**Gabriela Boggio**
**Leticia Hachuel**
**Guillermina Harvey**
**Julia Angelini**
**Brenda Niccolai**
*Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.*

# USO DE MODELOS LINEALES GENERALIZADOS PARA EL ANÁLISIS DE TIEMPOS HASTA EL EVENTO AGRUPADOS. UNA APLICACIÓN EN EL ESTUDIO DE LA DEMORA EN LA ELECCIÓN DE LA CARRERA UNIVERSITARIA

**Resumen:**
En estudios donde la respuesta de interés es el tiempo hasta la ocurrencia de un evento particular es común que los datos disponibles resuman lo acontecido durante un intervalo de tiempo específico. El análisis de este tipo de datos de duración se puede abordar a través del ajuste de ciertos modelos lineales generalizados. En este trabajo se consideran dos enfoques alternativos según se modelen probabilidades acumuladas o probabilidades condicionales considerando el tiempo como una variable discreta. La función de enlace utilizada permite interpretar los coeficientes de los modelos en términos de razones de hazards tal como ocurre en los análisis clásicos de supervivencia para tiempos continuos. Se ilustra esta metodología en un estudio sobre la demora de los estudiantes de la Facultad de Ciencias Económicas y Estadística de la UNR en la posibilidad de elegir la carrera cuando está estipulado un ciclo inicial común a las carreras de Ciencias Económicas. Se pudieron identificar las características de los estudiantes que favorecen o dificultan la elección de la carrera definitiva.

Palabras claves: censura por intervalos, razón de hazards, modelos lineales generalizados

**Abstract:**
In studies where the response of interest is the time to some event, the available data commonly summarizes what happened over a specific time interval. Generalized linear models can be used to analyze such duration data. In this paper two alternative approaches are considered. The first models cumulative probabilities and the second approach, conditional probabilities considering the time as a discrete variable. The link function used allows to interpret the regression coefficients in terms of hazard ratios as in the classic survival analysis for continuous time. This methodology is illustrated in a study about the student delay in the Facultad de Ciencias Económicas y Estadística – UNR in order to choose the career after a common initial cycle. It was possible to identify patterns of students that increase or reduce this time.

Keywords: interval censoring, hazard ratio, generalized linear models

## 1. Introducción

En el análisis de datos de supervivencia la variable respuesta es el tiempo que transcurre desde un origen determinado hasta la ocurrencia de un evento de interés. En general este tiempo se asume en escala continua y se supone que el evento en estudio puede ocurrir en cualquier punto de ese tiempo. Sin embargo puede suceder que los datos disponibles resuman lo ocurrido durante un intervalo de tiempo específico; de ahí que se hable de datos de supervivencia agrupados o con censura por intervalos pudiendo utilizarse los números enteros positivos para denotar estos tiempos.
En principio los enfoques tradicionales de modelización para tiempos continuos se pueden aplicar también a este tipo de datos agrupados, pero además son factibles de modelar dentro del marco de los modelos lineales generalizados (MLG).
Específicamente en este trabajo se consideran dos enfoques bajo dicho marco, a saber:

- Tratamiento del tiempo de supervivencia como una variable ordinal la cual puede ser censurada a la derecha o no.
- Tratamiento del tiempo de supervivencia como un conjunto de variables dicotómicas que indican si el evento ocurrió o no en determinados períodos de tiempo hasta el momento de ocurrencia del evento o censura.

Se ilustran los alcances y diferencias de estos enfoques de modelización mediante su aplicación a datos sobre rendimiento académico de alumnos de la Facultad de Ciencias Económicas y Estadística de la UNR.

