas las covariables $\mathbf{Z}$, se expresa de la forma siguiente:

$$\lambda(t|\mathbf{Z}, \nu_k) = \lambda_0(t) \exp(\beta' \mathbf{Z} + \nu_k) \tag{4.1}$$

donde $\lambda_0(t)$ es la función de riesgo basal no especificada, y β es el vector de parámetros de regresión para las covariables $\mathbf{Z}$ y ν es el término de frailty.

La introducción del término de frailty en el modelo de Cox puede conducir a tener riesgos que no sean proporcionales. Esto significa que este modelo puede ser una alternativa para analizar tiempos con riesgos no proporcionales. Un modelo con un término de frailty solo puede inducir asociación positiva dentro del cluster o grupo. Sin embargo, hay algunas situaciones en las que los tiempos de supervivencia para los individuos dentro del mismo grupo se asocian negativamente, con lo cual estos modelos no serían adecuados.

4.1.2. Modelo de Cox con dos o más terminos de frailty

En algunas situaciones, la estructura del conjunto de los datos es más complicada, por lo cuál hay la necesidad de tomar en cuenta dependencias dentro de los grupos, siendo necesario más de un término de frailty. Así, estudios previos (Ha et al., 2007; Ripatti & Palmgren, 2000; Yau & McGilchrist, 1997) han considerado modelos con más de un término de frailty.

Al igual que el modelo frailty univariado, en este caso la función de riesgo contiene tres componentes: una función de riesgo basal, un término que modela la influencia de las

covariables observadas y dos (o más) términos frailty:

$$\lambda(t|\mathbf{Z}, \nu_j, \mu_k) = \lambda_0(t) \exp(\beta'\mathbf{Z} + \nu_j + \mu_k) \tag{4.2}$$

donde λ_0 es una función de riesgo basal no especificada β es un vector de parámetros de regresión para covariables de efectos fijos $\mathbf{Z}$ y ν_j y μ_k son los términos de frailty que se asumen independientes.

4.2. Modelos de análisis de supervivencia en tiempos discretos con frailty

Como fué explicado en el capítulo 3, la verosimilitud cuando los tiempos son discretos es equivalente a la verosimilitud para una muestra de una variable respuesta binaria, basada en el conjunto de datos expandidos. Se puede hacer un argumento similar cuando se considera un término de frailty.

4.2.1. Verosimilitud para datos discretos con un término de frailty

Supongamos que tenemos K grupos. Sea n_k el número de individuos del k-ésimo grupo, $k = 1, \ldots, K$. Entonces la muestra puede escribirse: $\{(t_{ik}, \delta_{ik}, \mathbf{Z}_{ik}), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K\}$ donde t_{ik} es el tiempo del individuo i del grupo k, δ_{ik} el indicador de censura y $\mathbf{Z}_{ik}$ sus covariables.

La verosimilitud es proporcional a

$$L \approx \prod_{k=1}^{K} \prod_{i=1}^{n_k} \left[P(T = t_{ik}|\mathbf{Z}_{ik})^{\delta_{ik}} P(T > t_{ik}|\mathbf{Z}_{ik})^{(1-\delta_{ik})} \right]. \tag{4.3}$$

Siguiendo el mismo argumento que en el capítulo 3, utilizando los tiempos diferentes de

fallo $u_1 < u_2 < \ldots < u_r$, podemos reescribir la ecuación (4.3) utilizando la función de riesgo condicionada a las covariables $\mathbf{Z}_{ik}$ y al efecto aleatorio ν_k:

$$L \approx \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left\{ \prod_{i=1}^{n_k} \left[h(t_{ik}|\mathbf{Z}_{ik}, \nu_k)^{\delta_{ik}} (1 - h(t_{ik}|\mathbf{Z}_{ik}, \nu_k))^{(1-\delta_{ik})} \prod_{m;u_m<t_{ik}} (1 - h(u_m|\mathbf{Z}_{ik}, \nu_k) \right] \right\} dF_{\nu_k}.$$

Utilizando la variable $r_{ikm} = \delta_{ik} \mathbb{1}(u_m = t_{ik})$, obtenemos:

$$L \approx \prod_{k=1}^{K} \int_{-\infty}^{+\infty} \left\{ \prod_{i=1}^{n_k} \prod_{m;u_m \leq t_{ik}} \left[h(u_m|\mathbf{Z}_{ik}, \nu_k)^{r_{ikm}} (1 - h(u_m|\mathbf{Z}_{ik}, \nu_k))^{(1-r_{ikm})} \right] \right\} dF_{\nu_k}$$

lo que de nuevo es la verosimilitud de una muestra de Bernouillis independientes con probabilidad de éxito condicionada a $\mathbf{Z}$ y ν_k igual a $h(t|\mathbf{Z}, \nu_k)$.

4.2.2. Modelos logit y clog-log con un término de frailty

En orden de introducir la heterogeneidad no observada, se adiciona el término de frailty, a los modelos logit y clog-log considerados en el capítulo 3, como se indica en las siguientes ecuaciones:

$$\text{logit}(h\,(t|\mathbf{Z}, \nu_k)) = (\alpha_1 D_1 + \cdots + \alpha_J D_J) + (\beta_1 Z_1 + \cdots + \beta_p Z_p) + \nu_k$$

$$\text{clog-log}(h\,(t|\mathbf{Z}, \nu_k)) = (\gamma_1 D_1 + \cdots + \gamma_J D_J) + (\beta_1 Z_1 + \cdots + \beta_p Z_p) + \nu_k \tag{4.4}$$

donde ν_k corresponde al efecto aleatorio (término de frailty).

De igual forma se definirían los modelos logit y clog-log con dos o más términos de frailty.

4.3. Distribuciones del término de frailty

Los modelos de frailty dependerán de la selección de la distribución que se considere para el término frailty o su exponenciación. Una revisión detallada de las diferentes distribuciones utilizadas se puede encontrar en Duchateau & Janssen (2008).

Dos distribuciones para la exponencial del término de frailty han sido ampliamente utilizadas: la distribución gamma y la distribución log-normal. Inicialmente el modelo con distribución gamma fue el más popular debido a su conveniencia matemática ya que conduce a una forma analítica cerrada de la función de verosimilitud (Clayton (1978); Meyer (1990); Vaupel et al. (1979); Wienke et al. (2003)).

El modelo de frailty con distribución log-normal es muy relevante por su fuerte vínculo con modelos lineales mixtos. Sin embargo la estimación del modelo log-normal es complicada por el hecho que no permite una expresión analítica para la función de verosimilitud por lo que se requieren metodologías de integración numérica. De todas formas, un estudio reciente de Ha et al. (2012), adoptó tanto la distribución gamma como la log-normal para el término de frailty.

4.3.1. Estimación de parámetros

Cuando se consideran el modelo de Cox con frailty los procedimientos tradicionales para maximizar la verosimilitud no son apropiados. Si se considera el término de frailty como una covariable no observada seria apropiado utilizar el algoritmo EM (ver Guo & Rodriguez (1992); Klein (1992)). No obstante este algoritmo es lento y el cálculo de las varianzas puede comportar computaciones intensivas y no están implementadas en la mayoría de software disponibles. Sin embargo, se pueden utilizar en su lugar las metodologías para modelo de Cox penalizado (Therneau & Grambsch, 2000) que permiten también interpretar el término de frailty como un efecto aleatorio. Otras metodologías han sido propuestas por (Ha et al., 2012) que utilizan la verosimilitud jerárquica para la estimación de los parámetros. Para el caso de los modelos logit y clog-log se utilizan las metodologías de maximización para los modelos con respuesta binaria implementados en los software estándar.

Cuando el término de frailty proviene de una estructura jerárquica de los datos, es de interés el coeficiente de correlación intraclase (ρ). Para los modelos que asumen que los efectos aleatorios están distribuidos normalmente, la varianza estimada del término de frailty se puede relacionar como la correlación intraclase (Hedeker et al., 2000), que es la proporción de la varianza total atribuible al grupo (Paccagnella, 2006):

$$\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma^2} \tag{4.5}$$

.

donde σ_ν^2 representa la varianza del término aleatorio y σ^2 es un término de varianza igual a $\pi^2/3$, si la función link es la función logit; y $\sigma^2 = \pi^2/6$ si la función link es la función clog-log (ver Rabe-Hesketh & Skrondal (2008), Paccagnella (2006)).

4.4. Software estadísticos

En esta sección se presentan los procedimientos o funciones de los software estadísticos R, Stata y SAS, utilizadas para modelos con término de frailty; una comparación de los mismos se encuentra en el capitulo 7.

Software R

coxph esta función permite seleccionar la distribucción de la exponencial del término frailty como log-normal, gamma o t de student con la opción `dist=` " ". El método de tratamiento de empates utilizado cuando se incluye un término de frailty es siempre Breslow. La desventaja de esta función es que no proporciona el error estándar de la estimación de la varianza para el término de frailty.

También puede especificarse la variable `cluster` cuando el término frailty representa el efecto del grupo.

coxme En esta función ajusta un modelo de Cox con efectos aleatorios normales. Tiene dos opciones para los empates Efron y Breslow. La sintaxis es estándar para los efectos aleatorios.

frailtyHL Esta función forma parte del paquete del mismo nombre *frailtyHL* (Ha et al., 2012). Sirve para ajustar modelos de Cox con efectos aleatorios. Considera tanto la distribución normal como la log-gamma para el efecto aleatorio. Esta función proporciona estimadores basándose en la verosimilitud jerárquica (HL) y suministra los errores estándar de la varianza del término frailty. El unico método para tratamiento de empates es Breslow.

glmer Esta función pertenece al paquete *lme4*. Ajusta modelos lineales generalizados con efectos aleatorios. Al igual que para las funciones `coxph` y `coxme`, no suministra valores del error estandár de la varianza.

Software Stata

xtlogit / xtcloglog Estas funciones requieren de la matriz de datos expandida. Asumen que el término de frailty está distribuido normalmente. Reportan la desviaciones estándar del efecto aleatorio y la correlación intraclase, para ambos `link` logit y clog-log.

gllamm Esta función sirve para ajustar modelos multinivel y permite añadir efectos aleatorios en la pendiente y en el término independiente. Reporta la desviación estándar y no la varianza del término aleatorio.

Software SAS

phreg Este procedimiento permite adicionar un término frailty con los diferentes tratamientos de empates.

glimmix Este procedimiento permite ajustar modelos lineales generalizados con términos aleatorios. Una variedad de métodos de estimación basados en técnicas de pseudo-verosimilitud se encuentran disponibles en ésta función. En el caso de modelos de supervivencia con datos discretos, se puede utilizar para ajustar los modelos de variable respuesta binaria con enlaces logit y clog-log utilizando la matriz de datos extendida. En este caso el término de frailty se incorpora en la opción random de forma habitual.

Capítulo 5

APLICACIÓN DE MÉTODOS DE ANÁLISIS DE SUPERVIVENCIA CON TIEMPOS DISCRETOS

5.1. Introducción

El principal objetivo de este capítulo es aplicar los métodos de análisis de supervivencia para abordar modelos en tiempos discretos a un conjunto de datos reales relacionados a lesiones musculoesqueléticas de caballos de carreras pura sangre.
Esta aplicación ha permitido, por una parte analizar tiempos discretos con los diferentes enfoques utilizados y por otra parte evaluar los factores asociados al riesgo de presentarse el evento de interés teniendo en cuenta que la variable tiempo es discreta. Así, se han considerado específicamente el modelo de Cox con tratamientos para empates (Efron, Breslow, Exact y Average) y modelos para respuesta binaria logit y clog-log (descritos en el capítulo 3). Sin embargo es importante recordar que el modelo clog-log no es un modelo de riesgos proporcionales aunque las estimaciones de los parámetros pueden interpretarse como cocientes de riesgos y que el modelo logit es de odds proporcionales. Estos modelos pueden ser ajustados a partir de una expansión del conjunto de datos originales y la utilización de métodos estándar para respuestas dicotómicas.

En el presente estudio las covariables con un nivel de significación estadística del 1 %, 5 % y 10 % fueron consideradas en los modelos estudiados, así tambien se incluyeron aquellas covariables que presentaron interés clínico o anatomopatológico. En la mayoría de los análisis realizados en este capítulo se utilizó el software R versión 2.15.1, adicionalmente el Stata fué utilizado para obtener el modelo de Cox con tratamiento de empate Average, ya que el software R no lo contempla, el método Average también está disponible en el software SAS.
El capítulo esta estructurado en las siguientes secciones: la intoducción presentada en la sección 5.1; la metodología empleada en este capítulo se muestra en la sección 5.2; la presentación del conjunto de datos es incluida en la sección 5.3; con su respectiva estadística descriptiva en la sección 5.4; también la conversión del conjunto de datos es mostrada en la sección 5.5; las diferentes formas de considerar la base de datos para su estudio es mostrada en la sección 5.6, y finalmente una sección 5.7 que corresponde con los resultados de la aplicación de los métodos estadísticos para abordar tiempos discretos en análisis de supervivencia.

5.2. Metodología

A continuación se presenta un resumen de dos enfoques de análisis de supervivencia, para abordar los tiempos discretos cuando se requiere analizar e identificar factores asociados al evento de interés, los cuales han sido descritos y discutidos previamente en el capítulo 3.

Modelo de Cox con técnicas para tratamiento de empates

Como se discutió en el capítulo 2, las técnicas de análisis de supervivencia se aplican en investigaciones donde la variable de interés es el tiempo hasta que ocurra el evento de interés. Cuando se quieren evaluar el efecto de un conjunto de variables explicativas y el tiempo es continuo, se puede utilizar el modelo de Cox, donde la función de riesgo se asume como:

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}}$$

donde $\mathbf{Z}' = (Z_1, \ldots, Z_p)$ es el vector de covariables y $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ el vector de los coeficientes (lo cual implica que en cada momento del tiempo la función de riesgo cambia proporcionalmente cuando cambian las covariables $\mathbf{Z}$).

Cuando el tiempo es discreto se puede utilizar también el modelo de Cox si para la estimación de los parámetros se utilizan técnicas de tratamientos de empates. Tal como se ha explicado en el capítulo 3 hay cuatros diferentes enfoques o tratamientos de empates: Efron, Breslow, Exact, Average.

Modelos de Análisis de Supervivencia en tiempo discreto

En los modelos logit y clog-log, la función de riesgo se modeliza de la forma:

$$\begin{aligned}
\text{logit}\left(h(t|\mathbf{Z})\right) &= \log\left(\tfrac{h(t|\mathbf{Z})}{1-h(t|\mathbf{Z})}\right) \\
&= \alpha_1 D_1 + \cdots + \alpha_J D_J + \beta_1 Z_1 \cdots + \beta_p Z_p \\
\text{clog-log}\left(h(t_i|\mathbf{Z})\right) &= \log\left(-\log[1 - h(t_i|\mathbf{Z})]\right) \\
&= \gamma_1 D_1 + \cdots + \gamma_J D_J + \beta_1 Z_1 \cdots + \beta_p Z_p
\end{aligned}$$

Para la estimación de los parámetros se utilizan técnicas de los modelos para una respuesta binaria. Para ello el conjunto de datos debe ser convertido o transformado a un conjunto de datos ampliados, como requisito necesario para utilizar software estadístico para regresión logística.

