ANÁLISIS DE SUPERVIVENCIA APLICADO AL PROBLEMA DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD TECNOLÓGICA DE PEREIRA

MAURICIO BARRERA REBELLON

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERÍA INDUSTRIAL

Pereira, Julio de 2008

ANÁLISIS DE SUPERVIVENCIA APLICADO AL PROBLEMA DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD TECNOLÓGICA DE PEREIRA

MAURICIO BARRERA REBELLON

Director de Tesis.

MSc. Álvaro Trejos Carpintero

Proyecto presentado como requisito parcial para optar al título de

MAGISTER EN INVESTIGACION DE OPERATIVA Y ESTADISTICA

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERÍA INDUSTRIAL

Pereira, Julio de 2008

Nota de aceptación

_____

_____

_____

_____

_____

Jurado

_____

Jurado

RESUMEN

En este trabajo se presenta la metodología de análisis de supervivencia aplicada al problema de deserción estudiantil en la Universidad Tecnológica de Pereira, se utiliza la metodología de Kaplan-Meier y el modelo de regresión de COX. Se emplea la cohorte de estudiantes del primer semestre del año 2003 de todos los programas de la Universidad. Se construyeron modelos de supervivencia individuales para cada tipo de programa académico: de ingenierías, licenciaturas y tecnologías. Los resultados de este trabajo son útiles tanto para la planeación y diseño de políticas para evitar la deserción en la población estudiantil, como para el estudio de la metodología y su posterior aplicación.

ABSTRACT

In this research work applied survival analysis methodology to the problem of student desertion in the Universidad Tecnológica de Pereira, Kaplan-Meier's methodology is utilized and COX's model regression. The cohort of students of the first semester of the 2003 year is employed of all the programs of the University. Models of individual survival for each type of academic program were built: engineering, degrees and technologies. The results of this work are so much equipment for the planning and design of so much politics, to avoid the desertion in the student population, as for the study of the methodology and their subsequent application.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE GRAFICAS

1. INTRODUCCION

El fenómeno de la deserción estudiantil es un problema de gran complejidad debido no solo al gran numero de variables que intervienen sino también por la dificultad para describir al fenómeno. La deserción implica problemas a varios niveles, en lo personal el estudiante desertor ve frustradas sus aspiraciones profesionales además de perder oportunidades para obtener un mejor empleo, a nivel institucional la deserción logra grandes impactos económicos y financieros sobre las universidades debido a que se invierten recursos en personas que no cumplen sus proyectos de educación, en lo social es claro que los impactos son muy negativos: crece el ciclo de pobreza, incrementa subempleo, no se aportan conocimientos nuevos a la sociedad, entre otros.

Aunque en general se tiene conciencia de la problemática de la deserción, no es muy común encontrar trabajos que den luces sobre este fenómeno, en Colombia y más precisamente la Universidad Tecnológica de Pereira se está empezando a abordar el tema con algunos logros interesantes pero todavía no suficientes para la magnitud de las consecuencias de la deserción.

De lo mencionado anteriormente se tiene que analizar y resolver el problema de la deserción es de suma importancia. Poder identificar los años de mayores riesgos de abandono posibilita diseñar políticas que prevengan la deserción y, a su vez, lograr una administración más eficiente de los recursos escasos. También sería importante saber si el riesgo de desertar de un estudiante con unas características dadas es mayor que el de otro estudiante con otras características.

Se propone entonces realizar un análisis del fenómeno de la deserción en la UTP a través de una herramienta estadística llamada análisis de supervivencia que permita calcular el periodo en que es más probable que un estudiante de la universidad deserte además de que permita identificar y cuantificar las causas que aumentan el riesgo de desertar.

La información que permitirá realizar los análisis propuestos es proporcionada por la oficina de deserción y corresponde a la información de los programas con mayores niveles de deserción de la Universidad Tecnológica de Pereira. El análisis fue realizado sobre una base de datos que contenía la información referente a todos los matriculados por primera vez en los programas de pregrado de la Universidad y pertenecientes a la cohorte del primer semestre del 2003. El periodo de análisis es de 11 semestres (2003-1 hasta 2008_1).

La base de datos que se trabajo está compuesta por 1.264 registros de todos los programas académicos y con variables como se muestra a continuación:

Variables	**Descripción**
Documento	Es el código que identifica al estudiante
Codprg	Corresponde al programa al que el estudiante pertenece
Nombre	Nombre del estudiante
EDAD2	Edad del estudiante
Sexo	Genero del estudiante (1=Hombre; 0=Mujer)
Naturalezacol	Naturaleza jurídica del colegio (1=Privado; 0=Publico)
Estrato	Estrato socioeconómico
Depto	Departamento de origen (1=Risaralda; 0=Resto)
Estcivil	Estado civil del estudiante (1=Soltero; 0=Otro)
FECHAGRADO	Es la fecha en la que obtuvo el titulo
TIEMPO	Tiempo de permanencia en la institución
DESERCION	Se refiere al estado de deserción (1=Deserto; 0=Censura**)**

Esta base de datos fue desagregada y dividida en tres partes, una conteniendo los estudiantes de Ingenierías, otra los estudiantes de Tecnologías y otra los estudiantes de Licenciatura. Para cada una de estas bases se realizaron análisis independientes pudiéndose obtener conclusiones diferenciadas para cada tipo de programas de la Universidad Tecnológica de Pereira.

En cuanto al procedimiento de análisis inicialmente se realizaran análisis no parametricos (Kaplan-Meier), para cada tipo de programa académico, con esto se extraerán algunas conclusiones interesantes sobre todo en lo referente a la diferenciación de los estudiantes que desertan versus los que no lo hacen, también se pretende describir la situación de supervivencia presente en la Universidad. Después de realizado la fase anterior se procede a construir modelos semiparametricos (Regresión de Cox) para cada uno de los tipos de programas académicos, obteniéndose formas funcionales para la variable deserción de la Universidad Tecnológica de Pereira en sus diferentes tipos de programas.

Para realizar los cálculos referentes al modelamiento estadístico se utilizaran dos poderosos paquetes estadísticos SPSS 15.0 y SAS 9.0, en el SPSS se realizara todo lo concerniente al modelo de Kaplan-Meier, para los cálculos del modelo de Cox se utilizara el SAS debido básicamente a la imposibilidad que tiene el SPSS de realizar estos cálculos para variables de tiempo discreta como es el caso de este trabajo.

2. ANTECEDENTES

Los principales estudios sobre la permanencia de los estudiantes en la universidad reposan sobre dos teorías sociológicas; *el modelo de integración del estudiante* (Student Integration Model) Spady (1970), Tinto (1975); el cual indica que la integración del estudiante en el ambiente académico y social contribuye fuertemente a la decisión del estudiante de permanecer o desertar; y *el modelo de desgaste del estudiante* (Student Attrition Model) Bean (1980); el cual da mayor importancia relativa a los factores externos a la institución en la decisión de desertar.

Las investigaciones realizadas sobre el tema y basadas en los dos modelos sociológicos muestran que los alumnos desertores, comparados con los que permanecen, tienden a tener menores notas académicas y a tener padres con menores niveles educativos e inferiores ingresos economicos. Estas investigaciones también estiman que los estudiantes con un mayor nivel de interacción con los profesores y con otros estudiantes tienen menor probabilidad de desertar, así como también que la probabilidad de desertar es mayor al inicio de la carrera universitaria Robinson, R. (1990).

Sin embargo, las investigaciones anteriores abordan el problema de la deserción bajo un marco estático. En otras palabras, sólo se investiga la ocurrencia del evento así como los factores de los cuales depende pero sin tener en cuenta el tiempo en el que ocurre.

La herramienta estadística más utilizada en estas investigaciones ha sido el análisis de regresión, método que no permite captar la evolución del evento a lo largo del tiempo.

En Colombia el tema ha sido tratado básicamente desde el análisis de las características de la población que deserta, la construcción de índices de deserción y en análisis estadísticos descriptivos del fenómeno al interior de las instituciones[1].

En la Universidad Tecnológica de Pereira en el año 2004 Carvajal y Trejos[2] adelantaron un estudio a través de dos enfoques cuantitativo y cualitativo permitiesen explicar el fenómeno de la deserción. El enfoque cualitativo se basaba en la aplicación de la metodología de matriz de marco lógico donde participan todos los actores involucrados en el proceso de formación de los estudiantes y que contribuyen en el análisis con su percepción y experiencia. El enfoque cuantitativo utiliza herramientas estadísticas que permiten explorar tanto las causas del problema como la relación entre ellas.

[1] Castaño, E. "Determinantes de la deserción estudiantil en la universidad de Antioquia" 2003.
[2] Carvajal, P. Trejos, A. Caro, C. "Identificar las causas de deserción en la Universidad Tecnológica de Pereira usando la técnica multivariada análisis de correspondencias" 2004.

Uno de los inconvenientes del enfoque cuantitativo planteado por el estudio de Carvajal y Trejos fue el hecho de que los datos analizados presentaban censuras[3] que hacían que las herramientas estadísticas (análisis de correspondencias) que se aplicaron en su momento no fueran las más adecuadas.

Sin embargo el trabajo realizado por Carvajal y Trejos es muy valioso en la medida que logra descubrir cosas interesantes, y hábilmente logran hacer una caracterización del estudiante desertor de la UTP, así por ejemplo ellos concluyen que el estudiante desertor es hombre con promedio de edad de 19 años que estaba en primer semestre de carrera universitaria, proveniente de los estratos más bajos, que estudio en colegio público y de tipo académico, que el nivel académico de los padres llega máximo al bachillerato y adicionalmente el nivel de ingresos es de cerca de un salario mínimo, entre otras características.