## 2. Metodología

En un estudio de datos de supervivencia o duración agrupados se dispone de un resumen de lo que ha ocurrido durante un intervalo de tiempo. Formalmente el tiempo continuo se divide en intervalos:

$$[0, a_1), [a_1, a_2), \dots\dots, [a_q, \infty).$$

Se pueden utilizar números enteros positivos para denotar estos intervalos. Así, si se define la variable T: tiempo hasta la ocurrencia del evento de interés, $T = t \quad (t = 1,2, \dots q + 1)$ significa que el evento ha ocurrido en el intervalo $[a_{t-1}, a_t)$, también llamado período de tiempo $t$.
El proceso de modelización del comportamiento estocástico de esta variable T puede considerarse a través de:

$$P(T \leq t) \quad \text{ó} \quad P(T = t/T \geq t)$$

dando lugar a los dos enfoques que se presentan a continuación.

### 2.1. Enfoque ordinal

Bajo el supuesto de que el tiempo puede tomar sólo valores positivos discretos $t = 1,2,, \dots., q$, a cada individuo $i = 1, \dots., n$ se le asigna el tiempo $t_i$ en el cual ocurre el evento o bien la observación resulta censurada. La censura, elemento clave en el análisis de datos de supervivencia, indica en este caso que el i-ésimo individuo fue observado hasta $t_i$ pero no en $t_{i+1}$.

Se define $P_{it}$ como la probabilidad de que el i-ésimo individuo presente el evento antes o durante el intervalo de tiempo *t*, esto es:

$$P_{it} = Pr(T_i \leq t).$$

Por lo tanto, la probabilidad de supervivencia más allá del intervalo de tiempo t es simplemente $1 - P_{it}$ .

Debido a que $1 - P_{it}$ representa la función de supervivencia, McCullagh (1980) propuso la siguiente versión discretizada del conocido *modelo de hazards proporcionales para tiempo continuo* o más sencillamente *modelo de regresión de Cox*:

$$ln[-ln(1 - P_{it})] = \alpha_t + \boldsymbol{x'}_i\boldsymbol{\beta}.$$

Se trata de un MLG con enlace "log-log del complemento" para las probabilidades acumuladas de una respuesta ordinal.

En términos de esta probabilidad acumulada:

$$P_{it} = 1 - exp[-exp(\alpha_t + \boldsymbol{x'}_i\boldsymbol{\beta})].$$

En esta expresión el vector $\boldsymbol{x}_i$ incluye variables explicativas que no varían con el tiempo, es decir no varían a través de las categorías de respuesta ordenadas. Sin embargo pueden representar el promedio de una variable a través del tiempo o el valor de la covariable en el tiempo de ocurrencia del evento.

Debido a que $-ln(1 - P_{it})$ es la función hazard acumulada, $H(t)$, este modelo permite interpretar el efecto de cada covariable en términos de este concepto básico del análisis de supervivencia (Tutz&Schmid, 2015).

### 2.2. Enfoque binario

El segundo enfoque consiste en tratar cada tiempo de supervivencia individual como un conjunto de observaciones dicotómicas que incluyen variables indicadoras acerca de si un individuo presenta o no el evento en cada tiempo hasta que ese individuo experimente el evento o sea censurado.

Esta reestructuración de los datos es particularmente útil para manejar covariables dependientes del tiempo, es decir que el valor de esas covariables para un individuo particular cambia a través del tiempo. De esta manera es posible ajustar modelos que no verifiquen el supuesto de hazards proporcionales.

En este enfoque se define $p_{it}$ como la probabilidad de ocurrencia del evento en el intervalo de tiempo t condicional a que no se produjo antes de t:

$$p_{it} = P(T_i = t/T_i \geq t) \ .$$

Luego, $1 - p_{it}$ es la probabilidad condicional de sobrevivir más allá de $t$. Bajo este enfoque un MLG con enlace "log-log del complemento" es:

$$ln[-ln(1 - p_{it})] = \gamma_t + \boldsymbol{x'}_i\boldsymbol{\beta}.$$

En él, el intercepto dependiente del tiempo $\gamma_t$ representa un intercepto fijo más el efecto del tiempo asumido como una variable categórica a través de la consideración de $q - 1$ variables de diseño.