5.3. Presentación de la base de datos: caballos de carrera de la raza pura sangre con riesgo de experimentar una CMI

Un total de 214 caballos de carreras de la raza pura sangre proveniente del hipódromo *La Rinconada*, localizado en la región norte de Venezuela fué analizada en este estudio. El conjunto de datos considerados, corresponde a los caballos que resultaron lesionados durante la carrera en que estaban compitiendo. Estos fueron registrados entre Enero del año 2000 hasta el final de seguimiento en Mayo del 2011.
En el hipódromo *La Rinconada* el examen clínico se realiza antes del inicio de la carrera para evaluar la condición física del caballo. Posterior a la carrera se realiza otro examen físico a los caballos que llegan lesionados al hospital del hipódromo, donde se diagnostica la lesión sufrida por el caballo y de acuerdo a la severidad de la lesión y al compromiso vascular, se procede a decidir entre intervenir quirúrgicamente al caballo o aplicar eutanasia. Por su parte la eutanasia se decide cuando el daño del hueso es irreversible, la lesión presenta un compromiso vascular severo con interrupción en el riego sanguíneo al resto del miembro, lo cual conlleva a hemorragia en la lesión y una isquemia del miembro, y se considera que el caballo no tendrá posibilidades de recuperarse con la intervención quirúrgica. En cuyo caso, se diagnostica que el caballo sufrió una lesión musculoesquelética catastrófica (CMI). En el hipódromo *La Rinconada*, los caballos a los 2 años de edad son considerados aptos para iniciar las competiciones.

5.4. Estadística descriptiva

5.4.1. Eventos y observaciones censuradas

De los 214 caballos de carreras, 112 caballos (que representan 52.3 %) experimentaron el evento de intéres, es decir CMI y 102 (47.7 %) caballos de carreras no presentaron CMI y sobrevivieron a este evento, por lo tanto fueron considerados datos censurados. El tiempo hasta que ocurrió el evento de interés (CMI) se midió en años, con una mediana estimada de 4 años de edad (IC=[3;4]).

Tabla 5.1: Eventos y observaciones censuradas.

	Tiempo (edad en años)					
	2	3	4	5	6	7
eventos	42	27	15	15	7	6
censuras	47	22	19	9	2	3

5.4.2. Descripción de las covariables

Las variables disponibles en el conjunto de la base de datos pueden clasificarse como:
(1) características propias del caballo, dadas por género (masculino y femenino); peso físico (kg.); miembros (anteriores y posteriores); lesiones preexistentes; niveles de fósforo; leucocitos y niveles de calcio sérico (en dos categorías);
(2) características de la carrera, dadas por el número de carreras en las que el caballo ha participado; longitud o distancia de la carrera donde participaba en el momento de la lesión, que pueden dividirse en carreras cortas (800-1200 m.), medianas (1300-1600 m.) y largas (1800-2400 m.); la distancia proporcional al accidente, definida como la distancia desde el sitio del accidente hasta el final de la carrera con respecto al total de la longitud de la carrera en la cual participó; y la época clasificada en seca y lluviosa, debido al clima tropical que tiene Venezuela.
En la Tabla 5.2 se presenta la estadística descriptiva y los niveles de cada una de las covariables que fueron registradas durante el estudio, sin embargo, no todas fueron incluídas en los análisis.

5.4.3. Distribución anatómica de las lesiones musculoesqueléticas

De acuerdo al sitio anatómico donde fue diagnosticada la lesión musculoesquelética se registraron diferentes sitios anatómicos, siendo los más frecuentes: hueso sesamoide proximal (PSB) y el hueso metacarpiano (MIII), hueso del carpo, hueso radio-cubito; otros

Tabla 5.2: Estadística descriptiva para las covariables en caballos de carrera con CMI en el hipódromo *La Rinconada*

Covariables	*media*	*D.E.*	*mediana*	*min*	*max*
Carreras	8.82	7.72	7.50	1.00	37.00
Dist.Accid.	0.33	0.32	0.22	0.03	1.00
Fósforo (mg/dl)	10.24	1.16	10	9.00	12.60
Leucocitos (10^6 UI)	8.60	0.29	8.6	8.10	10.00
Peso Fisco (kg)	437.00	40.78	423	350.00	540.00
Covariables	*n*	*%*			
Lesión preexistente					
Normal (0)	160	74.8			
Lesionado (1)	54	25.2			
Longitud de Carrera					
Corta (1)	111	51.9			
Mediana (2)	73	34.1			
Larga (3)	30	14.0			
Época					
Lluviosa (1)	127	59.3			
Seca (2)	87	40.7			
Género					
Masculino (1)	91	42.5			
Femenino (2)	123	57.5			
Calcio					
Normal (0)	164	76.6			
Elevado (1)	50	23.4			

sitios anatómicos con menor frecuencia fueron: tibia, pelvis, metatarso, vertebras y costillas. Los miembros anteriores presentaron mayor proporción (86.2 %) de lesiones que los miembros posteriores, donde el miembro anterior derecho mostró una mayor proporción de lesiones (51 %) que el miembro anterior izquierdo (35 %). Los sitios de las lesiones musculoesqueléticas más comunes resultaron ser el PSB (64.3 %), y el MIII (24.1 %).

5.5. Conversión del conjunto de datos

Conjunto de datos estándar individuo-nivel

En el conjunto de datos original, como se muestra en la Tabla 5.3, cada caballo (individuo) tiene un único registro. El conjunto de la base de datos es a menudo conocido como un conjunto de datos referido a personas, sin embargo para efecto del desarrollo de esta tesis, donde las unidades de estudio son animales, le llamaremos conjunto individuo-nivel. Así, para cada caballo de carrera se registra la siguiente información:

1. Variable tiempo (T): la edad (en años) del caballo cuando le ocurrió la lesión musculoesquelética. Puede observarse que los caballos inician su participación en las carreras a los dos años de edad.
2. Indicador de censura (C): variable binaria, con valor igual a 1 si el caballo ha sufrido una lesión musculoesquelética catastrófica (CMI) y 0 si la lesión no ha sido catastrófica.
3. Covariables (Z): conjunto de variables explicativas de interés.

Tabla 5.3: Tres individuos del conjunto de datos individuo-nivel, de nuestra base de datos con la covariable época.

ID	*T*	*Censura*	*Época*
8	4	1	2
132	5	1	2
190	5	0	1

Nuevo conjunto de datos individuo-periodo

Antes de ajustar los modelos en tiempo discreto por medio de los modelos lineales generalizados (GLM), se requiere la transformación del conjunto de datos estándar individuo-nivel, a una nueva base de datos llamada conjunto de datos individuo-periodo. De esta forma cada caballo (individuo) tendrá múltiples registros en el nuevo conjunto de datos, con una estructura como se muestra en la Tabla 5.4, es decir, tendrá un registro (fila)

por cada periodo de tiempo observado hasta que ocurra el evento de interés o quede censurado.

Tabla 5.4: Conjunto de datos Individuo-Periodo con indicadores de tiempo y la covariable época, para los tres individuos de la Tabla 5.3

ID	T	Y	D_2	D_3	D_4	D_5	D_6	D_7	*Época*
8	2	0	1	0	0	0	0	0	2
8	3	0	0	1	0	0	0	0	2
8	4	1	0	0	1	0	0	0	2
132	2	0	1	0	0	0	0	0	2
132	3	0	0	1	0	0	0	0	2
132	4	0	0	0	1	0	0	0	2
132	5	1	0	0	0	1	0	0	2
190	2	0	1	0	0	0	0	0	1
190	3	0	0	1	0	0	0	0	1
190	4	0	0	0	1	0	0	0	1
190	5	0	0	0	0	1	0	0	1

De esta manera, se obtuvo un conjunto de datos individuo-periodo, que tiene la siguiente información:

1. Los indicadores de tiempo $D_2, \cdots, D_7$ son una secuencia de variables binarias (o dummy) de los tiempos registrados. Es importante recordar que en la práctica es usual encontrar que el primer indicador de tiempo es D_1, pero en ésta base de datos, debido a que el caballo de carreras inicia las competiciones a los 2 años de edad, se utilizó como primer indicador D_2. Por lo tanto, estos indicadores de tiempo son usados para representar cada periodo de tiempo, desde los 2 años de edad a los 7 años de edad, momento en el cual el caballo deja de participar en las carreras del hipódromo *La Rinconada* para ser destinado a un centro de recría o a otro hipódromo de menor categoría.

2. Las Covariables. En todo los registros del i-ésimo caballo, el vector de covariables Z es constante, ya que no se trata de covariables cambiantes en el tiempo.

3. La variable Y. Para un caballo i, Y_{ij} es el indicador de que el evento ocurrió en el tiempo j:

$$Y_{ij} = \begin{cases} 1 & \text{cuando el caballo } i \text{ experimentó el evento en el tiempo } j; \\ 0 & \text{en caso contrario.} \end{cases}$$

Por ejemplo, en la Tabla 5.4 del conjunto de datos individuo-periodo, el caballo de carrera $i = 132$ fue observado por 4 periodos de tiempo consecutivos, desde el 2 hasta el 5. Este mismo caballo experimentó el evento de interés (CMI) y el valor de la covariable registrada fue época.

Los resultados mostrados a continuación muestran las cuatro formas consideradas en este estudio de involucrar los caballos de carreras, como fue explicado en la sección 5.6.

5.6. Bases de datos consideradas para este estudio

El propósito de esta subsección es explicar el interés que se tiene de analizar la base de datos en cuatro subconjuntos. En la práctica se tiene interés en identificar factores de riesgo asociados a la edad en ocurre una CMI en caballos de carrera pura sangre que compiten en carreras planas en el Hipódromo *La Rinconada* en Venezuela.
Primero se ha analizado toda la base de datos de caballos, donde se ha evaluado el ajuste de diferentes modelos, el interés se centra en conocer los factores de riesgo en presentarse una CMI, considerando el total de la población que resultó afectada durante el periodo de tiempo 2000 a 2011, ya que en Venezuela no se tienen estudios anteriores referentes a este tema.
Luego, se consideró la lesión a nivel de miembros; el interés se basa en hallazgos encontrados en otros estudios. La mayor proporción de las fatalidades en carreras planas están referidas a CMI en miembros, como han sido reportadas en Reino Unido (74 %) por Boden et al. (2006) y en Norte América (89 %) por Johnson et al. (1994). Finalmente, se tiene interés en analizar los factores de riesgos que afectan diferentes sitios anátomicos, pues se conoce que las lesiones musculoesqueléticas en huesos sesamoide proximal son las mas comunes en Estados Unidos (Johnson et al., 1994; Mohammed et al., 1991; Peloso

et al., 1994), y por otra parte las CMI en los huesos metacarpo y carpo son las más comunes en el Reino Unido en los diferentes tipos de carreras (Parkin et al., 2004).

5.7. Resultados

5.7.1. Considerando toda la base de datos

Esta base de datos, contiene 214 caballos, que sufrieron una lesión musculoesquelética dentro del periodo 2000-2011. De estos 214 caballos, 112 caballos, que representan 52.3 % experimentaron el evento de intéres, es decir CMI y 102 (47.7 %) caballos no presentaron CMI y sobrevivieron a este evento, por lo tanto fueron considerados datos censurados.

Tiempo considerado

Para el análisis de este conjunto de datos fué considerado como tiempo discreto el tiempo (en anõs) hasta que ocurre una CMI en los caballos de carrera. Esto es, la edad del caballo medida en años.

Modelo de Cox con tratamientos de empates

Como primer análisis, presentamos los resultados del modelo de Cox con los diferentes tratamientos de empates: Efron, Breslow, Exact y Average. En la Tabla 5.5 se muestran las covariables que resultaron estadísticamente significativas que fueron: la lesión preexistente, el número de carreras y la época.

Los valores mostrados para e^{β} en los modelos con el método Efron y Average indican que el riesgo de CMI que presenta un caballo que fue previamente diagnosticado con lesión preexistente, es más del doble en comparación a un caballo que no presentó lesión preexistente, mientras que según el valor e^{β} del método Exact el riesgo es casi el triple y para el método Breslow no llega al doble.

En relación a la variable número de carreras, se puede apreciar que resultó estadísticamente significativa en los todos los métodos de tratamientos de empates. Así, se muestra que por cada carrera adicional que tenga un caballo, el riesgo de padecer una CMI dis-

minuirá.
La covariable época por su parte, resultó estadísticamente significativa al 10 % en los modelos con tratamiento de empates Efron, Exact y Average; esta situación no ocurrió con el método Breslow, donde no resultó estadísticamente significativa.
Sin embargo, la covariable longitud de carrera y la covariable género no resultaron estadísticamente significativas, lo que nos lleva a pensar que no son factores de riesgos en relación a que el caballo sufra una CMI.
Así, una diferencia que se observa entre los modelos con diferentes métodos de tratamiento de empates, es en relación a la magnitud de los coeficientes. Otra situación se presenta al observar los errores estándar, donde se encontró que el método Exact es el que muestra los valores mayores en las estimaciones obtenidas.

Tabla 5.5: Parámetros estimados del Modelo de Cox con diferentes tratamiento de empates con toda la base de datos.

	Modelo Efron		*Modelo Breslow*		*Modelo Exact*		*Modelo Average*	
$LesionP_1$								
Coeficiente	0.772		0.589		1.026		0.866	
exp(coef)	2.165		1.803		2.791		2.376	
EE	0.221		0.214		0.289		0.234	
p-valor	<0.001	(***)	0.006	(**)	<0.001	(***)	<0.001	(***)
Carreras								
Coeficiente	-0.154		-0.133		-0.206		-0.165	
exp(coef)	0.857		0.875		0.814		0.848	
EE	0.021		0.020		0.029		0.023	
p-valor	<0.001	(***)	<0.001	(***)	<0.001	(***)	<0.001	(***)
$LongC_2$								
Coeficiente	-0.283		-0.218		-0.235		-0.277	
exp(coef)	0.754		0.804		0.791		0.758	
EE	0.227		0.225		0.285		0.236	
p-valor	0.210		0.330		0.410		0.240	
$LongC_3$								
Coeficiente	-0.405		-0.296		-0.396		-0.451	
exp(coef)	0.667		0.744		0.673		0.637	
EE	0.311		0.309		0.394		0.320	
p-valor	0.190		0.340		0.310		0.160	
$Epoca_2$								
Coeficiente	0.335		0.260		0.417		0.365	
exp(coef)	1.398		1.296		1.517		1.441	
EE	0.195		0.194		0.249		0.201	
p-valor	0.086	(·)	0.180		0.093	(·)	0.069	(·)
$Genero_2$								
Coeficiente	0.051		0.023		0.118		0.092	
exp(coef)	1.052		1.024		1.125		1.097	
EE	0.196		0.195		0.248		0.203	
p-valor	0.800		0.900		0.640		0.650	

Nota: (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

Modelos en tiempo discreto: logit y clog-log, con toda la base de datos

En la Tabla 5.6 se muestran las estimaciones de los coeficientes de los parámetros y los errores estándar para los indicadores de tiempo y las covariables consideradas, correspondientes a los modelos discretos logit y clog-log, utilizando la base de datos completa.