En este estudio también se logran establecer causas de deserción dependiendo del programa y la jornada en la que estudia, así por ejemplo ellos plantean que en la jornada especial el retiro se da por dificultades que se presentan con las metodologías de enseñanza de los docentes. En programas como el de ingeniería mecánica la causa de deserción se debe principalmente al rendimiento académico. En ciencias de la salud se tiene como causa de deserción problemas de salud debidos en su mayor parte a la carga académica y a la presión que origina sobre los estudiantes. En ciencias de la educación el fenómeno se debe en su mayor parte a los embarazos no planeados que obedecen a una planificación de la sexualidad deficiente de las estudiantes de estas áreas.

Del estudio mencionado anteriormente surgen una serie de programas de mejoramiento, algunos puestos ya en práctica, como son:

- Mejoramiento del rendimiento académico de los estudiantes de la UTP
- Universidad saludable: Aprendizajes básicos para la vida
- Fortalecimiento de los recursos económicos de los estudiantes
- Creación de una gerencia para administrar la política orientada a mitigar la deserción.

[3] Casos censurados son aquellos a los que el evento de interés (deserción) no les ha ocurrido.

3. PLANTEAMIENTO DEL PROBLEMA

3.1 DIAGNÓSTICO DEL PROBLEMA DE INVESTIGACIÓN

Los estudiantes que abandonan sus estudios generan situaciones en los ámbitos personal, institucional y social que se pueden describir como sigue: En plano personal: el estudiante ve frustradas sus aspiraciones profesionales y la pérdida de oportunidades de empleos mejor remunerados que le permitan un ascenso social tanto a él como al grupo familiar que le rodea. En el plano social, es evidente que se incrementa el subempleo, creciendo el círculo de pobreza y la posibilidad de aporte intelectual que el profesional puede hacerle a la sociedad. En el plano institucional se genera mayor impacto económico en la universidad puesto que se invierte recursos en personas que luego abandonan sus proyectos de educación4.

3.2 FORMULACION DEL PROBLEMA

¿Qué factores influyen en la decisión de un estudiante de abandonar la universidad y en que semestre tiene mayor riesgo de hacerlo?

3.3 IMPORTANCIA DE LA INVESTIGACIÓN

La investigación aquí planteada es importante en dos sentidos, uno académico y uno institucional. En lo académico, la investigación proporciona una metodología estadística que permite describir el comportamiento de datos que corresponden al tiempo o duración desde un origen bien definido hasta la ocurrencia de algún evento o punto final, situación que técnicas clásicas como el análisis de regresión o el análisis discriminante no logran resolver adecuadamente.

A nivel institucional la investigación será de suma importancia ya que con el cálculo de la probabilidad de desertar de la universidad dadas ciertas características, permitirán diseñar políticas de permanencia, maximizando así el uso de los recursos disponibles en la universidad y minimizando los costos sociales.

3.4 POBLACION OBJETIVO

Los datos procesados corresponden a una muestra de estudiantes de la cohorte del año 2003 primer semestre. El nivel de significancia con el que se interpretaran los resultados es del 95%.

[4] Carvajal, P. Trejos, A. Caro, C. “Identificar las causas de deserción en la Universidad Tecnológica de Pereira usando la técnica multivariada análisis de correspondencias” 2004.

4. OBJETIVOS DE LA INVESTIGACIÓN

4.1 GENERAL

Realizar un análisis de supervivencia para los estudiantes de la Universidad Tecnológica de Pereira

4.2 ESPECÍFICOS

- Estimar la función de supervivencia a través del estimador de Kaplan y Meier para los estudiantes de la UTP.
- Seleccionar el mejor modelo de regresión de riesgos proporcionales (Modelo de Cox), para obtener los mejores predictores de la supervivencia de los estudiantes en la UTP.
- Interpretar los resultados de los mejores modelos de Cox obtenidos.
- Dar a la dirección un primer acercamiento a la caracterización del fenómeno de la deserción presente en la Universidad.

5. TEORIA DEL ANALISIS DE SUPERVIVENCIA[5]

El análisis de supervivencia constituye una serie de procedimientos y técnicas estadísticas para analizar datos en los cuales la variable de interés representa el tiempo de duración de un evento (El termino supervivencia se debe a que la técnica se utilizo inicialmente en aplicaciones donde el evento de interés es la muerte).

En general, cualquier variable puede medirse en forma instantánea, sin embargo, en temas de *"supervivencia"* la medición de la ocurrencia del evento de interés depende del tiempo de duración hasta que ocurra el evento, el cual suele ser bastante grande como para esperar a que el evento ocurra para todas las unidades experimentales al terminar el estudio. Es este hecho el que hace al análisis de supervivencia una herramienta fundamental y diferente para estudiar este tipo de fenómenos de duración.

En análisis de supervivencia cuando se tienen casos o individuos a los que no les ha ocurrido el evento que se estudia se dice que estos casos son censurados. Las causas de esta situación pueden ser: por ejemplo, cuando el evento (deserción) no tiene lugar antes de que finalice el estudio; en otros, el investigador puede haber perdido el seguimiento de su estado en algún momento anterior a que finalice el estudio; y existen además casos que no pueden continuar por razones ajenas al estudio (como el caso en el que un estudiante se enferme y deje de asistir a clases). El tipo de censura ejemplificado arriba es conocido como censura por la derecha, existe también censura por la izquierda en la cual por ejemplo se conoce al momento de iniciar la investigación la situación final del individuo pero no puede saberse cuando inició, y por lo tanto, no se tiene completa la información de la duración del evento.

5.1 ELEMENTOS BÁSICOS

En análisis de supervivencia es de suma importancia la variable aleatoria *T*, que representa el tiempo de duración hasta la ocurrencia de un evento (Muerte, deserción, falla de un sistema, entre otros). La variable aleatoria *T*, posee una función de densidad de probabilidad $f(t)$, y una función de distribución de probabilidad acumulada $F(t) = P(T \leq t)$, que representa la probabilidad de que el evento de interés ocurra en un tiempo menor o igual que *t.* Para el análisis de supervivencia, es de especial interés la función de supervivencia $S(t) = 1 - F(t) = P(T > t)$, la cual expresa la probabilidad de que el evento de interés ocurra en un tiempo mayor a *t.*

[5] La mayoría del trabajo aquí mostrado está basado en el libro de Hosmer, D. W. y S. Lemeshow. 1999

Otra función de interés para este análisis es la llamada función de riesgo, $h(t)$, que representa el riesgo instantáneo de que un evento ocurra en un intervalo infinitamente pequeño de tiempo $(t, t+\Delta t)$, dado que no ha ocurrido hasta el tiempo t, que se puede escribir como:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} \quad (1)$$

Desarrollando la probabilidad condicional de la expresión anterior se obtiene:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T < t + \Delta t)}{P(T > t)} \quad (2)$$

Del cálculo de probabilidades y del hecho de que $P(T > t) = S(t)$, se tiene que:

$$h(t) = \frac{1}{S(\mathrm{t})} \lim_{\Delta \mathrm{t} \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \quad (3)$$

De la definición de derivada se tiene que la expresión (3) se convierte en:

$$h(t) = f(t)/\mathrm{S(t)} \quad (4)$$

Se tiene entonces una expresión que relaciona la función de supervivencia y la función de riesgo. Si se integra la expresión (4), se obtiene entonces la función de riesgo acumulada:

$$H(t) = \int_0^t h(s)ds = \int_0^t \frac{f(s)}{S(\mathrm{s})} ds \quad (5)$$

Que se puede escribir como: $H(t) = \int_0^t \frac{f(s)}{1-F(\mathrm{s})} ds$, haciendo la sustitución $u = 1 - F(s)$, y $du = -f(s)ds$, se tiene que la expresión (4) se convierte en $H(t) = \int \frac{-du}{u}$, de donde resulta que:

$H(t) = -\ln S(t)$ (5.1)
$S(t) = e^{-H(t)}$ (5.2)
$h(t) = -d(\ln S(t))/dt$ (5.3)
$f(t) = -dS(t)/dt$ (5.4)

5.2 ESTIMADOR DE KAPLAN-MEIER

El estimador de Kaplan-Meier (1958), de la función de supervivencia, es el estimador más común entre los paquetes estadísticos. El método de construcción del estimador es un método no parametrico ya que no asume ninguna estructura para la función de distribución de probabilidad del tiempo de vida.

El estimador Kaplan-Meier utiliza toda la información disponible, casos censurados y no censurados, para realizar la estimación de la función de supervivencia. El estimador en cualquier instante de tiempo es obtenido de la multiplicación de una secuencia de probabilidades condicionales de supervivencia estimadas. Cada probabilidad condicional estimada se obtiene del número de casos observados en riesgo y el número de "muertes" en un instante de tiempo y se calcula como (n-d)/n, donde n es el número de casos en riesgo y d es el numero de "muertes" observadas.

Supóngase una muestra de n observaciones independientes y sean $t_1 < t_2 <\ldots< t_k$ los tiempo de vida observados y T_l el tiempo de vida de la observación de mayor duración en la muestra. Se define entonces:

$n_i =$ Numero de sujetos en riesgo en el instante t_i
$d_i =$ Numero de muertes en el instante t_i
$c_i =$ Numero de censuras en el intervalo $[t_i, t_{i+1})$

Se puede ver que:

$n_0 =$ n
$t_0 = 0$
$d_0 = 0$
$c_0 = 0$
$t_{k+1} = \infty$
$n_{i+1} = n_i - d_i - c_i \qquad i = 1,2, \ldots, k-1$

El estimador de la función de supervivencia de Kaplan-Meier $\hat{S}(t)$ se calcula de la siguiente manera:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \tag{7}$$

De la ecuación (7), resulta que $\hat{S}\ (t_0) = 1$ y que $\hat{S}\ (t_k) = 0$ si no existen casos censurados.