Nótese que la diferencia con la expresión del enfoque ordinal es que participan probabilidades condicionales y no probabilidades acumuladas. Comparando ambos enfoques, en el primero cada observación consiste en sólo dos partes: el tiempo en escala ordinal hasta la ocurrencia del evento y si hay censura o no. En cambio en el enfoque dicotómico cada tiempo está representado por un vector de variables indicadoras, donde el tamaño del vector depende del tiempo de ocurrencia del evento o censura. Por lo tanto, el enfoque ordinal es

más fácil de implementar y ofrece ventajas en términos de tamaño de la base de datos. El enfoque binario, en cambio, es superior en su tratamiento de covariables que dependan del tiempo.

Además el modelo correspondiente a este último enfoque se puede ajustar, luego de reestructurados los datos, a partir de cualquier software convencional mientras que el ajuste del modelo ordinal, si bien no requiere ningún trabajo adicional con respecto a la disposición de los datos, impone reprogramar el ajuste de los modelos para poder considerar la presencia de datos censurados.

Otro punto importante a destacar es que en ambos modelos, en ausencia de covariables dependientes del tiempo, los coeficientes $\beta'$s son idénticos a los del modelo de hazards proporcionales para la variable tiempo subyacente continua. Esto no significa que las estimaciones resulten idénticas sino que estiman a un mismo parámetro de diferente manera. Además el uso de este enlace "log-log del complemento" habilita a interpretar los coeficientes en términos de razones de hazards como en el modelo de Cox.

## 3. Aplicación

### 3.1. Los datos

El plan de estudios de los ingresantes de las carreras de Contador Público, Licenciatura en Economía y Licenciatura en Administración de la Facultad de Ciencias Económicas y Estadística de la UNR comienza con un Ciclo Introductorio común a las tres carreras. Los alumnos están en condiciones de elegir carrera a partir de la aprobación de cuatro de las asignaturas que conforman este Ciclo Introductorio. Si bien en los planes de estudios se considera que esta meta puede alcanzarse en el término de un año, es notorio el retraso en el cumplimiento de este requisito que habilitará a los alumnos a iniciar el cursado específico de cada carrera. Con el objetivo de determinar cuáles son los factores que pueden incidir sobre esta demora se analizan datos de los alumnos ingresantes en el año 2008 obtenidos por el Programa de seguimiento de Planes de Estudios que lleva adelante la Secretaría Académica de esta Facultad.
Concretamente se utiliza información registrada de 1723 alumnos inscriptos al Ciclo Introductorio común en el 2008 que han mostrado algún tipo de actividad académica durante el primer año en la Facultad o se han reinscripto al año siguiente. La información disponible comprende el seguimiento de su actividad académica hasta marzo de 2013. De esta forma es que la variable *Tiempo hasta cumplir las condiciones para poder elegir carrera* toma los valores 1, 2 3, 4 y 5 según la cantidad de años académicos que haya necesitado para cumplir los requisitos a partir del año 2008, momento de ingreso a la Facultad [(1)]. Como el seguimiento se realizó durante 5 años, los alumnos que no cumplieron los requisitos a marzo de 2013 se consideraron con tiempos censurados. Así mismo hubo censura en distintos puntos del tiempo, lo cual se reconoce como un problema con censura intermitente o generalizada. Cabe señalar que hubo situaciones particulares derivadas por ejemplo de períodos de pasividad seguidos de reinscripción, las cuales fueron analizadas en forma particular para la asignación del correspondiente tiempo.

A los fines de ilustrar la metodología presentada, del conjunto de variables registradas se seleccionan las siguientes:

- Sexo (masculino, femenino)
- Año de egreso de secundario (antes de 2007, año 2007)
- Categoría ocupacional del alumno (trabaja, no trabaja)
- Clase de colegio (público, privado)

- Estado ocupacional de la madre (trabaja, no trabaja)
- Nivel educativo del padre (1: hasta primaria incompleta; 2: hasta secundario incompleto; 3: hasta terciario incompleto; 4: hasta terciario completo o universitario incompleto; 5: universitario completo)
- Rendimiento académico (1: promedio de notas [(2)]<6; 3: 6≤promedio de notas<7,5; 4: promedio de notas≥7,5)
- Discapacidad (sí, no)

(1) Cabe aclararse que si bien los alumnos pueden elegir carrera en tres oportunidades durante el año académico, en este trabajo se utiliza el registro de elección de carrera que figura en la reinscripción anual.
(2) El promedio de notas consiste en el promedio de las notas obtenidas en los exámenes finales, incluyendo los aplazos, acumulado hasta la finalización de cada año académico.