Tabla 5.6: Modelo logit y clog-log para el tiempo discreto, con la base de datos individuo-periodo.

	Modelo	**logit**		**Modelo**	**clog-log**	
IT	$\hat{\boldsymbol{\alpha}}$ **EE**	**p-valor**		$\hat{\boldsymbol{\alpha}}$ **EE**	**p-valor**	
d2	-0.512 (0.295)	0.083	(·)	-0.769 (0.239)	0.001	(**)
d3	0.620 (0.413)	0.133		0.108 (0.328)	0.742	
d4	1.113 (0.517)	0.031	(*)	0.458 (0.414)	0.268	
d5	2.985 (0.669)	<0.001	(***)	1.871 (0.507)	<0.001	(***)
d6	3.393 (0.816)	<0.001	(***)	2.225 (0.592)	<0.001	(***)
d7	6.118 (1.199)	<0.001	(***)	3.927 (0.802)	<0.001	(***)
Cov.	$\hat{\boldsymbol{\beta}}$ **EE**	**p-valor**		$\hat{\boldsymbol{\beta}}$ **EE**	**p-valor**	
$LesionP_1$	1.051 (0.291)	<0.001	(***)	0.862 (0.225)	<0.001	(***)
$Carreras$	-0.210 (0.030)	<0.001	(***)	-0.164 (0.023)	<0.001	(***)
$LongC_2$	-0.234 (0.288)	0.415		-0.285 (0.233)	0.222	
$LongC_3$	-0.402 (0.397)	0.312		-0.454 (0.319)	0.154	
$Epoca_2$	0.423 (0.251)	0.092	(·)	0.371 (0.199)	0.063	(·)
$Genero_2$	0.123 (0.250)	0.623		0.096 (0.201)	0.634	

Nota: $\hat{\boldsymbol{\alpha}}$= coeficientes de los indicadores de tiempo (**IT**); $\hat{\boldsymbol{\beta}}$ = coeficientes de las covariables (**Cov.**); (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

En la Tabla 5.6, se muestra que los modelos discretos logit y clog-log presentaron las mismas covariables significativas, sin embargo el modelo logit mostró valores mayores tanto en los coeficientes α como en los coeficientes β y en los errores estándar respectivos, pero en la misma dirección. Se puede ver también que el el riesgo aumenta con la edad del caballo. Es así, que en ambos modelos los coeficientes α de los indicadores

de tiempo presentan la misma dirección y resultaron estadísticamente significativos, a excepción del indicador de tiempo a los 2 años (en ambos modelos) y el de 4 años en el modelo clog-log.
En similitud con el modelo de Cox con los cuatro tratamientos de empates (Tabla 5.5), las variables lesión preexistente y número de carreras resultaron estadísticamente significativas al 5 %; de igual forma, la variable época resultó estadísticamente significativa al 10 %. Por otra parte, al comparar estos resultados, Tabla 5.6, con los modelos de la Tabla 5.5 , se puede observar que el modelo logit se asemeja al modelo Exact , en relación a los coeficientes de las covariables y que lo mismo ocurre entre el modelo clog-log y el modelo de Cox con tratamiento de empates utilizando el método Average.

Bondad de ajuste para modelos en tiempos discretos. Selección del modelo adecuado

Los criterios de Deviance, AIC y BIC fueron utilizados para comparar la bondad de ajuste de los modelos logit y clog-log, con diferentes conjuntos de covariables explicativas. Los resultados de éstos análisis para los otros subconjunto de datos de caballos fueron omitidos en este estudio, debido a la similitud que presentan en sus hallazgos.
En la Tabla 5.7, se muestran los submodelos (**SubM**), los cuales contemplan las siguientes covariables:
Submodelo (a) Indicadores de tiempo (IT); submodelo (b) IT, Lesión preexistente y género; submodelo (c) IT, número de carreras, distancia o longitud de carreras y época; submodelo (d) IT, Lesión preexistente, género, número de carreras, distancia de carreras y época. El **np** corresponde al número de parámetros de las covariables no tiempo, considerados en cada submodelo. También esta Tabla 5.7 contiene estadísticos para evaluar el ajuste de los diferentes submodelos.
En la Tabla 5.7, se aprecia que el submodelo (a) de los modelos logit y clog-log que solo incluyen indicadores de tiempo, por lo que tienen el menor número de parámetros, presentan altos valores de Deviance, AIC y BIC, y por lo tanto se espera que los valores de estos estadísticos de bondad de ajuste disminuyan al incorporar variables relevantes. Al comparar el submodelo (b) y (c), se observa un mejor ajuste para el modelo (c) lo que nos da una idea que las variables número de carreras, distancia o longitud de

carreras y época ajustan mejor a los datos que el modelo que solo incluyen las variables relacionadas a características propias del caballo, es decir lesión preexistente y género. Finalmente, es el submodelo (d) tanto para el modelo logit y clog-log, mostró menor valor de Deviance, AIC y BIC; luego parece que presenta el mejor ajuste entre los cuatros submodelos considerados, por lo tanto fue el modelo escogido, el cuál será utilizado en los demás análisis presentados en este capítulo.

Tabla 5.7: Modelo logit y clog-log para el tiempo discreto, con toda la base de datos.

		Modelo logit			**Modelo clog-log**		
SubM	**np**	**Deviance**	**AIC**	**BIC**	**Deviance**	**AIC**	**BIC**
a	0	508.15	522.15	15.40	508.15	522.15	15.40
b	2	507.17	529.17	26.78	506.97	528.97	26.58
c	3	427.97	453.97	-40.05	431.32	457.32	-36.70
d	5	414.66	448.66	-40.99	418.07	452.07	-37.58

Nota: Los Submodelos (**SubM**), corresponden a las siguientes covariables (a) Indicadores de tiempo (IT), (b) IT, Lesión preexistente y género; (c) IT, número de carreras, distancia o longitud de carreras y época; (d) IT, Lesión preexistente, género, número de carreras, distancia de carreras y época. El **np** corresponde al número de parámetros considerados en cada **SubM**.

5.7.2. Considerando los caballos con lesiones musculoesqueléticas a nivel de los miembros

Las lesiones en los miembros de los caballos tienen especial interés por las características anatómicas de los huesos que lo conforman, por lo tanto dentro de la base de datos de caballos de carrera que presentaron lesiones musculosqueléticas en el hipódromo *La Rinconada* en Venezuela del lapso 2000-2011, fueron considerados los 181 caballos que presentaron lesiones musculoesqueléticas en esta zona anatómica.

Tiempo considerado

Para el análisis de este subconjunto de datos fué considerado como tiempo discreto, el

tiempo en años hasta que ocurre una CMI a nivel de los miembros tanto anteriores como posteriores en los caballos de carreras.

Modelo de Cox con tratamientos de empates

En la Tabla 5.8 se presentan los diferentes resultados para los diferentes métodos de tratamientos de empates. Las covariables lesión preexistente y número de carreras, resultaron estadísticamente significativas en los diferentes métodos de tratamiento de empates utilizados.
Por otra parte la variable distancia de la carrera en la categoría mediana presenta diferencias estadísticamente significativas respecto a la distancia de categoría corta para los modelos de tratamientos de empates Efron (al 5 %), Average (al 5 %) y Exact (al 10 %), a diferencia del modelo Breslow que no detectó significancia en esta categoría de la variable. También, la variable distancia de la carrera en la categoría larga presenta diferencias significativas (al 10 %) respecto a la distancia de categoría corta para los modelos de tratamientos de empates Efron y Average .

Tabla 5.8: Parámetros estimados del Modelo de Cox con diferentes tratamientos de empates, en los miembros de los caballos.

	Modelo Efron		*Modelo Breslow*		*Modelo Exact*		*Modelo Average*	
$LesionP_1$								
Coeficiente	0.593		0.426		0.809		0.684	
exp(coef)	1.809		1.531		2.246		1.983	
EE	0.211		0.211		0.292		0.232	
p-valor	0.006	(**)	<0.001	(***)	<0.001	(***)	<0.001	(**)
Carreras								
Coeficiente	-0.173		-0.146		-0.233		-0.185	
exp(coef)	0.841		0.864		0.792		0.832	
EE	0.022		0.021		0.031		0.023	
p-valor	<0.001	(***)	<0.001	(***)	<0.001	(***)	<0.001	(***)
$LongC_2$								
Coeficiente	-0.474		-0.358		-0.512		-0.488	
exp(coef)	0.623		0.699		0.599		0.614	
EE	0.220		0.220		0.288		0.230	
p-valor	0.032	(*)	0.100		0.076	(·)	0.034	(*)
$LongC_3$								
Coeficiente	-0.539		-0.394		-0.623		-0.601	
exp(coef)	0.583		0.675		0.536		0.548	
EE	0.303		0.303		0.393		0.313	
p-valor	0.076	(·)	0.190		0.110		0.055	(·)
$Epoca_2$								
Coeficiente	0.265		0.195		0.338		0.290	
exp(coef)	1.303		1.215		1.402		1.337	
EE	0.195		0.194		0.260		0.203	
p-valor	0.180		0.320		0.190		0.150	
$Genero_2$								
Coeficiente	-0.118		-0.095		-0.076		-0.089	
exp(coef)	1.889		0.910		0.927		0.915	
EE	0.198		0.197		0.262		0.208	
p-valor	0.550		0.630		0.770		0.670	

Nota: (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

Modelos en tiempo discreto: logit y clog-log

En la Tabla 5.9 se presentan las estimaciones de los coeficientes α de los indicadores de tiempo. También se muestran las estimaciones de los coeficientes β de las covariables consideradas con sus respectivos errores estándar, tanto para el modelo logit como para el modelo clog-log.

Tabla 5.9: Parámetros estimados con los modelo logit y clog-log utilizando los datos de CMI en los miembros de los caballos.

	Modelo logit				**Modelo clog-log**			
IT	$\hat{\boldsymbol{\alpha}}$	**EE**	**p-valor**		$\hat{\boldsymbol{\alpha}}$	**EE**	**p-valor**	
d2	0.442	(0.353)	0.200		0.013	(0.263)	0.959	
d3	1.292	(0.446)	0.004	(**)	0.622	(0.334)	0.063	(·)
d4	1.886	(0.554)	<0.001	(***)	1.055	(0.423)	0.013	(*)
d5	3.990	(0.729)	<0.001	(***)	2.587	(0.526)	<0.001	(***)
d6	4.449	(0.880)	<0.001	(***)	3.010	(0.612)	<0.001	(***)
d7	7.429	(1.278)	<0.001	(***)	4.803	(0.835)	<0.001	(***)
Cov.	$\hat{\boldsymbol{\beta}}$	**EE**	**p-valor**		$\hat{\boldsymbol{\beta}}$	**EE**	**p-valor**	
$LesionP_1$	0.829	(0.295)	<0.01	(**)	0.681	(0.223)	0.002	(**)
$Carreras$	-0.238	(0.031)	<0.001	(***)	-0.185	(0.023)	<0.001	(***)
$LongC_2$	-0.517	(0.290)	0.075	(·)	-0.495	(0.229)	0.030	(*)
$LongC_3$	-0.632	(0.397)	0.110		-0.603	(0.313)	0.054	(·)
$Epoca_2$	0.342	(0.262)	0.019		0.296	(0.201)	0.141	
$Genero_2$	-0.074	(0.265)	0.780		-0.087	(0.204)	0.669	

Nota: $\hat{\boldsymbol{\alpha}}$= coeficientes de los indicadores de tiempo (**IT**); $\hat{\boldsymbol{\beta}}$ = coeficientes de las covariables (**Cov.**); (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

En la Tabla 5.9, en el modelo logit el coeficiente estimado para la covariable lesión pre-existente es estadísticamente significativo, resultando el valor de los odds proporcionales de padecer una CMI a nivel de miembros es mas del doble ($e^{0,829} = 2.29$) en caballos con

lesión preexistente respecto a los caballos que no tenían lesión.
Si consideramos el modelo clog-log, la lesión previa también resultó significativa; en este caso la exponencial del coeficiente puede interpretarse como un cociente de riesgos (com se ha explicado en el capítulo 3). En este sentido, tal como vemos en la Tabla 5.9 en el modelo clog-log, los caballos que presentaron lesión preexistente, tienen casi el doble de riesgo de una lesión CMI en los miembros en comparación a los caballos que no tenían lesión preexistente.
El ajuste del modelo logit y clog-log (ver Tabla 5.9) muestran la misma tendencia para los estimadores de los parámetros α y β. Sin embargo el modelo clog-log fue capaz de detectar significancias en la covariable distancia de la carrera categoría larga respecto a la distancia de categoría corta.

Con respecto a las estimaciones del los coeficientes del conjunto de indicadores de tiempo en ambos modelos, correspondientes a este subconjunto de la base de datos, se puede enfatizar lo siguiente:

1. El indicador de tiempo D_2, no resultó significativo, sin embargo debe permanecer en los modelos correspondientes debido a que todos los indicadores de tiempo deben ser tratados como un único conjunto, pues corresponden a la estimación de la función de riesgo basal.

2. En el modelo logit, los estimadores de los coeficientes α y sus respectivos errores estándar, toman valores mayores que los estimadores del modelo clog-log, sin embargo los p-valores son similares. Esta situación se presenta en los análisis obtenidos en este capítulo.

3. En ambos modelos estos estimadores tienen valores positivos, indicando que el riesgo de padecer una CMI incrementa a lo largo de tiempo.

5.7.3. Estudio según la localización de la lesión musculoesquelética

Dentro de la base de datos de caballos de carreras que presentaron lesiones musculosqueléticas, en el hipódromo *La Rinconada* en Venezuela del lapso 2000-2011, se focalizó el interés en 136 caballos de carrera que experimentaron lesiones musculoesqueléticas a nivel de miembros anteriores, dado que el 60 % del peso corporal recae en estos miembros. Diferentes sitios anátomicos de la lesión en los miembros anteriores se incluyeron en el análisis, según interés clínico; dando lugar a dos subpoblaciones:

1. Subpoblación I, comprende los 92 caballos con lesiones localizadas en el hueso sesamoide proximal (PSB). Esta subpoblación mostró 72 eventos, es decir, estos caballos padecieron una lesión que fue diagnosticada como CMI en el hueso sesamoide proximal, y para 20 caballos los tiempos fueron considerados censurados.
2. Subpoblación II, comprende los 44 caballos con lesiones localizadas en los huesos metacarpo, carpo y radio-cubito (OFS). En esta subpoblación, 29 caballos de carreras presentaron el evento de interés, es decir una CMI y 15 caballos fueron registrados con tiempos censurados.

5.7.4. Considerando los caballos con lesiones musculoesqueléticas a nivel del PSB: subpoblación I

El hueso sesamoide proximal, es un hueso que por su ubicación y tamaño es muy propenso a sufrir daños al momento de una lesión, por lo cual se tiene especial interés en su estudio.

Tiempo considerado

Para el análisis de este subconjunto de datos fué considerado como tiempo discreto el tiempo (en anõs) hasta que ocurre una CMI a nivel de PSB en los caballos de carreras.