La función de supervivencia estimada se expresa entonces como:

$$\hat{S}\ (t_1), \hat{S}\ (t_2), \dots, \hat{S}\ (t_k)$$

5.2.1 Varianza del Estimador Kaplan-Meier

Adicional a la estimación de la función de supervivencia es de gran interés construir intervalos de confianza para los estimadores de la función de supervivencia así como de estimaciones puntuales y de intervalo para algunos percentiles.

Aunque existen varios enfoques para tratar este problema, en este trabajo se tratara el enfoque del *método delta* el cual está basado en la expansión en series de Taylor de primer orden.

Método Delta

La idea básica del método es usar una aproximación en series de Taylor para obtener una función lineal que aproxime una función más complicada como es el caso del estimador de la función de supervivencia. Para poder aplicar el método delta es necesaria que la función pueda ser expresada en series de Taylor.

Sea $f(x)$ una función de densidad de probabilidad de una variable aleatoria X. La expansión en series de Taylor de primer orden alrededor de la media es:

$$f(x) \cong f(u) + (x - u)f'(u) \qquad (8)$$

De la ecuación anterior se tiene que la varianza de la función $f(x)$ es aproximadamente igual a:

$$Var\big(f(x)\big) \cong [f'(u)]^2\ Var(x - u)$$
$$Var\big(f(x)\big) \cong [f'(u)]^2\ \sigma^2$$

Donde σ^2 es la varianza de la variable aleatoria X. Como ejemplo se muestra la varianza para la función logaritmo natural y la función exponencial

Con la función logaritmo natural:

$$ln(X) \cong ln(u) + (X - u)\frac{1}{u} \quad (9)$$

$$\widehat{Var}\big(ln(X)\big) \cong \frac{1}{\hat{u}^2}\hat{\sigma}^2 \quad (10)$$

Con la función exponencial:

$$exp(X) \cong exp(u) + (X - u)\exp(u) \quad (11)$$

$$\widehat{Var}\big(exp(X)\big) \cong \hat{\sigma}^2(\exp(\hat{u}))^2 \quad (12)$$

Varianza función de supervivencia

Para calcular el estimador de la varianza a través del método delta del estimador Kaplan-Meier, primero se calcula el estimador delta del logaritmo natural así:

$$\hat{S}(t) = \prod_{t_i \le t} \frac{n_i - d_i}{n_i}$$

$$\ln\left(\hat{S}(t)\right) = \sum_{t_i \le t} ln\frac{n_i - d_i}{n_i} \quad (13)$$

Haciendo $\hat{p}_i = \frac{n_i - d_i}{n_i}$, se tiene que la ecuación (13) se vuelve:

$$\ln\left(\hat{S}(t)\right) = \sum_{t_i \le t} ln\,(\hat{p}_i) \quad (14)$$

Aplicando el operador varianza a ambos lados de la ecuación (14), se tiene que:

$$\widehat{\mathrm{Var}}(\ln\left(\hat{S}(t)\right)) = \sum_{t_i \le t} Var(ln\,(\hat{p}_i)) \quad (15)$$

Lo anterior debido a la suposición de independencia de las variables $\hat{p}_i$. Si además suponemos que las $\hat{p}_i$ se distribuyen como variables aleatorias Bernoulli con probabilidad constante p_i. Se tiene entonces como estimador de p_i a $\hat{p}_i$, y como estimador de la

varianza la siguiente expresión $\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}$, ahora utilizando la ecuación (9) y (10) se tiene que:

$$\widehat{\mathrm{Var}}(\ln(\hat{p}_i)) \cong \frac{1}{\hat{p}_i^2}\frac{\hat{p}_i(1-\hat{p}_i)}{n_i} \tag{16}$$

Utilizando el hecho de que $\hat{p}_i = \frac{n_i - d_i}{n_i}$, la expresión (16) se convierte en:

$$\widehat{\mathrm{Var}}(\ln(\hat{p}_i)) \cong \frac{d_i}{n_i(n_i - d_i)} \tag{17}$$

Reemplazando (17) en (15), se tiene que:

$$\widehat{\mathrm{Var}}(\ln\left(\hat{S}(t)\right)) \cong \sum_{t_i \le t}\frac{d_i}{n_i(n_i - d_i)} \tag{18}$$

La expresión (18) da una estimación de la varianza del logaritmo natural de la función de supervivencia, lo que se debe hacer ahora es aplicar nuevamente el método delta sobre la función exponencial (como se hizo en (11) y (12)) para eliminar el logaritmo natural de la expresión (18).

Utilizando las ecuaciones (11) y (12) y el hecho de que $\hat{S}(t) = \exp\left(\ln\left(\hat{S}(t)\right)\right.$, se tiene:

$$\widehat{\mathrm{Var}}(\exp(\ln\left(\hat{S}(t)\right))) \cong \{\exp\left[\ln\left(\hat{S}(t)\right)\right]\}^2 \sum_{t_i \le t}\frac{d_i}{n_i(n_i - d_i)} \tag{19}$$

Simplificando:

$$\widehat{\mathrm{Var}}\left(\hat{S}(t)\right) \cong [\hat{S}(t)]^2 \sum_{t_i \le t}\frac{d_i}{n_i(n_i - d_i)} \tag{20}$$

La desviación estándar se calcula como la raíz cuadrada de la varianza dada en la expresión (19).

5.2.2 Media muestral y varianza

El estimador usado para la media es debido a un resultado de la teoría de la probabilidad sobre variables aleatorias positivas el cual enuncia que la esperanza matemática de una variable aleatoria positiva es igual a la integral definida (área bajo la curva) desde cero hasta infinito de la función de supervivencia como se muestra a continuación:

$$u = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$$

Debido a que la variable T no está definida para valores mayores a t_l,

$$u(t_l) = \int_0^{t_l} S(t)dt$$

El valor del estimador de la media es entonces el área de la función escalón definida por el estimador de Kaplan-Meier de la función de supervivencia así:

$$\hat{u} = \sum_{i=0}^{k-1} \hat{S}(t_i)(\, t_{i+1} - t_i) \qquad si \quad T_l = t_k \tag{21}$$

Donde T_l es el tiempo observado más grande, ahora si $T_l > t_k$ entonces el cálculo de la media se hace con:

$$\hat{u} = \sum_{i=0}^{k-1} \hat{S}(t_i)(\, t_{i+1} - t_i) + \hat{S}(t_k)(t_l - t_k) \tag{22}$$

Cuando no se tienen casos censurados, la varianza de la media muestral se calcula como la varianza muestral dividida sobre el tamaño de la muestra. Cuando se presentan censuras el estimador de la varianza de la media muestral se calcula como:

$$\widehat{Var}(\hat{u}) = \frac{d}{d-1} \sum_{i=1}^{k-1} \frac{\{\sum_{i=0}^{k-1} \hat{S}(t_i)(\, t_{i+1} - t_i)\}^2 d_i}{n_i(n_i - d_i)} \quad si \quad T_l = t_k,$$

Donde $d = \sum d_i$, la cantidad total de muertes. En el caso en el que $T_l > t_k$, la varianza se calcula como:

$$\widehat{Var}(\hat{u}) = \frac{d}{d-1}\sum_{i=1}^{k-1}\frac{\{\sum_{i=0}^{k-1}\hat{S}(t_i)(t_{i+1}-t_i)+\hat{S}(t_k)(t_l-t_k)\}^2 d_i}{n_i(n_i-d_i)} \qquad (23)$$

5.2.3 Estimación de percentiles y varianza

La función de supervivencia estimada puede ser utilizada para la estimación de algunos percentiles. El cálculo de los percentiles se puede hacer en forma grafica simplemente trazando una línea horizontal desde el eje Y de la grafica de la función de supervivencia hasta que toque la línea vertical, seguidamente se proyecta ese punto sobre el eje X y se obtiene el valor del percentil deseado. A continuación se muestra un ejemplo para una función de supervivencia dada:

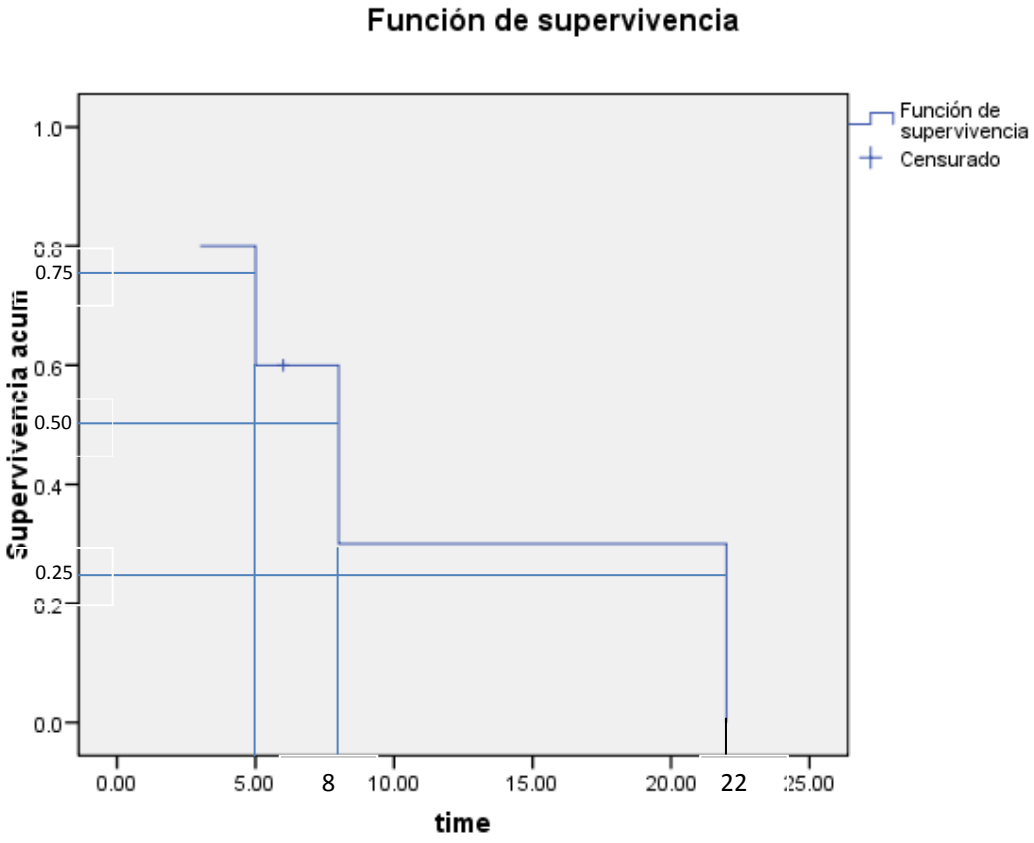

Gráfico 1 Función de supervivencia

El método grafico aunque ilustrativo se queda corto en precisión y en practicidad, por lo cual, se presenta un método más general y analítico que permite realizar cálculos más precisos. La fórmula para el cálculo se muestra a continuación:

$$\hat{t}_p = Min\{t: \hat{S}(t) \leq p\} \tag{24}$$

La varianza del estimador del $p - esimo$ percentil se estima con:

$$\widehat{Var}(\hat{t}_p) = \frac{\widehat{Var}(\hat{S}(\hat{t}_p))}{[\hat{f}(\hat{t}_p)]^2} \tag{25}$$

En la anterior expresión el numerador se calcula con la ecuación (19), y el denominador que es la estimación de la función de densidad de la distribución de los tiempos de supervivencia se calcula así:

$$\hat{f}(\hat{t}_p) = \frac{\hat{S}(\hat{u}_p) - \hat{S}(\hat{l}_p)}{\hat{l}_p - \hat{u}_p} \tag{26}$$

Donde $\hat{u}_p = Max(t: \hat{S}(t) \geq p + 0.05)$ y $\hat{l}_p = Min(t: \hat{S}(t) \leq p - 0.05)$, aunque se pudieran usar números inferiores al 0.05, este en la práctica funciona bien (Ver Hosmer, D. W., and S. Lemeshow. 1999).

5.2.4 Comparación funciones de supervivencia

Supóngase que se tienen dos grupos de estudiantes en una universidad, uno compuesto por estudiantes que han participado en un programa de acompañamiento con tutores y otro grupo que no ha participado en ningún programa y se desea realizar un análisis de supervivencia independiente para cada grupo y se desea comparar el efecto del programa de acompañamiento sobre la función de supervivencia. Inicialmente una comparación grafica de las funciones de supervivencia es importante e ilustrativa, sin embargo, a veces esta comparación grafica se vuelve muy subjetiva y en algunos casos difíciles de hacer. Para solucionar este inconveniente se hace necesario contar con pruebas estadísticas que permitan identificar si las diferencias observadas entre las funciones de supervivencia son significativas.

Las pruebas estadísticas están basadas en su mayoría en tablas de contingencia de grupo por estado de cada tiempo de supervivencia observado cómo se muestra en la siguiente tabla:

Tabla 1 Tabla usada para la prueba de igualdad de la función de supervivencia de dos grupos en el tiempo de observación t_i.

Evento	Grupo 1	Grupo 0	Total
Muerte	d_{1i}	d_{0i}	d_i
No muerte	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
En riesgo	n_{1i}	n_{0i}	n_i

El estadístico de prueba se construye con el número esperado de muertes y la varianza del número de muertes para alguno de los dos grupos, por ejemplo para el grupo 1 se tiene que el estimador del número de muertes es:

$$\hat{e}_{1i} = \frac{n_{1i} d_i}{n_i} \tag{27}$$

Y la varianza suponiendo una distribución hipergeometrica:

$$\hat{v}_{1i} = \frac{n_{1i} n_{0i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)} \tag{28}$$

Ya con los estimadores del número de muertes y su varianza se construye el estadístico de prueba Q.

$$Q = \frac{\sum_{i=1}^{k} w_i (d_{1i} - \hat{e}_{1i})}{\sum_{i=1}^{k} w_i^2 \hat{v}_{1i}} \tag{29}$$

Bajo la hipótesis nula de que las dos funciones de supervivencia son iguales y que el número observado de eventos es grande se puede demostrar que Q se distribuye como una chi-cuadrado con un grado de libertad ($p = P(\chi^2(1)) \geq Q$). Dependiendo del valor de w_i, se tienen diferentes tipos de contrastes como por ejemplo:

$w_i = 1$ Mantel y Haenzel
$w_i = n_i$ Breslow
$w_i = \sqrt{n_i}$ Tarone y Ware

El caso de más de dos grupos se trata en forma similar al caso ya visto de solo dos grupos. Se muestra a continuación la tabla asociada a varias funciones de supervivencia base para el cálculo de los estadísticos de prueba para la igualdad de las funciones:

Tabla 2 Tabla usada para la prueba de igualdad de la función de supervivencia de M grupos en el tiempo de observación $\boldsymbol{t_i}$.

Evento	1	2	…	m	…	M	Total
Muerte	d_{1i}	d_{2i}	…	d_{mi}	…	d_{Mi}	d_i
No muerte	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	…	$n_{mi} - d_{mi}$	…	$n_{Mi} - d_{Mi}$	$n_i - d_i$
En riesgo	n_{1i}	n_{2i}	…	n_{mi}	…	n_{Mi}	n_i

Inicialmente se debe estimar el número esperado de muertes de cada grupo bajo la suposición de igual función de supervivencia, es decir:

$$\hat{e}_{ki} = \frac{n_{ki} d_i}{n_i}, \qquad i = 1, \dots, m \tag{30}$$

Ya con varios grupos lo que se calcula ya no es una varianza sino una matriz de covarianzas del vector $\boldsymbol{d_i}$. La matriz de covarianzas es obtenida suponiendo que el número observado de eventos sigue una distribución hipergeometrica central multivariada. Los términos de la diagonal de la matriz denotada por $\widehat{\boldsymbol{V}}_{\boldsymbol{i}}$, se calculan de la siguiente manera:

$$\hat{v}_{mmi} = \frac{n_{mi}(n_i - n_{mi}) d_i (n_i - d_i)}{n_i^2 (n_i - 1)}, \qquad m = 1,2, \dots, M-1 \tag{31}$$

Los elementos fuera de la diagonal se calculan como:

$$\hat{v}_{lmi} = \frac{n_{li} n_{mi} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}, \qquad l, m = 1,2, \dots, M-1, \qquad l \neq m \tag{32}$$

Con los estimadores del número de muertes y la varianza se construye el siguiente estadístico:

$$Q = \left[\sum_{i=1}^{k} \boldsymbol{W_i} (\boldsymbol{d_i} - \hat{\boldsymbol{e}}_{\boldsymbol{i}}) \right]' \left[\sum_{i=1}^{k} \boldsymbol{W_i} \widehat{\boldsymbol{V}}_{\boldsymbol{i}} \boldsymbol{W_i} \right]^{-1} \left[\sum_{i=1}^{k} \boldsymbol{W_i} (\boldsymbol{d_i} - \hat{\boldsymbol{e}}_{\boldsymbol{i}}) \right], \tag{33}$$

Donde $\boldsymbol{W_i}$ es una matriz diagonal que tiene los pesos w_i, $\boldsymbol{d_i}$ es el vector que contiene la cantidad de muertes para cada grupo en el tiempo t_i, $\hat{\boldsymbol{e}}_{\boldsymbol{i}}$ es el vector que contiene el numero esperado de muertes para cada grupo en el tiempo t_i. Al igual que en el caso de dos grupos se tiene que Q para el caso de varios grupos, bajo la hipótesis nula de igualdad de funciones de supervivencia, se distribuye también como una chi-cuadrado con *M-1* grados de libertad ($p = P(\chi^2(1)) \geq Q$).

5.3 MODELO DE REGRESION DE COX

Supóngase que se quiere comparar la supervivencia de dos grupos de estudiantes, el primer grupo hace parte de un programa de acompañamiento con tutores el otro grupo no ha participado en ningún programa (caso ya enunciado en el apartado anterior), pero adicional a las diferencias que puedan tener los dos tipos de estudiantes en cuanto a la participación del programa o no, existen otras diferencias debido a la condición propia de cada estudiante, por ejemplo la edad, el estrato, el género, el nivel académico, entre otras. En la situación anterior el análisis de Kaplan-Meier se queda corto debido a su imposibilidad de tener en cuenta estas otras "covariables" o características diferentes al hecho de pertenecer o no a un programa de acompañamiento con tutores.

Lo que se pretende entonces es plantear un modelo para el riesgo $h(t)$, en función del tiempo y de variables explicativas o covariables. La idea básica del modelo es la misma del análisis de regresión, la diferencia ocurre cuando se presentan casos censurados que hacen que los modelos de regresión clásicos como el análisis de regresión lineal y logística fallen, de hecho si no existiera censura en los datos estos modelos de regresión serian adecuados y suficientes.

La presentación inicialmente se realizara sobre el supuesto de una sola covariable y después se generalizara a p variables. El modelo de regresión que se trata en este trabajo se puede escribir así:

$$h(t, x, \beta) = h_0(t) r(x, \beta) \tag{34}$$

La función de riesgo mostrada en (34), queda expresada por dos funciones. La función $h_0(t)$ que caracteriza como la función de riesgo cambia como una función del tiempo de supervivencia. Por otro lado la función $r(x, \beta)$ expresa como la función de riesgo cambia como una función de las covariables medidas a los sujetos. La función $h_0(t)$ que es la función de riesgo cuando $r(x, \beta) = 1$ es llamada la función base.