### 3.2. Resultados

A partir de esta información se ajusta un modelo de efectos principales asociados a las variables antes definidas bajo ambos enfoques. Cabe señalar que con respecto a rendimiento académico se le asignó a cada individuo el valor que asume esta covariable en el tiempo de ocurrencia del evento para el ajuste del modelo ordinal y los valores asumidos en cada año académico para el ajuste del modelo bajo el enfoque binario aprovechando la posibilidad de incluir covariables dependientes del tiempo.
Todas las covariables resultaron significativas y las estimaciones de sus coeficientes se presentan en las Tablas 1 y 2.

Tabla 1: Estimaciones del modelo correspondiente al enfoque ordinal

| Variable | Estimación | Error estándar | Razón de hazards |
|---|---|---|---|
| Interceptos | | | |
| Año 1 | -1,443 | 0,141 | 0,236 |
| Año 2 | -1,030 | 0,138 | 0,357 |
| Año 3 | -0,807 | 0,138 | 0,446 |
| Año 4 | -0,694 | 0,138 | 0,499 |
| Año 5 | -0,630 | 0,139 | 0,533 |
| Sexo | | | |
| Masculino | -0,159 | 0,073 | 0,853 |
| Egreso del secundario | | | |
| Antes del 2007 | -0,626 | 0,083 | 0,535 |
| Clase de escuela secundaria | | | |
| Público | -0,196 | 0,070 | 0,822 |
| Categoría ocupacional del alumno | | | |
| Trabaja | 0,199 | 0,070 | 1,220 |
| Estado ocupacional de la madre | | | |
| Trabaja | 0,379 | 0,080 | 1,461 |
| Nivel educativo del padre | 0,112 | 0,033 | 1,118 |
| Discapacidad | | | |
| Si | -1,503 | 0,145 | 0,222 |
| Rendimiento académico | 0,574 | 0,032 | 1,776 |

Los resultados obtenidos muestran que los estudiantes que tienen mayor riesgo de demorar la posibilidad de elegir carrera son los de sexo masculino, los egresados de una escuela secundaria pública, los que han terminado la escolaridad secundaria antes de 2007, los que presentan algún tipo de discapacidad, los estudiantes que no trabajan, los que tienen un peor rendimiento académico, el nivel educativo del padre es bajo y cuyas madres no trabajan.
También se puede medir el efecto de un factor en particular. Así por ejemplo, la oportunidad de alcanzar la condición para poder elegir carrera en forma más temprana es alrededor de un 20% mayor cuando el alumno trabaja. Parecería que las obligaciones que impone la actividad laboral favorecen el cumplimiento de las actividades académicas. Así también, haber finalizado la escuela secundaria inmediatamente antes del ingreso a la Facultad aumenta al doble (~1/0.5) las oportunidades de poder elegir más tempranamente la carrera.
Hay que destacar que a pesar de estas coincidencias en las estimaciones de los coeficientes de las covariables en ambos modelos con enlace “log-log del complemento”, interpretables en términos de razones de hazards (Läära, Matthews, 1985), los valores de los interceptos no son equivalentes. Ello se debe a que cada uno de los modelos predice diferentes probabilidades, uno probabilidades condicionadas y el otro, probabilidades acumuladas.