Modelo de Cox con tratamientos de empates

En la Tabla 5.10 se presenta la estimación de los coeficientes y los errores estándar de las covariables estudiadas que corresponden a los caballos que sufrieron la CMI a nivel

del PSB. Al igual que el análisis donde se consideraron todos los caballos (Tabla 5.10), el método Exact es el que muestra los valores mayores en las estimaciones del error estándar.
En estos resultados, se puede observar que para todos los tratamientos de empates del modelo de Cox, la variable número de carreras resultó estadísticamente significativa. Solo el modelo de Cox con tratamiento de empates Average detectó que la covariable época como significativa a nivel del 10 %. Así, es importante considerar que dada las condiciones de Venezuela de ser un país tropical la época de verano corresponde a los meses noviembre-abril, en la cual la pista se hace dura por compactación y probablemente durante la carrera hay una mayor sobreextensión del ligamento suspensorio, que afecta por su ubicación anatómica a los huesos sesamoides. Es decir, los huesos sesamoides (huesos cortos neumáticos), encargados de disipar las cargas (los vectores de fuerza durante el ejercicio), al momento del contacto con la superficie, (cuya dureza esta incrementada) aumentan el trabajo que unido a sobrepasar la capacidad fisiológica del aparato suspensor, pueden provocar la CMI; por ello el mayor número de incidencias involucra a los huesos sesamoideos.

Modelos en tiempo discreto logit y clog-log

En la Tabla 5.11 se presenta las estimaciones de los coeficientes α de los indicadores de tiempo y la estimaciones de los coeficientes β de las covariables consideradas, con sus respectivos errores estándar, tanto para el modelo logit como para el modelo clog-log, para el conjunto de datos correspondiente a los caballos que sufrieron la CMI a nivel del PSB.

Tabla 5.10: Parámetros estimados del Modelo de Cox con diferentes tratamiento de empates. Subpoblación I.

	Modelo Efron		*Modelo Breslow*		*Modelo Exact*		*Modelo Average*	
$LesionP_1$								
Coeficiente	0.005		-0.019		0.058		0.085	
exp(coef)	1.005		0.981		1.059		1.089	
EE	0.290		0.287		0.404		0.317	
p-valor	0.990		0.950		0.890		0.788	
Carreras								
Coeficiente	-0.160		-0.133		-0.250		-0.188	
exp(coef)	0.852		0.876		0.779		0.829	
EE	0.025		0.024		0.042		0.030	
p-valor	<0.001	(***)	<0.001	(***)	<0.001	(***)	<0.001	(***)
$LongC_2$								
Coeficiente	-0.232		-0.169		-0.180		-0.263	
exp(coef)	0.793		0.845		0.836		0.769	
EE	0.284		0.283		0.396		0.301	
p-valor	0.410		0.550		0.650		0.380	
$LongC_3$								
Coeficiente	-0.467		-0.334		-0.529		-0.597	
exp(coef)	0.627		0.716		0.589		0.550	
EE	0.378		0.375		0.532		0.413	
p-valor	0.220		0.370		0.320		0.150	
$Epoca_2$								
Coeficiente	0.135		0.108		0.276		0.181	
exp(coef)	1.145		1.114		1.318		1.199	
EE	0.261		0.258		0.362		0.281	
p-valor	0.610		0.680		0.450		0.052	(·)
$Genero_2$								
Coeficiente	-0.274		-0.211		-0.272		-0.225	
exp(coef)	0.760		0.809		0.762		0.799	
EE	0.243		0.243		0.359		0.266	
p-valor	0.260		0.380		0.450		0.400	

Nota: (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

Los modelos logit y clog-log en la Tabla 5.11, revelan en los diferentes periodos de tiempo un incremento en el riesgo de padecer una CMI a nivel de hueso sesamoide, en la medida que incrementa la edad de los caballos. Para el modelo logit todos los indicadores de tiempo resultaron estadísticamente significativos, no así en el modelo clog-log donde el indicador de tiempo correspondiente a los 2 años resultó no significativo.
En referencia a las covariables, se muestra que solo la variable número de carreras resultó estadísticamente significativa en los dos modelos. Para el modelo clog-log podemos decir que el riesgo disminuye un 17 % (e^{β}=0.828), al aumentar una carrera. Por lo tanto, esta variable es un factor protector relacionado a padecer una CMI a nivel de PSB.

Tabla 5.11: Parámetros estimados con los modelo logit y clog-log. Subpoblación I.

	Modelo	**logit**		**Modelo**	**clog-log**	
IT	$\hat{\boldsymbol{\alpha}}$	**EE**	**p-valor**	$\hat{\boldsymbol{\alpha}}$	**EE**	**p-valor**	
d2	0.954	(0.481)	0.047	(*)	0.349	(0.334)	0.296	
d3	2.072	(0.595)	<0.001	(***)	1.157	(0.403)	0.004	(**)
d4	3.029	(0.741)	<0.001	(***)	1.822	(0.509)	<0.001	(***)
d5	5.703	(1.054)	<0.001	(***)	3.562	(0.680)	<0.001	(***)
d6	6.657	(1.404)	<0.001	(***)	4.363	(0.878)	<0.001	(***)
d7	9.515	(1.741)	<0.001	(***)	6.707	(1.114)	<0.001	(***)
Cov.	$\hat{\boldsymbol{\beta}}$	**EE**	**p-valor**		$\hat{\boldsymbol{\beta}}$	**EE**	**p-valor**	
$LesionP_1$	0.077	(0.411)	0.852		0.112	(0.310)	0.718	
$Carreras$	-0.260	(0.043)	<0.001	(***)	-0.190	(0.030)	<0.001	(***)
$LongC_2$	-0.182	(0.402)	0.649		-0.252	(0.298)	0.398	
$LongC_3$	-0.541	(0.540)	0.316		-0.613	(0.409)	0.134	
$Epoca_2$	0.294	(0.368)	0.425		0.191	(0.273)	0.484	
$Genero_2$	-0.272	(0.363)	0.455		-0.221	(0.263)	0.399	

Nota: $\hat{\boldsymbol{\alpha}}$= coeficientes de los indicadores de tiempo (**IT**); $\hat{\boldsymbol{\beta}}$ = coeficientes de las covariables (**Cov.**); (***) 0,1 %; (**) 1 %; y (*)5 % de significación estadística.

5.7.5. Considerando los caballos con lesiones musculoesqueléticas a nivel del OFS: subpoblación II

Tiempo considerado

Para el análisis de este subconjunto de datos fué considerado como tiempo discreto el tiempo (en anõs) hasta que ocurre una CMI a nivel de OFS en los caballos de carreras.

Modelo de Cox con tratamientos de empates

En la Tabla 5.12 se presenta la estimación de los coeficientes y los errores estándar de las covariables estudiadas que corresponden a los caballos que sufrieron la CMI a nivel de OFS.

Los resultados obtenidos, en este punto, donde se consideró los caballos que sufren una lesión nivel de los huesos metacarpo, carpo y radio-cubito, han mostrado mayor número de variables estadísticamente significativas: lesión preexistente, número de carreras, distancia de carrera en categoría largas y época.

Por otra parte, en estos resultados en similitud con los casos estudiados en las subsecciones 5.7.1, 5.7.2 y 5.7.4 muestran que el modelo de Cox con el método Exact muestra los valores mayores en las estimaciones de los errores estándar. A diferencia de los resultados de los casos anteriores, la covariable época en el modelo de Cox con tratamiento de empates Efron, Exact y Average mostró diferencias estadísticamente significativas al 10 % y 5 % y la covariable distancia con categoría larga resultó estadísticamente significativa en todos los métodos de tratamiento de empates.

Tabla 5.12: Parámetros estimados del Modelo de Cox con diferentes tratamiento de empates. Subpoblación II.

	Modelo Efron		*Modelo Breslow*		*Modelo Exact*		*Modelo Average*	
$LesionP_1$								
Coeficiente	1.004		0.813		1.950		1.431	
exp(coef)	2.729		2.254		7.029		4.183	
EE	0.440		0.433		0.797		0.615	
p-valor	0.020	(*)	0.060	(·)	0.014	(*)	0.020	(*)
Carreras								
Coeficiente	-0.168		-0.134		-0.274		-0.231	
exp(coef)	0.846		0.874		0.761		0.794	
EE	0.053		0.051		0.090		0.074	
p-valor	<0.001	(***)	0.009	(**)	0.002	(**)	0.002	(**)
$LongC_2$								
Coeficiente	-0.710		-0.472		-0.988		-0.744	
exp(coef)	0.492		0.624		0.372		0.475	
EE	0.504		0.513		0.729		0.598	
p-valor	0.160		0.360		0.180		0.213	
$LongC_3$								
Coeficiente	-2.115		-1.695		-3.441		-2.619	
exp(coef)	0.121		0.184		0.032		0.073	
EE	0.797		0.799		1.117		0.893	
p-valor	<0.01	(**)	0.034	(*)	0.002	(**)	0.003	(**)
$Epoca_2$								
Coeficiente	0.847		0.642		1.662		1.276	
exp(coef)	2.332		1.901		5.270		3.581	
EE	0.458		0.448		0.770		0.587	
p-valor	0.065	(·)	0.150		0.031	(*)	0.030	(*)
$Genero_2$								
Coeficiente	-0.535		-0.391		-0.809		-0.769	
exp(coef)	0.586		0.676		0.445		0.463	
EE	0.415		0.411		0.652		0.522	
p-valor	0.200		0.340		0.210		0.141	

Nota: (***) 0,1 %; (**) 1 %; (*)5 % y (·) 10 % de significación estadística.

Modelos en tiempo discreto: logit y clog-log

En la Tabla 5.13 se presentan los modelos logit y clog-log con sus respectivas estimaciones de los coeficientes α y β y sus estimaciones de los errores estándar; estos resultados corresponden a los caballos que sufrieron la CMI a nivel del OFS.
A diferencia del resultado en caballos que padecieron una CMI a nivel de PSB (ver subsección 5.7.4), no presentan una tendencia ni creciente ni decreciente a lo largo del tiempo. En esta subpoblación OFS el indicador de tiempo 7 no fue estimado debido al pequeño número de caballos en este periodo. Además, en el modelo logit los indicadores de tiempo correspondientes a los 3, 5 y 6 años son estadísticamente significativos, a diferencia del modelo clog-log que mostró a los indicadores de tiempo 5 y 6 como significativas.
Todas las covariables consideradas resultaron estadísticamente significativas, con la excepción de la covariable género, lo que parece indicar que no afecta el riesgo de padecer una CMI a nivel de OFS.
Los resultados del modelo clog-log (ver Tabla 5.13), muestran que en los caballos que presentaron lesión preexistente se incrementa mas de cuatro veces ($e^{1,458} = 4.30$) el riesgo de padecer una CMI en comparación a los caballos que no tenían lesión preexistente. En esta estudio no se consideraron los tipos de lesiones preexistente registradas durante el estudio, sin embargo se podría pensar que la presencia de osteoartritis o la existencia de una alteración conformacional como el defecto angular de los huesos del carpo (conocido como rodilla hacia atrás o distal) puede predisponer a fracturas del carpo, por lo tanto, a una lesión del sistema locomotor. La misma predisposición pudiera ocurrir en los casos de periostitis y osteítis a nivel de la corteza dorsal del metacarpo, que conlleva a una disminución del grado de mineralización y en casos donde el daño es permanente se presenta una excesiva respuesta por parte de los osteoclastos, lo que se manifiesta con microfracturas en la fase más avanzada. Además, en muchos casos puede ser determinante la enfermedad degenerativa articular, tendinitis crónica, desmitis, enfermedad periosteal del tercer metacarpiano/metacarpo principal y miopatías de esfuerzo. Por otra parte, la inflamación de la zona distal del radio afecta a caballos jóvenes y a caballos donde debido a que los huesos de carpo o metacarpo se encuentran afectados, toda la carga repercute sobre este hueso. También, las lesiones del radio inclusive del cubito, posiblemente se deben al grado de inclinación de la pista y la prolongación de las curvas

de la pista, aunado a las fallas en la disipación de las cargas (vectores de fuerza) de los huesos sesamoides proximales, aparato suspensor, metacarpo y huesos del carpo.
Por otra parte, en este análisis, se encontró que el número de carreras fue estadísticamente significativa por lo tanto el número de carreras está asociado con el riesgo de padecer una CMI a nivel de OFS; por cada carrera adicional que tenga un caballo, el riesgo disminuye un 20 %, donde $e^{-0,226} = 0.80$, esto nos lleva a pensar que la experiencia adquirida por los caballos durante las carreras (sin olvidar la necesidad de entrenamiento) es fundamental para favorecer la fortaleza y estabilidad de los miembros del animal mediante la adaptación y acondicionamiento del sistema músculoesquelético del animal, disminuyendo la posibilidad de riesgo en que se produzcan CMI.
La covariable distancia de la carrera también representa un factor protector para la CMI a nivel de OFS. En los caballos de carrera que compiten en distancias de categoría largas (1800 - 2400 m) se tiene un 94 % $e^{-2,786}$ más del riesgo en comparación con caballos que compiten en carreras de distancia cortas (800 - 1200 m). Lo que nos está sugiriendo que bajo estas condiciones los caballos en carreras de distancias cortas están sometidos a mayores esfuerzos debido a la exigencia de mayor velocidad en menor tiempo, que a su vez genera en el caballo un mayor estrés y fatiga, lo que se traduce en un factor predisponente a sufrir una CMI.
En este mismo orden de ideas, el análisis estadístico reveló que caballos que compiten en época de verano tienen tres veces más del riesgo de sufrir una CMI a nivel de OFS que caballos que compiten en época de lluvia. Posiblemente esto podría ser explicado por la influencia que ejerce las condiciones climáticas durante el verano sobre la composición y firmeza de la pista de carreras (que son de tierra en su totalidad), la cual se torna dura, por lo que el caballo bajo estas condiciones estaría ejerciendo un mayor impacto sobre las estructuras óseas que conforman la parte distal del carpo.

Tabla 5.13: Parámetros estimados con los modelo logit y clog-log. Subpoblación II.

	Modelo	**logit**		**Modelo**	**clog-log**	
IT	$\hat{\boldsymbol{\alpha}}$ **EE**	**p-valor**		$\hat{\boldsymbol{\alpha}}$ **EE**	**p-valor**	
d2	0.809 (0.944)	0.392		0.239 (0.638)	0.708	
d3	2.220 (1.192)	0.063	(·)	1.238 (0.840)	0.140	
d4	0.939 (1.618)	0.562		0.258 (1.331)	0.846	
d5	4.813 (1.869)	0.010	(*)	3.487 (1.409)	0.013	(*)
d6	6.642 (2.179)	0.002	(**)	4.961 (1.504)	<0.001	(***)
Cov.	$\hat{\boldsymbol{\beta}}$ **EE**	**p-valor**		$\hat{\boldsymbol{\beta}}$ **EE**	**p-valor**	
$LesionP_1$	2.036 (0.823)	0.013	(*)	1.458 (0.589)	0.013	(*)
$Carreras$	-0.287 (0.093)	0.002	(**)	-0.226 (0.072)	0.002	(**)
$LongC_2$	-1.092 (0.771)	0.157		-0.727 (0.591)	0.219	
$LongC_3$	-3.717 (1.202)	0.002	(**)	-2.786 (0.905)	0.002	(**)
$Epoca_2$	1.746 (0.799)	0.029	(*)	1.319 (0.560)	0.019	(*)
$Genero_2$	-0.872 (0.675)	0.197		-0.800 (0.495)	0.106	

Nota: $\hat{\boldsymbol{\alpha}}$= coeficientes de los indicadores de tiempo (**IT**); $\hat{\boldsymbol{\beta}}$ = coeficientes de las covariables (**Cov.**); (***) 0.1 %, (**) 1 % ,(*) 5 % y (·) 10 % de significanción estadística.