Con el modelo dado en (34), la razón de riesgos de dos sujetos con valor de covariable x_1 y x_2 respectivamente es:

$$RR(t, x_1, x_2) = \frac{h(t, x_1, \beta)}{h(t, x_2, \beta)}, \tag{35}$$

Utilizando (34) y reemplazando en (35) se obtiene:

$$RR(t, x_1, x_2) = \frac{r(x_1, \beta)}{r(x_2, \beta)}, \tag{36}$$

De la expresión anterior se tiene que la razón de funciones de riesgo de dos individuos depende solo de la función relacionada con la covariable $r(x, \beta)$. Si adicionalmente se supone una forma exponencial a la función de la covariable estamos ante el modelo de regresión de Cox (1972) o modelo de riesgos proporcionales debido a que la razón de riesgos depende solamente del valor que tome la covariable. El modelo de Cox se puede escribir así:

$$h(t, x, \beta) = h_0(t)e^{x\beta}, \tag{37}$$

En otras palabras el modelo de Cox es un modelo semiparametrico en el cual se supone una forma conocida para la función de las covariables y una desconocida para la función que depende solo del tiempo de supervivencia (ver modelo Kaplan-Meier). La razón de funciones de riesgo se escribe como:

$$RR(t, x_1, x_2) = e^{\beta(x_1 - x_2)}, \tag{38}$$

Uno de los atractivos del modelo de Cox planteado es lo que resulta de la ecuación (38), por ejemplo si la variable X indicara el género de un estudiante codificado con X=1 si es hombre y X=0 si es mujer, se tendría que la razón de riesgos entre hombres y mujeres sería igual a:

$$RR(t, x_1, x_2) = e^{\beta}$$

Ahora si de los cálculos resulta que $\beta = \ln(3)$, quiere decir que los hombres tienen un riesgo de desertar tres veces mayor que el riesgo de las mujeres.

La cuestión ahora será calcular la función de supervivencia para un modelo que parte de la función de riesgo. De la ecuación (5.2), y con la función de riesgo dada en (37) se obtiene:

$$S(t, x, \beta) = e^{-H(t,x,\beta)}, \tag{39}$$

Donde $H(t, x, \beta)$ es la función de riesgo acumulada que se puede calcular como:

$$\begin{aligned} H(t, x, \beta) &= \int_0^t h(u, x, \beta)\, du \\ &= r(x, \beta) \int_0^t h_0(t)\, du \\ &= r(x, \beta) H_0(t), \end{aligned} \tag{40}$$

Sustituyendo (40) en (39) se tiene:

$$S(t, x, \beta) = e^{-r(x,\beta)H_0(t)}, \tag{41}$$

De donde se sigue que:

$$\begin{aligned} S(t, x, \beta) &= \left[e^{-H_0(t)}\right]^{r(x,\beta)} \\ &= [S_0(t)]^{r(x,\beta)} \end{aligned} \tag{42}$$

Donde $S_0(t)$ es la función de supervivencia base. Ahora bajo la suposición del modelo de Cox (41) se convierte en:

$$S(t, x, \beta) = [S_0(t)]^{e^{x\beta}}, \tag{43}$$

5.3.1 Estimación de parámetros modelo de Cox

El procedimiento más utilizado para realizar la estimación de los parámetros de un modelo de regresión es el procedimiento relacionado con la verosimilitud que básicamente consiste en estimar los parámetros que maximice el logaritmo natural de la verosimilitud, en el caso del modelo de Cox este procedimiento no funciona básicamente debido al desconocimiento de la función de riesgo base que aparece en el modelo.

Cox en 1972 desarrolla un método que permite estimar los parámetros sin necesidad de asumir una forma particular para la función de riesgo base, propone entonces usar una expresión llamada "*función de verosimilitud parcial*" que depende solamente de los parámetros de interés. La función de verosimilitud parcial se puede escribir como:

$$l_p = \prod_{i=1}^{k} \frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \tag{44}$$

Donde el producto se hace sobre K tiempos de supervivencia distintos y $x_{(i)}$ corresponde con el valor de la covariable del individuo que tiene el tiempo t_i, y $R(t_i)$ se refiere a todos los individuos en riesgo en el tiempo t_i, y corresponde a todos los individuos con tiempos de supervivencia o censura superiores a t_i.

Calculando el logaritmo natural de la función de verosimilitud parcial se tiene:

$$L_p = \sum_{i=1}^{k} \left\{ x_{(i)}\beta - \ln \left[\sum_{j \in R(t_i)} e^{x_j \beta} \right] \right\}, \tag{45}$$

Derivando con respecto a β se obtiene:

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{k} \left\{ x_{(i)} - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right\} \tag{46}$$

La ecuación (46) se debe igualar a cero y resolverse, pero como se puede notar la ecuación resultante, debido a su complejidad, debe ser resuelta a través de un método numérico como el de Newton-Raphson con algún criterio de parada:

$$\sum_{i=1}^{k} \left\{ x_{(i)} - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right\} = 0 \tag{47}$$

El estimador de la varianza del estimador del parámetro es obtenido de la segunda derivada. Tomando la derivada en (46) se tiene:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{k} \left\{ \frac{\left[\sum_{j \in R(t_i)} e^{x_j \beta}\right]\left[\sum_{j \in R(t_i)} {x_j}^2 e^{x_j \beta}\right] - \left[\sum_{j \in R(t_i)} x_j e^{x_j \beta}\right]^2}{\left[\sum_{j \in R(t_i)} e^{x_j \beta}\right]^2} \right\} \tag{48}$$

El negativo del lado derecho de la ecuación (48) es llamada *información observada* (cuando se presenta más de una covariable se llama *matriz de información observada*) y se denota por:

$$I(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2} \tag{49}$$

El estimador de la varianza del coeficiente estimado es el inverso de (49) evaluado para $\hat{\beta}$ y se calcula:

$$\widehat{Var}(\hat{\beta}) = I^{-1}(\hat{\beta}) \tag{50}$$

La desviación estándar $\widehat{SE}(\hat{\beta})$, se calcula como la raíz cuadrada de la varianza dada en (50). Para determinar la significancia del coeficiente se presentan las tres estadísticas más usadas en la práctica: La prueba de la razón de verosimilitud, la prueba de Wald y la prueba de puntajes.

La prueba de la razón de verosimilitud denotada por G, se calcula como dos veces la diferencia entre el logaritmo natural de la verosimilitud parcial del modelo que contiene la variable y el logaritmo de la verosimilitud parcial del modelo sin la variable, es decir:

$$G = 2\{L_p(\hat{\beta}) - L_p(0)\} \tag{51}$$

Donde

$$L_p(0) = -\sum_{i=1}^{k} \ln(n_i) \tag{52}$$

Y n_i denota el número de individuos en riesgo en el tiempo de supervivencia observado t_i. Bajo la hipótesis nula que el coeficiente es igual a cero, el estadístico G se distribuye chi-cuadrado con un grado de libertad.

La prueba de Wald se calcula como la razon entre el coeficiente estimado y la desviación estándar de ese coeficiente como se muestra a continuación:

$$Z = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})} \tag{53}$$

Bajo la hipótesis nula que el coeficiente es igual a cero, el estadístico de Wald se distribuye normal estándar.

La prueba de puntajes se calcula como la razón de la derivada del logaritmo natural de la verosimilitud parcial y la raíz cuadrada de la información observada evaluadas las dos en $\beta = 0$ como se muestra abajo:

$$\dot{Z} = \frac{{\partial L_p}/{\partial \beta}}{\sqrt{I(\beta)}} \quad \beta = 0 \tag{54}$$

Bajo la hipótesis nula que el coeficiente es igual a cero, el estadístico de puntaje se distribuye normal estándar.

Los intervalos de confianza se construyen de la siguiente manera:

$$\hat{\beta} \pm Z_{1-{\alpha}/{2}} \widehat{SE}(\hat{\beta}) \tag{55}$$

Lo anterior visto corresponde al caso en que existe solo una covariable, en adelante se generalizara el procedimiento visto al caso de muchas covariables.

Sean p covariables medidas para el individuo i que deban ser incluidas en el modelo de regresión y denotadas por $\boldsymbol{X}_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$. Estas variables pueden ser de tipo continuo, nominales o dummy. El nuevo modelo multivariable se construye de manera similar que el del caso univariado, solo que ahora se deben estimar p parametros. En la expresión (37) se sustituye la variable unidimensional x por la variable multidimensional $\boldsymbol{X}$, quedando así:

$$h(t, \boldsymbol{X}, \boldsymbol{\beta}) = h_0(t) e^{\boldsymbol{X'\beta}}, \tag{55}$$

La función de verosimilitud parcial se construye de la misma manera a como se construyo el modelo, simplemente se reemplaza en (44), la variable unidimensional x por la variable multidimensional $\boldsymbol{X}$, quedando así:

$$l_p = \prod_{i=1}^{k} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} \tag{56}$$

Después de aplicar logaritmo natural a (56) y calcular p derivadas parciales, una por cada covariable en el modelo, se tienen p ecuaciones como se muestra a continuación:

$$\frac{\partial L_p(\boldsymbol{\beta})}{\partial \beta_m} = \sum_{i=1}^{k} \left\{ x_{(im)} - \frac{\sum_{j \in R(t_i)} x_{jm} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} \right\} \qquad m = 1, \dots, p \tag{57}$$

Entonces el estimador de máxima verosimilitud parcial se denota por $\widehat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m)$.