Tabla 2: Estimaciones del modelo correspondiente al enfoque binario

| Variable | Estimación | Error estándar | Razón de hazards |
|---|---|---|---|
| Intercepto | -1,367 | 0,140 | 0,255 |
| Año | | | |
| Año 2 | -0,704 | 0,101 | 0,495 |
| Año 3 | -1,020 | 0,158 | 0,361 |
| Año 4 | -1,542 | 0,274 | 0,214 |
| Año 5 | -2,003 | 0,416 | 0,135 |
| Sexo | | | |
| Masculino | -0,172 | 0,073 | 0,842 |
| Egreso del secundario | | | |
| Antes del 2007 | -0,543 | 0,084 | 0,581 |
| Clase de escuela secundaria | | | |
| Público | -0,202 | 0,070 | 0,817 |
| Categoría ocupacional del alumno | | | |
| Trabaja | 0,183 | 0,070 | 1,201 |
| Estado ocupacional de la madre | | | |
| Trabaja | 0,342 | 0,081 | 1,407 |
| Nivel educativo del padre | 0,113 | 0,033 | 1,119 |
| Discapacidad | | | |
| Si | -1,559 | 0,144 | 0,210 |
| Rendimiento académico | 0,567 | 0,033 | 1,763 |

Para ilustrar la diferencia en el tipo de probabilidad predicha por cada uno de los enfoques se muestran en la Tabla 3 los valores obtenidos para un perfil particular -perfil (a)-, a saber: alumnas que no presentan ningún tipo de capacidad diferente, egresadas en el año 2007 de una escuela pública, con promedio de notas en su trayectoria universitaria entre 6 y 7,5 (en

una escala de 0 a 10), que no trabajan, cuya madre tampoco trabaja y cuyo padre tiene estudios secundarios completos y/o ha iniciado algún estudio terciario.

Tabla3: Probabilidades estimadas bajo ambos enfoques para el perfil (a)

| Año | Probabilidad acumulada | Probabilidad condicionada |
|---|---|---|
| Año 1 | 0,78 | 0,78 |
| Año 2 | 0,90 | 0,55 |
| Año 3 | 0,94 | 0,44 |
| Año 4 | 0,96 | 0,29 |
| Año 5 | 0,97 | 0,19 |

Según los resultados la Tabla 3, el modelo ordinal estima que la probabilidad de que una alumna con el perfil (a) alcance los requisitos para elegir carrera después de un año en la Facultad es de 0,78, valor que naturalmente coincide con el estimado a partir del enfoque binario. En cambio, a los tres años, el modelo ordinal estima con en 0,94 la probabilidad de cumplir los objetivos en 1, 2 ó 3 años en la Facultad. Por su parte el modelo correspondiente al enfoque binario estima con 0,44 la probabilidad de alcanzar los objetivos en el tercer año no habiéndolo logrado antes.

## 4. Consideraciones Finales

En este trabajo se presentan dos formas alternativas de enfrentar un problema de estudio sobre datos de duración desde la perspectiva de los MLG. Ellas suponen un tiempo continuo hasta la posibilidad de elegir carrera pero donde la ocurrencia de este evento se toma en cuenta una vez por año académico, lo que transforma la escala en puntos discretos en el tiempo. El enlace del MLG apropiado para este tipo de situaciones es el que fue utilizado en este trabajo. Sin embargo, si se consideran los tiempos como intrínsecamente discretos, en el sentido que sólo se puede elegir la carrera al inicio de cada año académico resultaría apropiado también utilizar el enlace logit. Los resultados se espera sean similares a los hallados en este trabajo pero las interpretaciones de los coeficientes de los modelos se realizan en términos de razones de odds.

## Agradecimientos

Se agradece especialmente a la Lic. Luciana Ruiz por su asesoramiento y asistencia en el tratamiento de la base de datos.

## REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2013). *Analysis of ordinal categorical data.* 2º edition. John Wiley & Sons, New York, USA.

Allison, P. (2006). *Survival analysis using SAS.A practical guide*. SAS Institute Inc. Cary, North Carolina, USA.

Bennet, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2: 273-277.

Hedecker, D.; Ohidul, S.; Frank, B.(2009). Random effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*9: 161-179.

Läärä, E.; Matthews, J. (1985). The equivalence of two models for ordinal data. *Biometrika*, 72 (1): 206-207.

Tutz G., Schmid, M. (2016). *Modeling discrete time to event data*. Springer Series in Statistics, Switzerland.