Capítulo 6

APLICACIÓN DE MÉTODOS DE ANÁLISIS DE SUPERVIVENCIA EN TIEMPO DISCRETO CON FRAILTY

6.1. Introducción

En análisis de supervivencia a menudo los individuos en estudio forman grupos o *cluster*, como es el caso de animales que están agrupados en granjas, piaras, hatos o zonas geográficas. En estos casos, se considera que existe heterogeneidad no observada. Para tomar en cuenta que los datos forman *cluster*, se puede plantear un modelo de análisis de supervivencia con efecto aleatorio, el cual proporciona un enfoque útil para estimar simultáneamente los parámetros de las covariables y la varianza del efecto aleatorio correspondiente al *cluster*.

Uno de los objetivos de esta tesis es comparar algunas funciones utilizadas en software estadísticos para analizar datos de análisis de supervivencia para tiempos discretos con uno o dos términos de frailty. Para ello se han analizado una muestra de 437

vacas de producción de leche, provenientes de diferentes zonas geográficas (Mérida, Trujillo y Zulia) de Venezuela. El interés clínico es estudiar el número de lactancias hasta el primer diagnóstico de mastitis en estas vacas. Dado que la vacas se encuentran agrupadas dentro de los rebaños y estos últimos a su vez forman parte de diferentes zonas geográficas, es conveniente el uso de un o dos términos de frailty en los modelos. Con estos datos se han ajustado los modelos logit y clog-log con uno o dos términos de frailty y el modelo de Cox con tratamiento de empates también con uno o dos términos de frailty; estimando las covariables fijas y la varianza de los términos de frailty.
Los resultados se han obtenido utilizando los software R, Stata y SAS, lo que ha permitido realizar una comparativa de las distintas metodologías implementadas en cada una de ellos. Entre los software utilizados, está la función frailtyHL del R (Ha et al., 2012), que permite elegir la disribución normal y la log-gamma para el término de frailty y también proporciona los indicadores de bondad de ajuste AIC y Deviance, además de las estimaciones de los efectos fijos, de la varianza de los efectos aleatorios así como sus errores estándar. Otros métodos que han sido utilizados son: del software Stata el xtlogit, xtcloglog y gllamm, donde se requiere el conjunto de datos individuo-periodo y del software SAS, los procedimientos proc phreg y proc glimmix.

La organización de este capítulo es la siguiente: una sección 6.2 donde se describen los métodos utilizados, el conjunto de datos correspondiente al número de lactancias hasta el primer diagnóstico de mastitis en vacas productoras de leche y la conversión dl conjunto individuo-nivel al conjunto individuo-periodo. En la sección 6.3 se muestran los resultados de la aplicación de los diferentes modelos a este conjunto de datos.

6.2. Metodología y datos

6.2.1. Métodos

Un resumen de los modelos discutidos en el capítulo 4 que serán utilizados durante el desarrollo de este capítulo, se presentan a continuación.

6.2.2. Modelo de Cox con un término de frailty

En este caso la función de riesgo contiene tres componentes: una función de riesgo basal, un término que modela la influencia de las covariables observadas y dos (o más) términos frailty:

$$\lambda(t|\mathbf{Z}, \nu_j, \mu_k) = \lambda_0(t) \exp(\beta'\mathbf{Z} + \nu_j) \tag{6.1}$$

donde λ_0 es una función de riesgo basal no especificada β es un vector de parámetros de regresión para covariables de efectos fijos $\mathbf{Z}$ y ν_j es el término de frailty. De igual forma se procedería con dos o más términos de frailty.

6.2.3. Modelos logit y clog-log con un término de frailty

El término de frailty se incorpora a los modelos logit y clog-log de la forma siguiente:

$$\text{logit}(h\,(t|\mathbf{Z}, \nu_k)) = (\alpha_1 D_1 + \cdots + \alpha_J D_J) + (\beta_1 Z_1 + \cdots + \beta_p Z_p) + \nu_k$$

$$\text{clog-log}(h\,(t|\mathbf{Z}, \nu_k)) = (\gamma_1 D_1 + \cdots + \gamma_J D_J) + (\beta_1 Z_1 + \cdots + \beta_p Z_p) + \nu_k \tag{6.2}$$

donde ν_k corresponde al término de frailty De igual forma se definirían los modelos logit y clog-log con dos o más términos de frailty.

6.2.4. Presentación del conjunto de datos

Descripción del conjunto de datos: diagnóstico de mastitis en vacas lecheras

La base de datos contiene la información para una muestra de 437 vacas, provenientes de diferentes zonas geográficas (Mérida, Trujillo y Zulia) de Venezuela. El interés clínico era determinar el número de lactancias hasta el primer diagnóstico de mastitis en las

vacas en producción de leche. Es importante resaltar que la mastitis es la inflamación de la glándula mamaria y es considerada una enfermedad que afecta a nivel mundial. Tiene un alto impacto en las pérdidas económicas derivadas de la baja productividad y el deterioro en la calidad de la leche, por el costo de los antibióticos empleados, por los servicios veterinarios (Guiraudo et al., 1997), así como por las horas extras de trabajo a los jornaleros y por el sacrificio obligado y reemplazo de los animales.
En este estudio, el tiempo de interés fue definido como el número de lactancias (tiempo discreto) hasta que ocurre el primer diagnostico de infección de mastitis, representando el evento de interés. De las 437 vacas, 338 vacas (77,3 %) resultaron positivas al diagnostico de mastitis y para las 99 vacas restantes, sus tiempos se consideraron observaciones censuradas. El tiempo de seguimiento tiene una mediana de 6 lactancias (IC=(6;7)). Las covariables incluidas en este estudio son: la enfermedad previa, entre ellas neumonía, enfermedades podales y reproductivas (retención de placenta y metritis); el tipo de ordeño utilizado (mecánico o manual); y el indicador de si se les suministra suplemento alimenticio o no. La estación o época climática (época seca y lluviosa por ser Venezuela un país trópical), fue utilizada para estratificar el conjunto de datos. Los resultados de la sección 6.3 corresponden al estrato época seca; los resultados correspondientes al estrato época lluviosa no son mostrados, debido a los pocos casos registrados en esta época. Los animales están agrupados en 15 rebaños los cuales corresponden a 3 zonas geográficas de Venezuela: Mérida, Trujillo y Zulia (ver Figura 6.1). La Tabla 6.2 muestra la estadística descriptiva de estas covariables.

Tabla 6.1: Descripción de las covariables consideradas para el riesgo de mastitis en vacas lecheras de tres zonas geográficas de Venezuela.

Covariable	Descripción
EnfermedadP.	Enfermedad previa, sin (1) y con (2)
Estación	Estación seca: desde Mayo a Octubre (1) y Estación lluviosa: desde Noviembre a Abril (2)
T. ordeño	Tipo de ordenño: mecánico (1) y manual (2)
Suplemento	suplemento alimenticio para optimizar la producción: con (1) y sin suplemento (2)
Rebaño	Grupo de animales con características especificas (del 1 al 15).
ZonaG.	Zona geográfica: Mérida (1), Trujillo (2) y Zulia (3)

Tabla 6.2: Estadística descriptiva de las covariables usadas para nuestra base de datos de vacas en producción de leche.

Covariables	*Mastitis*		*Total*	
	no	*si*	*n*	*%*
EnfermedadP.				
No	37	132	169	38.7
Si	62	206	268	61.3
T. ordeño				
mecánico	5	66	71	16.2
manual	94	272	366	83.8
Suplemento				
con	44	132	176	40.3
sin	55	206	261	59.7
Estación				
seca	92	281	373	85.4
lluviosa	7	57	64	14.6

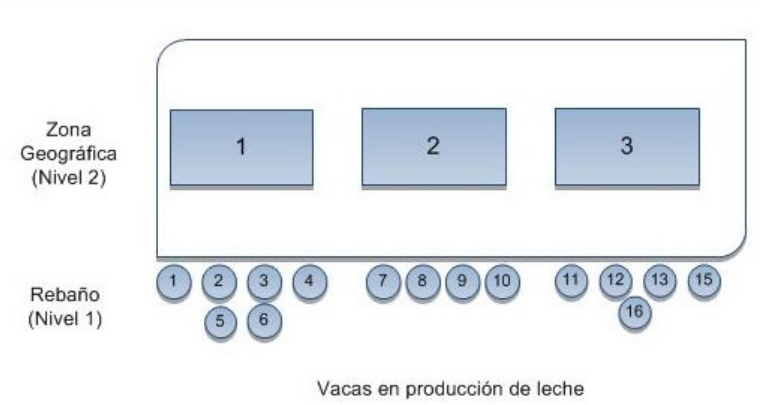

Figura 6.1: Estructura en rebaños y zonas geogáficas de la muestra de vacas en producción de leche

Preparación del conjunto de datos individuo-periodo

Para poder ajustar los modelos discretos logit y clog-log con término de frailty, igual que hemos hecho en los capítulos 3 y 5, se requiere la conversión del conjunto de datos originales (individuo-nivel), a un conjunto de datos expandidos (individuo-periodo).

La Tabla 6.3 muestra los datos correspondientes a tres vacas en producción de leche, la segunda columna muestra sus respectivos tiempos de interés; la tercera columna muestra si la vaca fue diagnósticada de mastitis o por el contrario esa observación es una censura; la cuarta columna indica la condición de la vaca en función si presentó enfermedad previa y las dos últimas columnas muestran el rebaño y zona geográfica al que pertenece la vaca. A partir de esta base de datos originales, se procede a realizar una conversión de toda la base de datos.

1. Duración (T): el número de lactancias observadas hasta el primer diagnostico de mastitis o hasta el final del seguimiento.
2. Indicador de censura (C): variable binaria, con valor igual a 1 si la vaca ha sido diagnostica de mastitis y 0 en caso contrario.
3. Covariables fijas (Z): son el conjunto de variables explicativas de interés.
4. Efectos aleatorios: rebaño y zona geográfica.

Tabla 6.3: Conjunto individuo-nivel, vacas en producción de leche al primer diagnóstico de mastitis.

ID_i	T_i	C	*EnfermedadP*	*Rebaño*	*ZonaG.*
7	4	1	2	12	3
169	5	1	2	3	1
410	2	0	1	9	2

Así, a partir de la Tabla 6.3, se obtiene una base de datos con la estructura como se muestra en la Tabla 6.4, para el caso especifico de nuestro conjunto de datos, donde cada vaca tendrá tantas filas como lactancias hasta la mastitis o hasta la última lactancia registrada. De igual forma, en el nuevo conjunto de datos individuo-periodo, se muestra un conjunto de indicadores de tiempo, donde cada indicador proporciona información de si el individuo estaba en riesgo en el periodo correspondiente a la fila y una variable binaria que registra si la vaca experimentó el evento de interés (mastitis) en aquel periodo. Por ejemplo en la Tabla 6.4 se muestran tres vacas del conjunto de datos individuo-periodo, así, la vaca $i = 169$ ha sido observado por cinco periodos de tiempo desde 1 a 5, ésta experimento el evento de interés y con $Z = 2$ como valor de su covariable.

Tabla 6.4: Conjunto de datos individuo-periodo, vacas en producción de leche al primer diagnóstico de mastitis.

ID_i	T_i	y_{ijk}	D_{i1}	D_{i2}	D_{i3}	D_{i4}	D_{i4}	$\cdots$	D_{i8}	*EnfermedadP*	*Rebaño*	*ZonaG.*
7	1	0	1	0	0	0	0		0	2	12	3
7	2	0	0	1	0	0	0		0	2	12	3
7	3	0	0	0	1	0	0		0	2	12	3
7	4	1	0	0	0	1	0		0	2	12	3
169	1	0	1	0	0	0	0		0	2	3	1
169	2	0	0	1	0	0	0		0	2	3	1
169	3	0	0	0	1	0	0		0	2	3	1
169	4	0	0	0	0	1	0		0	2	3	1
169	5	1	0	0	0	0	1		0	2	3	1
410	1	0	1	0	0	0	0		0	1	9	2
410	2	0	0	1	0	0	0		0	1	9	2

6.3. Resultados

6.3.1. Aplicación de los Modelos en tiempo discreto con término frailty

El primer lugar se compararon los modelos clog-log con y sin término de frailty, con la finalidad de conocer si existe un cambio de magnitud y significancia de las covariables y evaluar el efecto del término frailty. Para esto se consideraron las funciones clog-log y xtcloglog del software estadístico Stata. Los resultados de este análisis se muestran en la Tabla 6.5. Aquí, los coeficientes de los indicadores de las lactancias (L1, L2,...,L8) son de menor magnitud en comparación al modelo clog-log en ausencia de efecto aleatorio, sin embargo la significancia estadística se mantiene.
También se puede observar un cambio en la magnitud de los coeficientes correspondientes a las covariables fijas, donde los valores incrementan.

Tabla 6.5: Modelos ajustados clog-log sin y con frailty

Indicadores	*Modelo clog-log*		*Modelo clog-log con frailty*	
de lactancia	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
L1	-2.670 (0.285)	<0.001	-2.934 (0.338)	<0.001
L2	-2.182 (0.246)	<0.001	-2.431 (0.306)	<0.001
L3	-2.377 (0.274)	<0.001	-2.608 (0.328)	<0.001
L4	-0.934 (0.186)	0.001	-1.121 (0.259)	0.001
L5	-1.967 (0.273)	<0.001	-2.072 (0.325)	<0.001
L6	-1.544 (0.250)	<0.001	-1.602 (0.307)	<0.001
L7	-0.675 (0.211)	0.001	-0.602 (0.277)	0.030
L8	0.584 (0.198)	0.003	1.003 (0.275)	<0.001
Covariables	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
EnfermedadP.2	0.063 (0.166)	0.125	0.135 (0.129)	0.296
T. ordeño2	-0.672 (0.166)	<0.001	-0.644 (0.178)	<0.001
Suplemento2	0.084 (0.127)	0.510	0.164 (0.136)	0.227
Efecto aleatorio			*Estimación (EE)*	*I.C.95 %*
Rebaño ν(sigma)			0.644 (0.141)	0.418-0.989
ρ			0.201 (0.071)	0.096-0.373

Para ambos modelos la covariable fija tipo de ordeño resultó estadísticamente significativa, lo que esta indicando que las vacas que son ordeñadas en forma manual tienden a presentar menos riesgo de padecer mastitis en comparación a vacas que son ordeñadas en forma mécanica. Una posible explicación de esto, es que las vacas con ordeño manual están sometidas a menos estrés y la práctica de ordeño manual deja al becerro al lado de su madre lo que contribuye a que la saliva del becerro permita con mayor efectividad y de forma natural el sellado de los pezones. Además hay que considerar que actualmente existen problemas de mantenimiento del equipo de las maquinas ordeñadoras, especialmente el cambio de las gomas de las pezoneras que pudieran afectar a la presión de las maquinas ordeñadoras.