La matriz de información es obtenida de la matriz Hessiana de la siguiente forma:

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \tag{58}$$

Donde los elementos de la diagonal vienen dados por:

$$\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \beta_m^2} = -\sum_{i=1}^{k} \sum_{j \in R(t_i)} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} \left[x_{jm} - \sum_{j \in R(t_i)} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} x_{jm} \right]^2$$

Los elementos fuera de la diagonal vienen dados por:

$$\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_m} = -\sum_{i=1}^{k} \sum_{j \in R(t_i)} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} \left[x_{jl} - \sum_{j \in R(t_i)} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} x_{jl} \right] \left[x_{jm} - \sum_{j \in R(t_i)} \frac{e^{\boldsymbol{X'}_{(i)}\boldsymbol{\beta}}}{\sum_{j \in R(t_i)} e^{\boldsymbol{X'}_j \boldsymbol{\beta}}} x_{jm} \right]$$

El estimador de la matriz de covarianzas de los estimadores de máxima verosimilitud parcial de los parámetros del modelo vienen dados por:

$$\widehat{\boldsymbol{Var}}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{I}^{-1}\left(\widehat{\boldsymbol{\beta}}\right) \tag{59}$$

Las pruebas estadísticas vistas para el caso univariado se pueden generalizar para el caso multivariado realizando unos ajustes.

La prueba de la razón de verosimilitud G, se calcula igual que al caso univariado, como dos veces la diferencia entre el logaritmo natural de la verosimilitud parcial del modelo que contiene todas las variables y el logaritmo de la verosimilitud parcial del modelo sin las variables, es decir:

$$G = 2\left\{L_p\left(\widehat{\boldsymbol{\beta}}\right) - L_p(\mathbf{0})\right\} \tag{60}$$

Donde $L_p(0)$representa la verosimilitud del modelo con cero variables igual que en (51). En este caso G también se distribuye, bajo la hipótesis nula de que los coeficientes son iguales a cero, como una chi-cuadrado con p grados de libertad (un grado por cada variable en el modelo).

Para calcular los estadísticos de Wald y de puntuación estadística implican cálculos matriciales. Si se denota el vector de primeras derivadas parciales de la función parcial de verosimilitud evaluado en **0** como $\mathbf{U}(\mathbf{0}) = \mathbf{U}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}}$, se tiene que bajo la hipótesis nula de que todos los coeficientes son iguales a cero y algunas otras condiciones de la función de verosimilitud parcial, este vector se distribuye normal multivariado con media igual a cero y matriz de covarianzas dada por la matriz de información evaluada en el vector **0**, $\boldsymbol{I}(\mathbf{0}) = \boldsymbol{I}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\mathbf{0}}$. La matriz de información es obtenida de la expresión (58).

La prueba de puntajes para el caso multivariado se calcula entonces como:

$$\boldsymbol{U}'(\mathbf{0})[\boldsymbol{I}(\mathbf{0})]^{-1}\boldsymbol{U}(\mathbf{0}) \tag{61}$$

Bajo la hipótesis nula de que todos los coeficientes son iguales a cero la expresión (61) se distribuye chi-cuadrado con p grados de libertad (un grado por cada variable en el modelo).

La prueba de Wald se obtiene del hecho de que $\widehat{\boldsymbol{\beta}}$, el estimador de los coeficientes se distribuirá asintóticamente normal con vector de medias igual a cero y matriz de covarianzas dada por la ecuación (59). El estadístico de Wald se escribe como:

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{I}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}} \tag{62}$$

Bajo la hipótesis nula de que todos los coeficientes son iguales a cero la expresión (62) se distribuye chi-cuadrado con p grados de libertad (un grado por cada variable en el modelo).

5.3.2 Modelo de Prentice-Gloecker

El modelo de regresión de Cox tal como fue expuesto en este trabajo es para tiempos continuos. Pero para el caso en que el evento de interés es observable solamente en un intervalo de tiempo discreto o bien la duración es intrínsecamente discreta, el modelo de Cox debe de ajustarse. En el año de 1978 Prentice y Gloecker construyen un modelo para tratar con datos en tiempos discretos.

Supóngase que el tiempo de supervivencia para el i-esimo individuo está limitado por dos valores conocidos, denotados por $a_i < T \leq b_i$. Sea la función de supervivencia en el periodo t para un individuo con vector de coavariables $\boldsymbol{x_i}$ y de vector de parámetros asociado $\boldsymbol{\beta}$, denotada por $S(t, \boldsymbol{x_i}, \boldsymbol{\beta})$. La probabilidad de que al individuo i-esimo le ocurra el evento en el intervalo $a_i < T \leq b_i$ es:

$$[S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta})]^{1-c_i}[S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta}) - S(b_i, \boldsymbol{x_i}, \boldsymbol{\beta})]^{c_i} \tag{62}$$

Donde c_i es el estado de censura del individuo i-esimo. La expresión arriba muestra que la probabilidad de que le ocurra el evento a casos censurados (a la derecha) es igual a $S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta})$ y la probabilidad de que le ocurra a casos no censurados es $S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta}) - S(b_i, \boldsymbol{x_i}, \boldsymbol{\beta})$.

La función de verosimilitud, a partir de la ecuación (62), se calcula entonces como:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta})]^{1-c_i} [S(a_i, \boldsymbol{x_i}, \boldsymbol{\beta}) - S(b_i, \boldsymbol{x_i}, \boldsymbol{\beta})]^{c_i} \qquad (63)$$

Se suponen J+1 intervalos denotados por $(t_{j-1}, t_j]$ para j=1,2,…, J+1 con $t_0 = 0$ y $t_{J+1} = \infty$ y donde el j-esimo intervalo (I_j) es denotado por$(t_{j-1}, t_j]$ así:

$$\{(0, t_1], (t_2, t_3], \dots, (t_j, t_j], \dots, (t_J, \infty)\}$$

Para especificar el intervalo de interés se define:

$$y_{ij} = \begin{cases} 1 \; si \; (a_i, b_i] = I_j \\ 0 \; otro \; caso \end{cases}$$

Reescribiendo la probabilidad de que el evento ocurra para el j-esimo intervalo se tiene:

$$S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}) - S(t_j, \boldsymbol{x_i}, \boldsymbol{\beta}) = [S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta})] \left\{1 - \frac{S(t_j, \boldsymbol{x_i}, \boldsymbol{\beta})}{S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta})}\right\} \qquad (64)$$

Donde $\left\{1 - \frac{S(t_j, \boldsymbol{x_i}, \boldsymbol{\beta})}{S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta})}\right\}$ es la probabilidad de que el evento ocurra en el intervalo $(t_{j-1}, t_j]$ dado que no ha ocurrido hasta el instante t_{j-1} en otras palabras la probabilidad dada por la expresión $P(t_{j-1} < T \leq t_j | T > t_j)$. Bajo el supuesto del modelo de Cox $S(t, \boldsymbol{x}, \boldsymbol{\beta}) = [S_0(t)]^{e^{\boldsymbol{x}'\boldsymbol{\beta}}}$ se tiene que:

$$\frac{S(t_j, \boldsymbol{x_i}, \boldsymbol{\beta})}{S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta})} = \frac{[S_0(t_j)]^{e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}}{[S_0(t_{j-1})]^{e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}}$$

$$= \left[\frac{S_0(t_j)}{S_0(t_{j-1})}\right]^{e^{\boldsymbol{x_i}'\boldsymbol{\beta}}} \qquad (65)$$

Sacando logaritmo natural a ambos de (65) y aplicando el operador exponencial se tiene que:

$$\frac{S(t_j, \boldsymbol{x_i}, \boldsymbol{\beta})}{S(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta})} = exp(-\exp(\boldsymbol{x_i'}\boldsymbol{\beta} + \tau_j)) \qquad (66)$$

Donde:

$$\tau_j = \ln\left[-\ln\left(\frac{S_0(t_j)}{S_0(t_{j-1})}\right)\right]$$

Reemplazando (66) en (64) y luego en (63) se tiene:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\left[S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right)\right]^{1-c_i}\left[S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right) - S\left(t_j, \boldsymbol{x_i}, \boldsymbol{\beta}\right)\right]^{c_i}\right\}^{y_{ij}}$$

$$= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\left[S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right)\right]\left\{1 - \frac{S\left(t_j, \boldsymbol{x_i}, \boldsymbol{\beta}\right)}{S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right)}\right\}^{c_i}\right\}^{y_{ij}}$$

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\left[S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right)\right]\left\{1 - exp\left(-\exp\left(\boldsymbol{x_i'\beta} + \tau_j\right)\right)\right\}^{c_i}\right\}^{y_{ij}} \quad (67)$$

Donde τ_j es como en (66).

La función de supervivencia se calcula en forma parecida al estimador de Kaplan-Meier obteniéndose:

$$S\left(t_{j-1}, \boldsymbol{x_i}, \boldsymbol{\beta}\right) = \prod_{l=1}^{j-1}\left[exp(-\exp(\boldsymbol{x_i'\beta} + \tau_l))\right] \quad (68)$$

Sustituyendo (68) en (67) se obtiene:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\prod_{l=1}^{j-1}\left[exp(-\exp(\boldsymbol{x_i'\beta} + \tau_l))\right]\left\{1 - exp\left(-\exp\left(\boldsymbol{x_i'\beta} + \tau_j\right)\right)\right\}^{c_i}\right\}^{y_{ij}} \quad (69)$$

Para el i-esimo sujeto denotamos con m_i el intervalo observado para este sujeto, es decir, $I_{m_i} = (a_i, b_i]$ se tiene que todos los términos en la productoria sobre j son siempre iguales a uno para $j \neq m_i$ con lo que se obtiene:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{1 - exp\left(-\exp\left(\boldsymbol{x_i'\beta} + \tau_{m_i}\right)\right)\right\}^{c_i}\prod_{j=1}^{m_i-1}\left[exp\left(-\exp\left(\boldsymbol{x_i'\beta} + \tau_j\right)\right)\right] \quad (70)$$

Si en la expresión anterior se realiza el cambio:

$$\theta_{ij} = 1 - exp\left(-\exp\left(\boldsymbol{x_i'\beta} + \tau_j\right)\right)$$

Se obtiene entonces:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{\theta_{im_i}\right\}^{c_i}\prod_{j=1}^{m_i-1}\left[1 - \theta_{ij}\right] \quad (71)$$

Definiendo a $z_{ij} = y_{ij} \times c_i$, y haciendo los cambios adecuados la ecuación (71) se puede escribir como:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{m_i - 1 + c_i} \left[1 - \theta_{ij}\right]^{1 - z_{ij}} \left[\theta_{ij}\right]^{z_{ij}} \tag{72}$$

La expresión anterior se puede ver como la función de verosimilitud de un modelo de regresión binario.

6. MODELO DE KAPLAN-MEIER

En esta sección se presentaran los resultados obtenidos de la aplicación de la metodología de Kaplan-Meier para los programas de Ingenierías, Licenciaturas y Tecnologías. Se asimilará el tiempo de supervivencia de un estudiante como el tiempo el cual aquel permanece en la universidad, la mortalidad se asimila a la deserción y la censura es equivalente a la salida del estudiante de la universidad por causas diferentes a la deserción (Graduación por ejemplo).

INGENIERIAS

Por Programa

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por programa académico y sus comparaciones:

Tabla 3 Resumen

Programa	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
INGENIERIA ELECTRICA	72	40	32	44,4%
INGENIERIA INDUSTRIAL	75	23	52	69,3%
INGENIERIA MECANICA	71	39	32	45,1%
INGENIERIA DE SISTEMAS Y COMPUTACION	76	30	46	60,5%
INGENIERIA FISICA	83	50	33	39,8%
Global	377	182	195	51,7%

En la tabla anterior *No Total* se refiere a la cantidad de estudiantes analizados en la cohorte que pertenecen al programa que se indica, para ingeniería eléctrica se estudiaron en total 72 estudiantes; *No Eventos* corresponde a la cantidad de estudiantes desertores durante el periodo de análisis, para el caso de ingeniería eléctrica se tiene que de la primera cohorte del año 2003 han desertado 40 estudiantes; Los casos censurados corresponden a los estudiantes que sin haber desertado han salido del programa académico (por ejemplo cuando se gradúa) o permanecen en el programa hasta el fin del periodo de análisis (alta permanencia), para el caso de ingeniería eléctrica se tiene que se han graduado, se han ido de la universidad y permanecen en la universidad 32 estudiantes.

Tabla 4 Medias del tiempo de supervivencia

PROGRAMA	Media			
	Estimación	Error típico	Intervalo de confianza al 95%	
			Límite inferior	Límite superior
INGENIERIA ELECTRICA	6,375	,522	5,352	7,398
INGENIERIA INDUSTRIAL	8,426	,472	7,502	9,351
INGENIERIA MECANICA	6,732	,516	5,720	7,744
INGENIERIA DE SISTEMAS Y COMPUTACION	7,645	,502	6,661	8,629
INGENIERIA FISICA	5,916	,485	4,966	6,866
Global	7,005	,228	6,558	7,452

En la tabla anterior *Estimación* se refiere al tiempo medio de permanencia de los estudiantes analizados en el estudio, así para ingeniería eléctrica el tiempo medio de permanencia en el programa es de 6,375 semestres; *Error Tipico* se refiere al error que se puede cometer en la estimación del tiempo medio de supervivencia; *Intervalo de confianza* se refiere a los limites entre los cuales se espera este el valor medio del tiempo de supervivencia de los estudiantes, para ingeniería eléctrica se espera la media de supervivencia este entre 5,352 y 7,398 semestres; *Confianza* se refiere a la probabilidad con la cual se espera que el intervalo de confianza contenga a la media de supervivencia del programa, en este trabajo se utilizo una confianza del 95%.

Tabla 5 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	17,771	4	,001
Breslow (Generalized Wilcoxon)	15,334	4	,004
Tarone-Ware	16,681	4	,002

En la tabla 5 se presentan tres pruebas estadísticas[6] para la igualdad de las 5 funciones de supervivencia, en estas pruebas estadísticas la hipótesis nula se refiere a la igualdad de las funciones de supervivencia, como se aprecia se puede concluir que existen diferencias significativas entre las funciones de supervivencia a un nivel de significancia del 5% (los niveles de significancia son demasiado bajos para todos los estadísticos).

[6] Estas pruebas son tratadas en detalle en la página 17 y 18 de este texto.

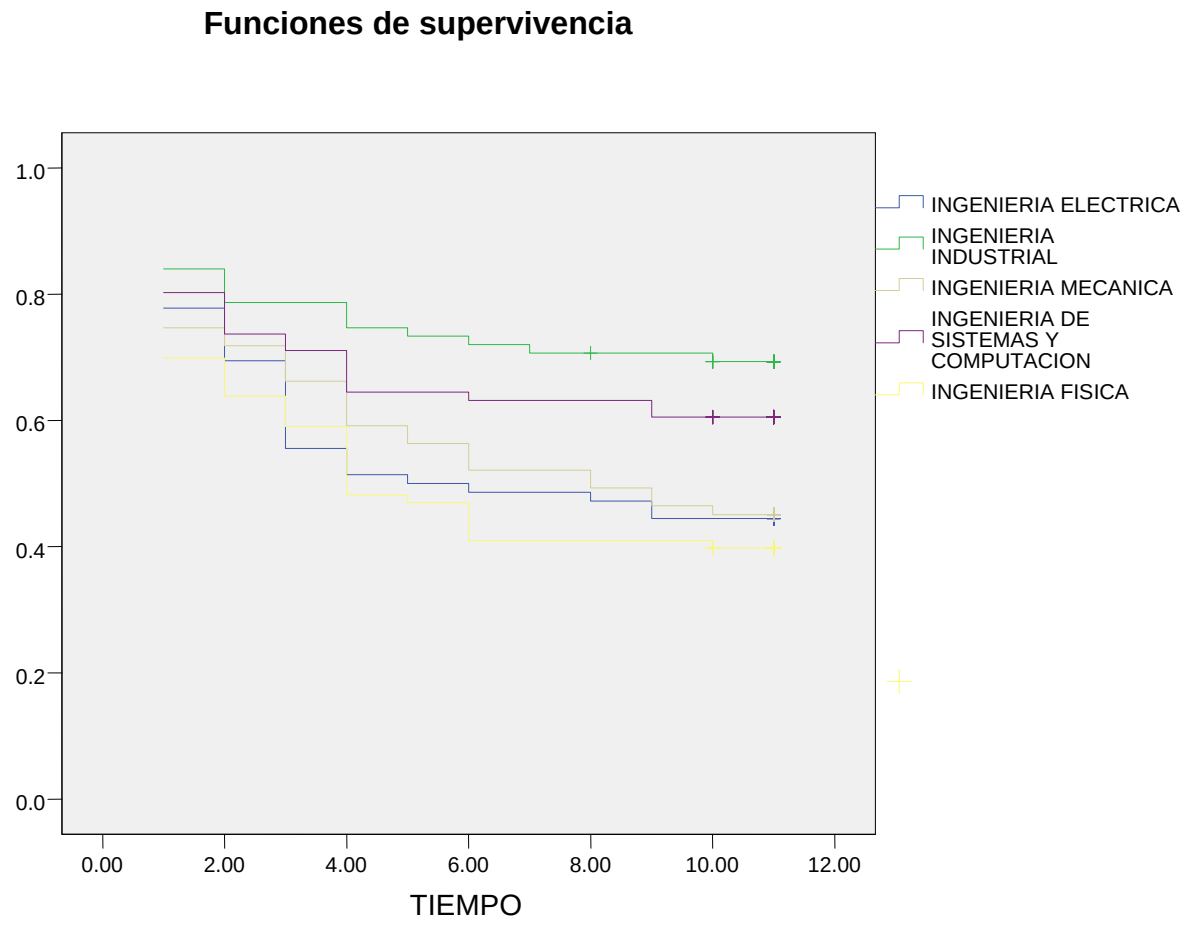

Gráfico 2 Funciones de supervivencia

En la grafica 2 se muestran las diferentes curvas de supervivencia para los diferentes programas de ingeniería. Existen diferencias significativas entre ingeniería industrial y de sistemas con ingeniería mecánica, eléctrica y física.

Por Género

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por género y sus comparaciones:

Tabla 6 Resumen

GENERO			Censurado	
	Nº total	Nº de eventos	Nº	Porcentaje
MUJER	106	42	64	60,4%
HOMBRE	271	140	131	48,3%
Global	377	182	195	51,7%

En la tabla anterior se muestra el resumen de casos analizados con sus respectivas estadísticas de censura. Al que más le ocurrió el evento fue a los hombres 51,7%

Tabla 7 Medias del tiempo de supervivencia

sexo	Media			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
MUJER	7,594	,424	6,763	8,425
HOMBRE	6,775	,269	6,247	7,303
Global	7,005	,228	6,558	7,452

En la tabla 7 se aprecia que los intervalos de confianza para esas estimaciones se traslapan, es decir, no es muy clara la diferencia de las funciones de supervivencia.

Tabla 8 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	3,774	1	,052
Breslow (Generalized Wilcoxon)	2,979	1	,084
Tarone-Ware	3,379	1	,066

En la tabla anterior se presentan las pruebas para la igualdad de las funciones de supervivencia de hombre y mujeres, y como se puede ver aunque existen diferencias estas no son muy significativas (al 5%).