Para el término frailty correspondiente al rebaño fue asumida una distribucción $\mathrm{N}(0, \sigma_\nu^2)$. La varianza σ_ν^2 resultó estadísticamente significativa, lo que nos indica que el efecto rebaño influye en el riesgo de que las vacas contraigan mastitis. Una característica relevan-

te en los modelos con frailty es el coeficiente de correlación intraclase ρ, definido en el capítulo 4 (ver ecuación 4.5), que se interpreta como la proporción de la varianza total explicada por los diferentes clústers. En la Tabla 6.5 se muestra la estimación de ρ que resultó estadísticamente significativo, por lo que se puede concluir que un 20 % de la varianza total es atribuible al rebaño.

Comparación de modelos logit y clog-log con un término frailty

El segundo lugar fueron comparados los modelos logit y clog-log con un término de frailty asociado al rebaño, usando las funciones xtlogit and xtcloglog (Stata), asumiendo una distribución normal para este término. Los resultados se muestran en la Tabla 6.6, donde al comparar los indicadores de tiempo en ambos modelos, se puede apreciar que en el modelo logit tanto las estimaciones de los coeficientes y sus respectivos errores estándar son mayores que los obtenidos para el modelo clog-log. El indicador de lactancia L7 resultó estadísticamente significativo solo en el modelo clog-log.
Para ambos modelos la covariable fija tipo de ordeño resultó estadísticamente significativa (Tabla 6.6). Además, en ambos modelos el efecto rebaño (varianza del término de frailty considerado), resultó estadísticamente significativo.

Comparación de modelos logit y clog-log con dos términos de frailty

Usando las mismas covariables y considerando dos efectos aleatorios fueron ajustados los modelos logit y clog-log asumiendo una distribución normal para los dos términos de frailty. Los resultados se obtuvieron utilizando la función `gllamm` del software Stata.

Como se puede apreciar en la Tabla 6.7, igual que en los otros modelos, solo la covariable tipo de ordeño presentó diferencias significativas. En comparación con los modelos con un solo factor aleatorio, la magnitud de los coeficientes son ligeramente menores para los indicadores de lactancia, y mayores para las covariables fijas. Por lo que respecta a los términos de frailty, el efecto rebaño disminuye pero se mantiene estadísticamente significativo mientras que el efecto zona geográfica no resulta estadísticamente significativo.

Tabla 6.6: Ajuste de los modelos logit y clog-log con un término de frailty

Indicadores	*Modelo logit con frailty*		*Modelo clog-log con frailty*	
de lactancia	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
L1	-2.821 (0.367)	<0.001	-2.934 (0.338)	<0.001
L2	-2.305 (0.339)	<0.001	-2.431 (0.306)	<0.001
L3	-2.474 (0.360)	<0.001	-2.608 (0.328)	<0.001
L4	-0.847 (0.303)	0.005	-1.121 (0.259)	0.001
L5	-1.910 (0.360)	<0.001	-2.072 (0.324)	<0.001
L6	-1.423 (0.345)	<0.001	-1.602 (0.306)	<0.001
L7	-0.291 (0.328)	0.375	-0.602 (0.277)	0.030
L8	1.808 (0.358)	<0.001	1.003 (0.274)	<0.001
Covariables	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
EnfermedadP.2	0.134 (0.152)	0.378	0.135 (0.129)	0.296
T. ordeño2	-0.780 (0.212)	<0.001	- 0.643 (0.178)	<0.001
Suplemento2	0.152 (0.157)	0.334	0.163 (0.135)	0.227
Efecto aleatorio	*Estimación (EE)*	I.C.95 %	*Valor (EE)*	I.C.95 %
Rebaño ν	0.717 (0.160)	0.463-1.108	0.643 (0.141)	0.418-0.989
ρ	0.135 (0.052)	0.061-0.272	0.201 (0.070)	0.096-0.373

Tabla 6.7: Ajuste de los modelos logit y clog-log con dos términos de frailty

	Modelo logit		*Modelo clog-log*	
Indicadores	*con dos términos de frailty*		*con dos términos de frailty*	
de lactancia	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
L1	-2.447 (0.612)	<0.001	-2.700 (0.540)	<0.001
L2	-1.933 (0.596)	0.001	-2.199 (0.521)	<0.001
L3	-2.105 (0.609)	0.001	-2.378 (0.534)	<0.001
L4	-0.477 (0.579)	0.410	-0.888 (0.494)	0.072
L5	-1.535 (0.605)	0.011	-1.837 (0.530)	0.001
L6	-1.051 (0.600)	0.080	-1.370 (0.522)	0.009
L7	0.078 (0.596)	0.896	-0.371 (0.507)	0.465
L8	2.187 (0.613)	<0.001	1.242 (0.504)	0.014
Covariables	*Coeficiente (EE)*	*p-valor*	*Coeficiente (EE)*	*p-valor*
EnfermedadP.2	0.138 (0.152)	0.365	0.137 (0.129)	0.288
T. ordeño2	-0.771 (0.211)	<0.001	- 0.639 (0.177)	<0.001
Suplemento2	0.184 (0.158)	0.244	0.200 (0.137)	0.144
Efectos aleatorios	*Estimación* (EE)		*Estimación (EE)*	
Rebaño ν	0.521 (0.153)		0.460 (0.116)	
ZonaG. μ	0.522 (0.296)		0.479 (0.243)	

6.3.2. Aplicación del Modelo de Cox con frailty

Los resultados de esta sección se han obtenido utilizando tanto el el software SAS como el R. Con SAS mediante el procedimiento `phreg` y la opción `ties=exact`, que equivale al método Averge, permitió obtener, además de los coeficientes del modelo, estimar el efecto rebaño sobre la función de riesgo. Por otro lado, con el paquete FrailtyHL de R, se obtuvieron los resultados para modelos con uno y dos términos de frailty, con el inconveniente que la única aproximación para el tratamiento de los empates es Breslow; no obstante permite tener una comparativa entre los diferentes modelos de Cox con frailty.

Modelo de Cox con tratamiento Average y un término de frailty

La estimación de los parámetros del modelo están en la Tabla 6.8. No se observan diferencias por lo que se refiere a la interpretación, con los modelos presentados anteriormente. La estimación de la desviación estándar del efecto aleatorio fue de 0.447 y su error estándar 0.193.

Tabla 6.8: Modelo de Cox con frailty usando el método Average

Covariables	*Coeficiente*	*exp(coef)*	*EE*	*p-valor*
EnfermedadP.2	0.135	1.145	0.129	0.295
T. ordeno2	-0.641	0.527	0.178	<0.001
Suplemento2	0.164	1.178	0.136	0.228
Rebaño (ν)	0.447		0.193	

En la Tabla 6.9 se puede observar el efecto del rebaño sobre la función de riesgo condicionando a unas misma características representadas por las covariables introducidas en el modelo. Se presentan también los correspondientes intervalos de confianza. Cuando estos contienen el valor 1 indican que el efecto rebaño no es estadísticamente significativo sobre el riesgo.

En la Tabla 6.9 podemos destacar tres grupos de rebaños:

1. Las vacas de los rebaños 4, 5, 8 y 9 tienen un menor riesgo de padecer mastitis. Por ejemplo en el rebaño 4, la estimación es e^{ν}=0.440 lo que significa que la pertenencia

Tabla 6.9: Estimación del rebaño (e^{ν}) junto con los I.C.

Rebaño	*Estimación de* e^{ν}	*I.C.(95 %) de* e^{ν}	
		Lim. Inf.	*Lim. Sup.*
1	1.742	0.987	3.077
2	1.842	1.039	3.264
3	1.040	0.581	1.862
4	0.440	0.248	0.783
5	0.552	0.325	0.937
6	1.651	0.944	2.885
7	0.723	0.425	1.232
8	0.479	0.278	0.824
9	0.357	0.206	0.618
10	0.554	0.304	1.008
11	2.435	1.354	4.378
12	1.239	0.665	2.311
13	1.364	0.760	2.449
14	1.908	1.118	3.255
15	1.389	0.873	2.210

a este rebaño conlleva tan sólo un 44 % del riesgo que tendría pertenecer a un rebaño de referencia.

2. Las vacas de los rebaños 2, 11 y 14 tienen un mayor riesgo de padecer mastitis. Por ejemplo para el rebaño 11, con e^{ν}=2.435 significa que la pertenencia a este rebaño es más del doble que el riesgo que tendría pertenecer a un rebaño de referencia.

3. El resto de los rebaños, es decir 1, 3, 6, 7, 18, 12, 13 y 15 el efecto rebaño no fue estadísticamente significativo.

Con estos resultados, indican la importancia de haber introducido el efecto rebaño como término de frailty. En la práctica seria interesante, para la mejora productiva, considerar las características relacionadas a los rebaños 4, 5, 8 y 9, debido a que son los rebaños que contribuyen a disminuir el riesgo de padecer mastitis en estas vacas de producción de leche.

Modelo de Cox con uno o dos términos de frailty

Se presentan a continuación los resultados obtenidos con el paquete *frailtyHL*. Este paquete supone que el tiempo es continuo, utiliza la técnica de empates Breslow y permite las distribuciones normal o gamma para el término frailty. Este paquete permite también el ajuste del modelo sin término de frailty con las opciones: `vrfixed=TRUE` y `varinit=0` para la función `frailtyHL` (ver Ha et al. (2012)).

Dado que las vacas del estudio pertenecen a uno de los 15 rebaños considerados y, a su vez, cada rebaño pertenece a una sola de las tres zonas geográficas (Mérida, Trujillo y Zulia), en los presentes análisis fueron introducidos dos efectos aleatorios, zona geográfica y rebaño, donde en forma natural, presentan esta estructura anidada. Así pues, se ajustó un modelo de Cox con tres covariables y uno o dos términos de frailty con distribución normal, mediante el paquete *frailtyHL* del software R.

Los resultados se muestran en la Tabla 6.10, done el efecto del tipo de ordeño es estadísticamente significativo igual como se obtuvo en los modelos discretos logit y clog-log con frailty, analizados en la subseccción anterior. Por otra parte, las variables suplemento y enfermedad preexistente no fueron estadísticamente significativas. Además, la estimación de la desviación típica del término de frailty correspondiente a la variable rebaño disminuyó al considerar un segundo término de frailty (zona geográfica). Además la variabilidad del efecto de la zona geográfica es mayor que la de rebaño.

Tabla 6.10: Estimaciones de los coeficientes de los parámetros y errores estándar (EE), para efectos fijos y efectos aleatorios, con un y dos érminos de frailty

Covariables	*Modelo 1* *Coeficiente (EE)*		*Modelo 2* *Coeficiente (EE)*		*Modelo 3* *Coeficiente (EE)*		*Modelo 4* *Coeficiente (EE)*	
EnfermedadP.2	0.064 (0.123)		0.108 (0.125)		0.106 (0.124)		0.113 (0.125)	
T. ordeno2	-0.571 (0.162)	***	-0.549 (0.169)	**	-0.538 (0.164)	***	-0.538 (0.168)	**
Suplemento2	0.051 (0.125)		0.113 (0.131)		0.223 (0.129)	•	0.156 (0.131)	
frailty			*Dev Est (EE)*		*Dev Est (EE)*		*Dev Est (EE)*	
Rebaño			0.488 (0.115)				0.326 (0.069)	
ZonaG.					0.469 (0.233)		0.458 (0.249)	
Criterio de selección								
AIC	2986.7		2943.0		2955.3		2941.4	
Deviance	2987.3		2955.3		2959.9		2951.0	

Nota: Los Modelos han sido ajustados usando la función *frailtyHL*, que usa el método breslow.
El Modelo 1 incluye solo efectos fijos, el modelo 2 y 3 incluyen un solo efecto aleatorio (rebaño o zona geográfica)
y el Modelo 4 incluye los dos términos aleatorios anidados: rebaño y zona geográfica.
Para *** 0,1 %; ** 1 %; *5 % y • 10 % de significación estadística.

Evaluación del término frailty, usando estadísticos de bondad de ajuste

El Criterio de Información de Akaike (AIC) y la Deviance, permiten la selección de modelos anidados. Estos estadísticos están incluídos en la Tabla 6.10. A partir de los resultados obtenidos, se puede indicar que el Modelo 4, es el que presenta menores valores de los dos estadísticos por lo tanto puede ser considerado el que mejor ajusta al conjunto de nuestros datos.

Para contrastar la hipótesis de ausencia del efecto del término de frailty, es decir contrastar si $\sigma = 0$, se utilizó la diferencia de estimaciones de los estadísticos Deviance (ver Tabla 6.10) entre los modelos correspondientes. Esta diferencia tiene una distribución asintótica dada por la mixtura χ^2_0 y χ^2_1 (ver Ha et al. (2012)) que tiene percentil 95 igual a 2.71.

Por ejemplo, para el efecto rebaño, la diferencia de Deviance entre el modelo de Cox sin frailty (Modelo 1) y el modelo de Cox con frailty (Modelo 2) seria: 2987.3 - 2955.3=32.0, lo que permite concluir que el efecto del rebaño es estadísticamente significativo. Paralelamente, para el efecto zona geográfica una vez se considera el efecto rebaño, la diferencia de Deviance entre los modelos 2 y 4 seria: 2955.3 - 2951.0=4.3 lo que permite concluir que añadir la zona geográfica además del efecto del rebaño, es estadísticamente significativa. De la misma forma se puede concluir que añadir el efecto del rebaño además de la zona geográfica, es estadísticamente significativo. De acuerdo a los resultados obtenidos, los dos efectos aleatorios están influyendo sobre el riesgo de padecer mastitis.

Capítulo 7

COMPARACIONES DE SOFTWARE PARA ABORDAR TIEMPOS DISCRETOS. Aplicación a un estudio sobre lactancia en vacas lecheras.

7.1. Introducción

Existen varios software estadísticos que se usan para ajustar modelos en análisis de supervivencia, en este estudio se han utilizado el Stata (v 11.2, StatCorp LP, Texas, USA), SAS (v 9.3, SAS Institute inc, Cary, NC, USA) y R (v 2.15.1, R Development Core Team, Vienna, Austria). La gran versatilidad de los procedimientos estadísticos disponibles para el análisis de supervivencia en tiempos discreto, hace necesario comparar los procedimientos incluidos en los software de interés, especialmente para modelos de análisis de supervivencia que cuentan pocas décadas de aplicación. Uno de los objetivos de esta tesis es comparar algunas funciones utilizadas en software estadísticos para ajustar modelos discretos con y sin frailty, utilizando un conjunto de datos referidos al primer

diagnóstico de mastitis en vacas productoras de leche (presentados en el capítulo 6), donde se han estimando las covariables fijas y uno o dos términos de heterogeneidad no observada (frailty).

Este capítulo se compone de una sección con el resumen de los procedimientos de software estadístico que se han utilizado en los capítulos 5 y 6 y otra donde se comparan los resultados obtenidos mediante los diferentes software para el análisis del número de lactancias hasta el diagnóstico de mastitis en vacas lecheras.

7.2. Software estadísticos utilizados

El software estadístico utilizado en este capítulo se muestran en la tabla 7.1. La descripción técnica de estos procedimientos o funciones se encuentran en los capítulos 3 y 4.

Tabla 7.1: Software estadísticos para modelos de análisis de supervivencia para tiempo discreto con y sin términos frailty.

Modelos/ Software	*R*	*Stata***	*SAS***
I: Modelos de Cox con empates, sin frailty	coxph: efron, breslow, exact	stcox: efron, breslow, exactp, exactm	proc phreg: efron, breslow, discrete, exact
II: Modelos en tiempos discretos sin frailty	glme: link=logit, cloglog	logit, clog-log	proc logistic: link=logit, cloglog proc genmod
III: Modelos de Cox con empates y frailty	coxph coxme frailtyHL	stcox*	proc phreg
IV: Modelos en tiempos discretos con frailty	lme4: glmer	xtlogit xtcloglog	proc glimmix: random
V: Modelos de Cox dos términos de frailty	frailtyHL	—	—
VI: Modelos discretos dos términos de frailty	lme4: glmer	gllamm logit, cloglog	proc glimmix random

Nota: $*$ No fue usado en este estudio, porque asume una distribución gamma para el término frailty. $**$ En el software SAS el método Exact es llamado discrete y el método Average es llamado exact. En el software Stata el método Exact es llamado exactp y el método Average es llamado exactm.

7.3. Comparación de las funciones para ajustar modelos en tiempo discreto

Para las estimaciones de los parámetros de los modelos planteados, se consideraron funciones disponibles en los software estadísticos R, Stata y SAS. Es importante considerar que los modelos logit y clog-log con y sin términos de frailty, pueden ser analizados utilizando los software para los modelos de variable respuesta binaria con y sin factores aleatorios (Rabe-Hesketh et al., 2001). Por otra parte se puede apreciar que el software R no incluye ninguna función para la estimación de los parámetros para el caso del modelo de Cox con tratamiento de empates Average. Adicionalmente, la función mostrada del software Stata para el caso del modelo de Cox con tratamientos de empates e incluyendo el término de frailty, sólo utiliza la distribución log-gamma para el término frailty.

En la Tabla 7.2, se muestran las estimaciones de los parámetros y errores estándar para el modelo de Cox con los diferentes tratamientos de empates y los modelos para el número de lactancias hasta el diagnóstico de mastitis. En estos aún no se ha considerado ningún efecto aleatorio.

Tabla 7.2: Estimaciones de los parámetros β y sus errores estándar para modelos sin frailty

Covariables	*Efron*	*Breslow*	*Exact*	*Average*	*logit*	*clog-log*
	$\hat{\beta}$ (EE)	$\hat{\beta}$ (EE)	$\hat{\beta}$ (EE)	$\hat{\beta}$ (EE)	$\hat{\beta}$ (EE)	$\hat{\beta}$ (EE)
EnfermedadP.	0.063	0.064	0.080	0.064	0.080	0.063
	(0.123)	(0.123)	(0.145)	(0.125)	(0.146)	(0.116)
T. ordeño	-0.647	-0.571	-0.790	-0.670	-0.793	-0.672
	(0.162)	(0.162)	(0.194)	(0.166)	(0.194)	(0.166)
Suplemento	0.0818	0.051	0.057	0.082	0.057	0.084
	(0.125)	(0.125)	(0.147)	(0.127)	(0.147)	(0.127)

Los resultados que se muestran en las primeras cuatro columnas de la Tabla 7.2 corresponden al modelo de Cox con los cuatro tratamientos de empates y las dos últimas a los modelos en tiempos discretos logit y clog-log. A partir de la similitud entre los resultados obtenidos para las estimaciones de los coeficientes y los correspondientes errores

estándar, las metodologías utilizadas se pueden agrupar en:

Grupo I formado por el modelo de Cox con métodos de empates Efron y Average y el modelo discreto clog-log;

Grupo II formado por el modelo de Cox con método de empate Exact y el modelo logit;

Grupo III formado por el modelo de Cox con método de tratamiento de empates Breslow.

Estos grupos ya han sido puestas de manifiesto por otros autores como Singer & Willett (2003) así como en el capítulo 5 de esta misma tesis.

Comparación de las funciones para ajustar modelos en tiempo discreto con término de frailty

En esta sección se comparan las estimaciones de los parámetros y del efecto del término de frailty, aplicando diferentes funciones de software estadísticos para analizar los datos discretos de supervivencia. Las funciones utilizadas son: `coxph` (Breslow), `coxme`, `frailtyHL` del software R; `xtlogit, xtcloglog` y `gllamm` del software Stata y `proc phreg` y `proc glimmix` del software SAS.

En la Tabla 7.3 se presentan los resultados del estudio de diagnóstico de mastitis en vacas lecheras. Es importante resaltar que en estos resultados se han omitido las estimaciones de los indicadores de lactancia, ya que estos solo se generan para los modelos discretos logit y clog-log.

Considerando la similitud entre las estimaciones de los coeficientes y sus errores estándar, así como las estimaciones de los efectos aleatorios o término de frailty, se presentan la formación de los mismos grupos que en el caso de los modelos sin frailty. Además se pone de manifiesto la coherencia entre las distintos software cuando estiman el mismo modelo:

Grupo I formado por el modelo de Cox con método de empates Average y el modelo discreto clog-log: `glmer(cloglog), phreg(exact), xtcloglog, gllamm(cloglog)`. Estos presentan una gran proximidad de sus resultados tanto para los coeficientes de efectos fijos como para los términos frailty

y su variabilidad.

Grupo II formado por el modelo de Cox con método de empate Exact y el modelo logit: `coxme`, `glmer(logit)`, `phreg(discrete)`, `xtlogit`, `gllamm(logit)`. Estos presentan una gran proximidad de sus resultados tanto para los coeficientes de efectos fijos como para los términos frailty y su variabilidad.

Grupo III formado por el modelo de Cox con método de tratamiento de empates Breslow: `coxph(frailty)`, `frailtyHL`, `proc phreg`. La función `coxph(frailty)` no proporciona la estimación del error estándar para el término de frailty.

Tabla 7.3: Estimaciones de los parámetros y sus errores estándar para modelos con frailty en diferentes software

		R		SAS*		Stata*	
Covariables	*coxme*	glmer(logit)	glmer(clog-log)	*Exact*	*Average*	*xtlogit*	*xtcloglog*
EnfermedadP.	0.133	0.134	0.135	0.133	0.135	0.134	0.135
	(0.126)	(0.152)	(0.128)	(0.151)	(0.129)	(0.152)	(0.129)
T. ordeño	-0.611	-0.780	-0.646	-0.769	-0.641	-0.780	-0.644
	(0.172)	(0.211)	(0.178)	(0.210)	(0.178)	(0.212)	(0.178)
Suplemento	0.158	0.152	0.164	0.151	0.164	0.152	0.164
	(0.132)	(0.157)	(0.134)	(0.156)	(0.136)	(0.157)	(0.136)
frailty							
Rebaño	0.619	0.716	0.644	0.738	0.669	0.717	0.644
	(—)	(—)	(—)	(0.239)	(0.193)	(0.160)	(0.141)
		R		SAS	R con glmer		Stata con gllamm
Covariables	*coxph(frailty)*	*frailtyHL*		*proc phreg*	*logit*	*clog-log*	logit	clog-log
EnfermedadP.	0.108	0.108		0.108	0.138	0.137	0.138	0.137
	(0.126)	(0.125)		(0.126)	(0.152)	(0.128)	(0.152)	(0.129)
T. ordeño	-0.549	-0.549		-0.549	-0.771	-0.642	-0.778	-0,639
	(0.170)	(0.169)		(0.169)	(0.210)	(0.177)	(0.212)	(0.177)
Suplemento	0.112	0.113		0.113	0.184	0.200	0.184	0.200
	(0.131)	(0.131)		(0.131)	(0.157)	(0.134)	(0.158)	(0.137)
frailty								
Rebaño	0.471	0.488		0.489	0.520	0.460	0.521	0.460
	(—)	(0.115)		(0.113)	(—)	(—)	(0.153)	(0.116)
ZonaG.					0.522	0.480	0.522	0.479
					(—)	(—)	(0.296)	(0.243)

Nota $*$ En el software SAS el método Exact es llamado discrete y el método Average es llamado exact.
En el software Stata el método Exact es llamado exactp y el método Average es llamado exactm.

Capítulo 8

DISCUSIÓN Y CONCLUSIONES

8.1. Discusión

Para abordar tiempos discretos fueron considerados modelos de Cox con tratamientos de empates y modelos para una variable respuesta binaria logit y clog-log. Adicionalmente, se han estudiado los mismos modelos con uno o dos términos de frailty. También se han comparado diferentes utilidades del software disponible para ajustar estos modelos.

En análisis de supervivencia, el modelo semi-parámetrico de Cox goza de gran popularidad, posiblemente esta sea la causa por la que la mayoría de los trabajos lo utilizan. Sin embargo, la ventaja de tomar en cuenta la naturaleza del tiempo discreto ya sea si el tiempo es estrictamente discreto o cuando presenta un gran número de empates, es evitar posibles sesgos en los efectos de las covariables. Por lo tanto los métodos de análisis de supervivencia con tiempo continuo, son menos satisfactorios y hay que utilizar modelos de supervivencia con tiempo discreto (Rabe-Hesketh et al., 2001).

En esta tesis doctoral, fue considerado el modelo de Riesgos Proporcionales de Cox, para datos agrupados, siguiendo la notación de Kalbfleisch & Prentice (1973) y las propuestas para datos discretos de Singer & Willett (2003) y Grilli (2005). Así, los métodos utilizados para abordar tiempos discretos fueron los modelos de Cox con diferentes tratamientos de empates, y los modelos para variable respuesta binaria con enlaces logit y clog-log, con la finalidad de estimar los coeficientes de las covariables involucradas en el estudio

y obtener la correspondiente estimación de la función de riesgo. Otros autores (Rabe-Hesketh et al., 2001; Singer & Willett, 2003) han desarrollado y aplicado estos modelos en el campo de las ciencias sociales.

Al igual que en otros estudios, nosotros hemos ajustado modelos discretos logit y clog-log utilizando métodos estándar para respuestas dicotómicas, pero como paso inicial se requiere expandir la matriz de datos originales a una matriz individuo-periodo. La obtención de ésta expansión de la matriz puede parecer un inconveniente, pero no lo es, ya que puede ser obtenida a través de un código simple.

En los primeros capítulos de esta tesis se presentan las metodologías del campo del análisis de la supervivencia necesarias para llevar a cabo los estudios aplicados que se presentan en los próximos capítulos.

En el capitulo 5 se presenta una aplicación de estos modelos estudiados a un conjunto de datos reales en caballos de carreras de la raza pura sangre, donde se consideró la naturaleza discreta de la variable tiempo, definida como la edad en años en que se presentó la lesión musculoesquelética catastrófica (CMI). En este estudio, los modelos discretos logit y clog-log fueron utilizados para identificar factores asociados con el riesgo de que el caballo padezca CMI durante una carrera. A partir del modelo logit, se obtienen odds proporcionales mientras que el modelo clog-log tiene la ventaja que permite la interpretación de las estimaciones de los parámetros como un cociente de riesgos (HR) (Allison, 2010; Singer & Willett, 2003), concepto muy usado en análisis de supervivencia, lo cual fue muy atractivo en la interpretación de nuestros resultados. Al aplicar estos modelos a nuestros datos, las covariables estadísticamente significativas fueron: la lesión preexistente, número de carreras y longitud de carrera. Se consideraron también el género y la época dado su interés desde el punto de vista veterinario. Adicionalmente, el modelo de Cox con distintos tratamientos de empates permitió también obtener estimaciones en términos de riesgos proporcionales, llegando a las mismas conclusiones con respecto a la identificación de los factores de riesgos estadísticamente significativos que en los modelos discretos.

Los resultados para la estimación de los coeficientes así como para los errores estándar son muy similares entre el modelo de Cox con tratamiento de empates Exact y el modelo logit; lo mismo resultó para el modelo de Cox con tratamientos de empates Average y

el modelo discreto clog-log. Tal como se dice en el capítulo 3, si se asume un modelo de Cox para la variable tiempo continua, el correspondiente modelo clog-log para los datos agrupados contiene los mismos coeficientes β; por tanto no es de extrañar que las estimaciones $\hat{\beta}$ sean similares. Por otro lado se obtuvieron resultados muy parecidos entre los métodos Efron y Average, lo que también ha sido reportada por Therneau & Grambsch (2000).

A partir de los resultados de este estudio, se pone de relieve que los modelos discretos logit y clog-log no sólo son útiles para abordar datos discretos sinó que son una alternativa cuando los datos vienen agrupados. Además se muestran dos fortalezas de estos modelos: la primera es que permiten obtener fácilmente la función de riesgo basal estimada a lo largo de los diferentes periodos de tiempos involucrados; la segunda es la rapidez en la obtención de resultados en comparación a los métodos Exact y Average para modelo de Cox.

En el capítulo 6, se han ajustado modelos de supervivencia para tiempos discretos con uno o dos término de frailty. Este modelo se aplicó para analizar datos de una muestra de vacas en producción de leche, donde el tiempo discreto considerado fue definido como el número de lactancias hasta el primer diagnóstico de mastitis. Los resultados obtenidos nos permitieron identificar, por una parte, que el tipo de ordeño es un factor de riesgo en padecer mastitis y, por otra parte, los resultados permitieron comparar las similitudes en los diferentes software utilizados. En todos los análisis se ha asumido una distribución normal para el término de frailty.

También se ha realizado una comparación de los resultados obtenidos por tres software: R, Stata y SAS.

Para el modelo de Cox con un término frailty, el software R con las funciones frailtyHL y coxph (frailty) con Breslow para el tratamiento de empates, sirven para estimar el mismo modelo que con el proc phreg (random) de SAS.

Las funciones xtlogit y xtcloglog de Stata, la función glmer con link: logit y clog-log del paquete lme4 del R y el proc glimmix (binary random) de SAS sirven para estimar los modelos de variable respuesta binaria con factores aleatorios. Además dan estimaciones similares a las obtenidas con el proc phreg (random) y el tratamiento de empates Exact y Averge (en SAS denominados discrete y exact, respectivamente) correspondientes al

modelo de Cox con un término frailty que tiene en cuenta la naturaleza discreta del tiempo.

Cuando el término de frailty corresponde al efecto grupo, una ventaja de utilizar el Stata es que aporta el coeficiente de correlación intraclase y los correspondientes intervalos de confianza para la varianza del término de frailty. Para este modelo, el SAS, permite obtener el efecto medio del factor aleatorio grupo.

8.2. Conclusiones

1. En datos donde el tiempo es estrictamente discreto o sus valores han sido agrupados, se ha de tener en cuenta al aplicar metodologías clásicas del análisis de la supervivencia.
2. Los modelos de tiempo discretos tienen en cuenta que el riesgo es una probabilidad, lo cuál es una ventaja al momento de interpretar resultados, por parte de los especialistas veterinarios y otras áreas de prácticas profesionales.
3. Con el uso de la matriz de datos expandida, es posible utilizar los métodos para modelos lineales generalizados y de fácil aplicación con cualquier software estadístico.
4. El ajuste de un modelo de Cox con tratamiento de empates Exact y Average proporciona estimaciones similares a las obtenidas al ajustar los modelos logit y clog-log, respectivamente.
5. Los modelos discretos logit y clog-log facilitan la obtención de la función de riesgo basal estimada. Lo que puede ser de utilidad en este tipo de análisis.
6. Cuando la variable tiempo considerada en el estudio es verdaderamente continua, pero los valores han sido agrupados, es recomendable utilizar el modelo discreto clog-log. Mientras en el caso en que el tiempo es estrictamente discreto, el uso del modelo logit es el más adecuado.
7. En el caso en que se tenga heterogeneidad no observada, recomendamos incluir un término de frailty como efecto aleatorio del modelo de análisis de supervivencia para evitar el sesgo de las estimaciones.
8. En presencia de frailty los modelos logit y clog-log pueden ser ajustados mediante los procedimientos estándar para modelos lineales generalizados con factores aleatorios que presentan los respectivos software estadísticos.
9. La función frailtyHL del software R, permite ajustar modelos con uno o dos términos de frailty.
10. La mayor parte de los estudios desarrollados en los temas de análisis de supervi-

vencia con aplicación al área de la veterinaria utilizan el modelo de Cox. Recomendamos tener en cuenta la posible naturaleza discreta del tiempo cuando sea el caso.

Capítulo 9

TRABAJOS FUTUROS APLICADOS AL ÁREA DE VETERINARIA

En virtud del desarrollo de esta tesis, de análisis de superviencia en tiempo discreto y en presencia de términos de frailty, con aplicación en el área de veterinaria, se plantean cuatro líneas básicas de investigación futura multidisciplinaria:

1. El efecto de la sobreexpresión de CDK4 sobre la incidencia de diabetes en ratones. En las últimas décadas la diabetes tipo 1 está adquiriendo dimensiones epidémicas en el mundo desarrollado, lo que hace necesario el estudio de esta enfermedad. La CDK4 (kinasa dependiente de cíclica 4) se ha demostrado que tiene una función relevante en la generación de la masa beta pancreática productora de insulina. Por lo tanto es importante conocer el efecto de la sobreexpresión mediante transgénesis de dicha kinasa bien en su forma salvaje (WT) o hiperactiva (R24C), en las células beta del páncreas de la cepa NOD (Non obese diabetic) de ratones, la cual está predispuesta genéticamente a desarrollar diabetes autoinmune o de tipo 1 (T1D). El estudio consta de una muestra de 125 ratones hembras, de la misma edad. La medida experimental consistió en determinar el momento en el cual los ratones resultan positivos a la enfermedad. Dentro del estudio son de interés dos covariables relacionadas con las características propias de estos animales, como es

el genotipo de la cepa en relación al tipo de transgén incorporado en el genoma de las células beta-pancreáticas (5 líneas transgénicas diferentes) y sus correspondientes genotipos controles en los que el genoma no había incorporado el transgén, o transgénicos negativos. También, en los análisis se cuantificará el grado de infiltración leucocitaria de los islotes pancreáticos o de Langerhans. Para ajustar el modelo de supervivencia, el tiempo será considerado como una variable discreta y definido como el número de semanas transcurridas hasta que los ratones sean diagnosticados con diabetes tipo 1. Aquellos ratones que al final de estudio, no desarrollen diabetes o mueran por cualquier causa diferente al evento estudiado, sus tiempos serán considerados observaciones censuradas. Dada la naturaleza discreta del tiempo, este problema se abordará mediante los modelos par datos de supervivencia discretos logit y clog-log.

2. Factores de riesgo que influyen en la longevidad de vacas de la raza Carora (Venezuela). En términos de mejorar el nivel de productividad lechera, es de interés determinar los factores de riesgo que afectan la longevidad de estos animales. Se propone analizar unos datos provenientes de registros lecheros del Estado Cojedes en Venezuela. Se han registrado variables tales como: porcentaje racial, lote animal, padre, madre, año de parto, mes de parto, año de nacimiento, días de lactancias, litros de producción de leche, intervalo parto-destete, número de lactancias (variable tiempo) y edad, durante los periodos 1995 a 2015. En el ajuste de los datos a un modelo de análisis de supervivencia, será considerado el tiempo discreto como el tiempo desde la primera lactancia hasta la última lactancia en que la vaca este productiva, también, el tiempo sera medido como el número de lactancias hasta que ocurre su descarte. Por el contrario todas aquellas vacas que se encuentren en lactancia, al momento del cierre del estudio, su tiempo será considerado censurado. Como la variable tiempo considerada es discreta, los modelos que se propondrán son los modelos discretos logit y clog-log con covariables dependientes del tiempo.

3. Deformación del borde dorsal del cuello en équidos. La deformación del borde dorsal del cuello representa un problema que afecta a la industria hípica en España, con repercusiones clínicas como ataxia, pérdida de movilidad del cuello y claudicaciones entre otras. En la actualidad un grupo de veterinarios de la Universidad de Córdoba, está trabajando en esta área, mediante una evaluación clínica, mor-

fológica y toma de muestras del cuello en équidos mediante biopsias, estudio histopatológico e inmunohistoquimico, pertenecientes a las Comunidades Autónomas de Andalucía y Extremadura, en un total de 627 equinos y 373 burros. Entre las variables registradas se encuentran: identificación del individuo, edad, raza, sexo, pelaje, localización geográfica (Andalucía-Extremadura), grado de deformación del cuello, diámetro del cuello, longitud del cuello, peso, actividad atlética, alimentación, genética, evaluación clínica, caracterización histológica y biopsia. El objetivo de esta futura investigación será determinar los factores de riesgo asociados al desarrollo de la deformación del borde dorsal en équidos, así como su evolución en el tiempo. En el análisis estadístico, la variable tiempo a considerar, es el tiempo (en años) hasta que ocurre el diagnostico de esta patología. Al ser el tiempo medido en años, los modelos discretos con términos de frailty son adecuados.

4. Síndrome Ulceroso Gástrico Equino. Actualmente se están desarrollando unos ensayos clínicos en la zona de Extremadura y Córdoba en España desde el año 2012, con fecha de cierre del estudio a finales del 2016, con un total 500 estómagos de caballo. La información se registra en base a variables como: identificación individual, edad, raza, sexo, pelaje, localización geográfica (Andalucía-Extremadura), grado de inflamación y ulceración gástrica, peso, actividad atlética, alimentación, genética, evaluación clínica, caracterización histológica y biopsia. El objetivo del futuro estudio será identificar los factores de riesgo asociados al desarrollo del Síndrome Ulceroso Gástrico Equino (EGUS). En el ajuste de los datos a un modelo de análisis de supervivencia, será considerado y el tiempo discreto es definido como el tiempo (en años) hasta que se presente esta enfermedad y el animal sea descartado. Igual que antes, al ser el tiempo medido en años, se propondrán los modelos discretos con términos de frailty.

5. Muerte súbita en caballos de carreras. Estudio realizado en Venezuela, en caballos de carreras de la raza pura sangre. La muerte súbita es un problema que afecta a la industria hípica a nivel mundial. Durante un periodo entre 2000 al 2012, se registraron datos mediante seguimiento de los caballos que participan en las carreras del hipódromo La Rinconada en Caracas (Venezuela). El tiempo será definido como la edad y el número de carreras en que ha participado el caballo hasta experimentar muerte súbita y el objetivo es determinar los factores de riesgo que determinan este

evento. El modelo que se propondrá en este caso, dado que la variable repuesta es discreta (número de carreras), es el modelo logit con covariables dependientes del tiempo.

Bibliografía

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View.* Statistics for Biology and Health. Springer.

Allison, P. D. (2010). *Survival Analysis using SAS. A practical guide.* SAS Institute Inc.

Barber, J., Murphy, S., Axinn, W., & Maples, J. (2000). Discrete-time multilevel hazard analysis. *Sociological Methodology*, 30, 201–235.

Boden, L., Anderson, G., Charles, J., Morgan, K., Morton, J., Parkin, T., Slocombe, R., & Clarke, A. (2006). Risk of fatality and causes of death of Thoroughbred horses associated with racing in Victoria Australia between 1989 and 2004. *Equine Vet. J.*, 38, 312–318.

Casellas, J., Caja, G., Such, X., & Piedrafita, J. (2007). Survival analysis from birth to slaughter of Ripollesa lambs under semi-intensive management. *J. Anim. Sci.*, 85, 512–517.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. *Biometrika*, 65, 141–151.

Clayton, D. G. (1991). A monte carlo method for bayesian inference in frailty models. *Biometrics*, 47, 467–485.

Collett, D. (2003). *Modelling survival data in medical research.* Chapman-Hall, USA.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society.*, 34, 187–220.

Duchateau, L. & Janssen, P. (2008). *The Frailty Model.* New York: Springer.

Ducrocq, V. & Casella, G. (1996). A bayesian analysis of mixed survival models. *Genet. Sel. Evol.*, 28, 505–529.

Evans, S. & Sayers, A. (2000). A longitudinal study of campylobacter infection of broiler flocks in Great Britain. *Preventive Veterinary Medicine*, 46, 209–223.

Famula, T. (1981). Exponential stayability model with censoring and covariates. *J. Dairy Sci.*, 64, 538–545.

Gómez, G. (2005). *Análisis de Supervivencia*. Barcelona: Impreso por Ahlens S.L.

Grilli, L. (2005). The random-effects proportional hazards model with grouped survival data: A comparison between the grouped continuous and continuation ratio versions. *Journal of the Royal Statistical Society Series A*, 168(1), 83–94.

Guiraudo, A., Calsolari, A., Romponr, A., Bogni, C., Larriestra, A., & Angel, R. (1997). Field trials of vaccine against bovine mastitis. *Journal Dairy Science.*, 80, 845–853.

Guo, G. & Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in guatemala. *Journal of the American Statistical Association*, 87, 969–976.

Ha, I., Lee, I., & MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine*, 26, 4790–4807.

Ha, I., Noh, M., & Lee, Y. (2012). FrailtyHL: a package for fitting frailty models with H-likelihood. *The R Journal*, 4, 28–36.

Han, A. & Hausman, J. (1990). Flexible parametric estimation of duration and competing risk model. *Journal of Applied Econometrics*, 5, 1–28.

Hedeker, D., Siddiqui, O., & Hu, F. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in medical Research*, 9, 161–179.

Hougaard, P. (2000). *Analysis of multivariate survival data.* New York: Springer Verlag.

Hudson, C., Huxley, J., & Green, M. (2014). Using simulation to interpret a discrete time survival model in a Complex Biological System: Fertility and Lameness in Dairy Cows. *Plos One*, 9, 1–7.

Johnson, B., Stover, S., Daft, B., Kinde, H., Read, D., Barr, B., Anderson, M., Moore, J., Woods, L., Stoltz, J., & Blanchard, P. (1994). Causes of death in racehorses over a 2 year period. *Equine Vet. J.*, 26, 327–330.

Kalbfleisch, J. D. & Prentice, L. R. (1973). Flexible parametric estimation of duration and competing risk model. *Biometrika*, 13(2), 267–278.

Kalbfleisch, J. D. & Prentice, L. R. (2002). *The Statistical Analysis of Failure Time Data.* Wiley Inter-Science. Wiley series in probability and statistics. New Jersey: John Wiley & Sons.

Klein, J. & Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* Statistics for Biology and Health. Springer.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics*, 48, 795–806.

Korsgaard, I., Madsen, P., & Jensen, J. (1998). Bayesian inference in the semi-parametric log normal frailty model using Gibbs sampling. *Genet. Sel. Evol.*, 30, 241–256.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939–956.

Legrand, C., Ducrocq, V., Janssen, P. Silvester, R., & Duchateau, L. (2005). A bayesian approach to jointly estimate centre and treatment by centre heterogeneity in a proportional hazards model. *Statistics in Medicine*, 24, 3789–3804.

McCullagh, P. (1980). Regression models for ordinal data (methodological). *Journal of the Royal Statistical Society. Series B*, 42, 109–142.

Meyer, B. (1990). Unemployment insurance and unemployment spell. *Econometrica*, 58, 757–782.

Mohammed, H., Hill, T., & Lowe, J. (1991). Risk factors associated with injuries in Thoroughbred horses. *Equine Vet. J.*, 23, 445–448.

Nicoletti, C. & Rondinelli, C. (2010). The (mis)specification of discrete duration models with unobserved heterogeneity:A Monte Carlo study. *Journal Econometrics*, 159, 1–13.

Paccagnella, O. (2006). Comparing vocational training courses through a discrete-time multilevel hazard model. *Statistical Modelling*, 6, 119.

Parkin, T., Clegg, P., French, N., Proudman, C., Riggs, C., Singer, E., Webbon, P., & Morgan, K. (2004). Risk of fatal distal limb fractures among thoroughbreds involved in the five types of racing in the United Kingdom. *Vet. Rec.*, 154, 493–497.

Peloso, J., Mundy, G., & Cohen, N. (1994). Prevalence of, and factors associated with, musculoskeletal racing injuries of Thoroughbred. *J. Am. Vet. Med. Ass*, 204, 620–626.

Popkowski, P. & Bass, F. M. (1998). Determinig the efects of observed and unobserved heterogeneity on consumer brand choice. *Appl. Stochastic Models Data Anal.*, 14, 95–115.

Prentice, R. L. & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34, 57–67.

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata.* STATA press.

Rabe-Hesketh, S., Yang, S., & Pickles, A. (2001). Multilevel models for censored and latent responses. *Statistical Methods in Medical Research*, 10(6), 409–427.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163.

Ripatti, S. & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56, 1016–1022.

Scheike, T. H. & Jensen, T. (1997). A discrete survival model with random effects: An application to time to pregnancy. *Biometrics*, 53, 318–329.

Singer, J. D. & Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 15(2), 1536–1547.

Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence.* Oxford University Press, New York.

Sölkner, J. & Ducrocq, V. (1996). The Survival Kit: a tool for analysis of survival data. *Genet. Sel. Evol.*, 28, 505–529.

Southey, B. R., Rodriguez-Zas, S., & Leymaster, K. (2003). Discrete time survival analysis of lamb mortality in a terminal sire composite population. *Journal of animal Science*, 81, 1399–1405.

Tarrés, J., Casellas, J., & Piedrafita, J. (2005). Bayesian inference in the semi-parametric log normal frailty model using Gibbs sampling. *Journal of Animal Science*, 83, 543–551.

Therneau, T. M. & Grambsch, P. (2000). *Modeling Survival Data. Extending the Cox Model.* Statistics for Biology and Health. Springer.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), 439–454.

Wienke, A., Holm, N., Christensen, K., Skytthe, A., Vaupel, J., & Yashin, A. (2003). The heritability of cause-specific mortality: a correlated gamma-frailty model applied to mortality due to respiratory diseases in danish twins born. *Stat. Med.*, 22, 3873–3887.

Yau, K. & McGilchrist, C. (1997). Use of generalized linear mixed models for the analysis of clustered survival data. *Biometrical Journal*, 39, 3–11.