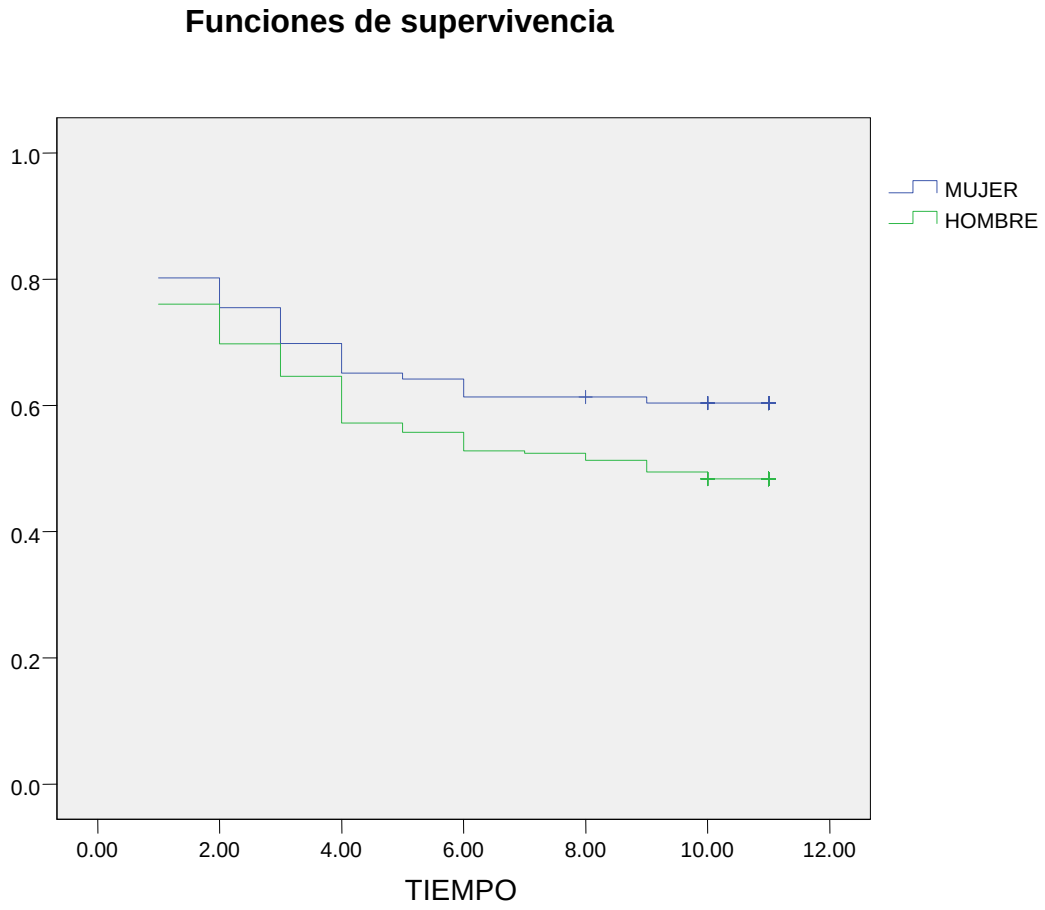

Gráfico 3 Funciones de supervivencia

En la anterior grafica se presentan las curvas de supervivencia para hombres y mujeres, se aprecia entonces como en los primeros semestres las dos curvas de supervivencia están cercanas mientras que después del quinto semestre se empieza a percibir un clara diferencia entre estas dos curvas.

Por Tipo de Colegio

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por tipo de colegio y sus comparaciones:

Tabla 9 Resumen

Tipo de colegio			Censurado	
	Nº total	Nº de eventos	Nº	Porcentaje
PUBLICO	228	114	114	50,0%
PRIVADO	149	68	81	54,4%
Global	377	182	195	51,7%

En la tabla anterior se presenta el resumen de casos analizados y sus censuras. Al que más le ocurrió el evento fue a los estudiantes de colegio público 50%.

Tabla 10 Medias del tiempo de supervivencia

	Media			
			Intervalo de confianza al 95%	
Tipo de Colegio	Estimación	Error típico	Límite inferior	Límite superior
PUBLICO	6,868	,293	6,294	7,443
PRIVADO	7,214	,363	6,503	7,925
Global	7,005	,228	6,558	7,452

Al igual que en el caso del genero aquí no es muy clara la superioridad de los estudiantes provenientes de los colegios privados sobre los estudiantes provenientes de los colegios públicos.

Tabla 11 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	,614	1	,433
Breslow (Generalized Wilcoxon)	,463	1	,496
Tarone-Ware	,540	1	,462

En la tabla anterior se presenta las estadísticas de prueba para la igualdad de las dos funciones de supervivencia, las tres estadísticas muestran que no existen diferencias significativas entre las dos funciones de supervivencia.

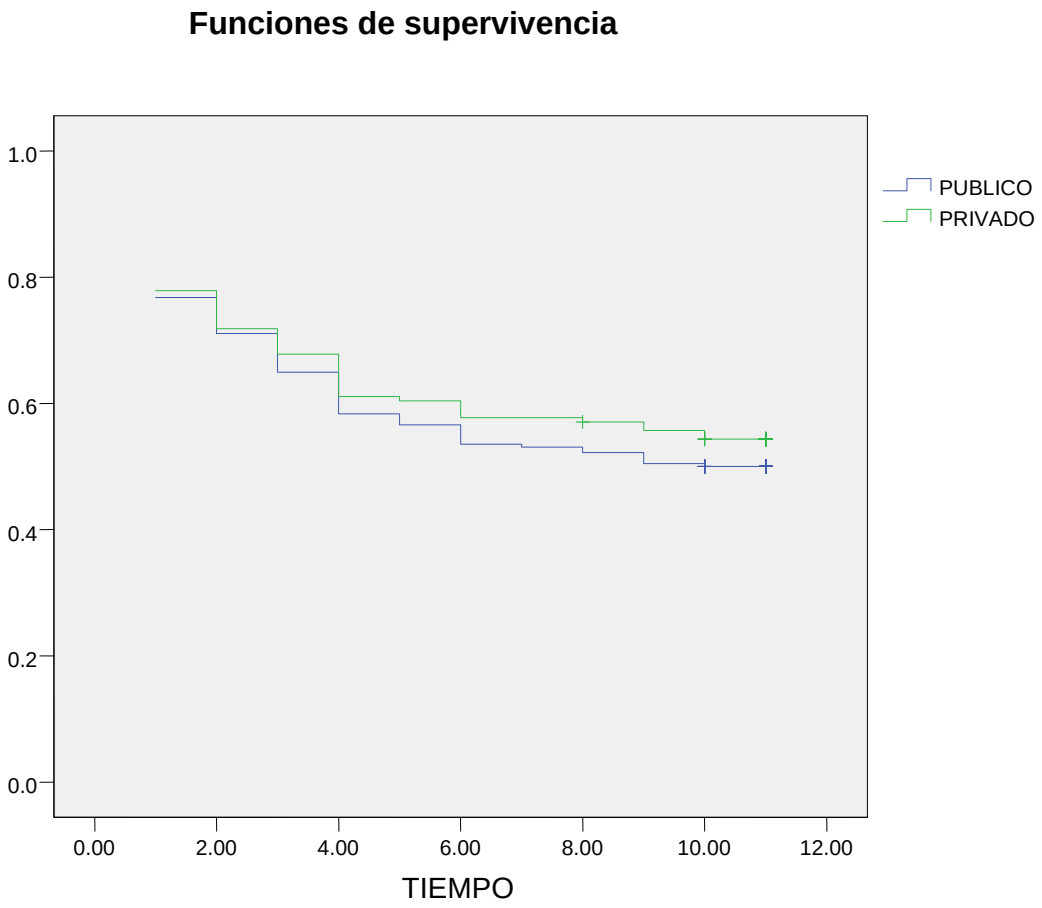

Gráfico 4 Funciones de supervivencia

La grafica anterior muestra diferencias no significativas entre las curvas de supervivencia.

LICENCIATURAS

Por Programa

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por programa académico y sus comparaciones:

Tabla 12 Resumen

Programa	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
LICENCIATURA EN MUSICA	42	31	11	26,2%
LICENCIATURA EN ARTES VISUALES	64	45	19	29,7%
LICENCIATURA EN MATEMATICAS Y FISICA	66	53	13	19,7%
LICENCIATURA EN ESPAÑOL Y LITERATURA	65	43	22	33,8%
ETNOEDUCACION Y DESARROLLO COMUNITARIO	51	23	28	54,9%
LICENCIATURA EN PEDAGOGIA INFANTIL	102	49	53	52,0%
Global	390	244	146	37,4%

En la tabla anterior se presenta el resumen de casos y sus respectivas censuras, se aprecia que licenciatura en matemáticas y física es el programa que mas deserción ha presentado en términos relativos 80.3% (100% - 19.7%).

Tabla 13 Medias del tiempo de supervivencia

Programa	Media			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
LICENCIATURA EN MUSICA	5,667	,623	4,445	6,888
LICENCIATURA EN ARTES VISUALES	5,219	,540	4,161	6,277
LICENCIATURA EN MATEMATICAS Y FISICA	3,652	,475	2,721	4,582
LICENCIATURA EN ESPAÑOL Y LITERATURA	4,969	,553	3,885	6,054
ETNOEDUCACION Y DESARROLLO COMUNITARIO	7,451	,601	6,272	8,630
LICENCIATURA EN PEDAGOGIA INFANTIL	7,097	,433	6,249	7,945
Global	5,743	,225	5,302	6,185

En la tabla 13 se observa como el programa de etnoeducacion y desarrollo comunitario presenta el intervalo de confianza para la media de supervivencia de (6,272, 8.630).

Tabla 14 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	34,641	5	,000
Breslow (Generalized Wilcoxon)	38,508	5	,000
Tarone-Ware	37,194	5	,000

En cuanto a los estadísticos de prueba se tiene que ninguno de ellos es significativo, es decir, existen diferencias significativas entre todas las funciones de supervivencia de las licenciaturas.

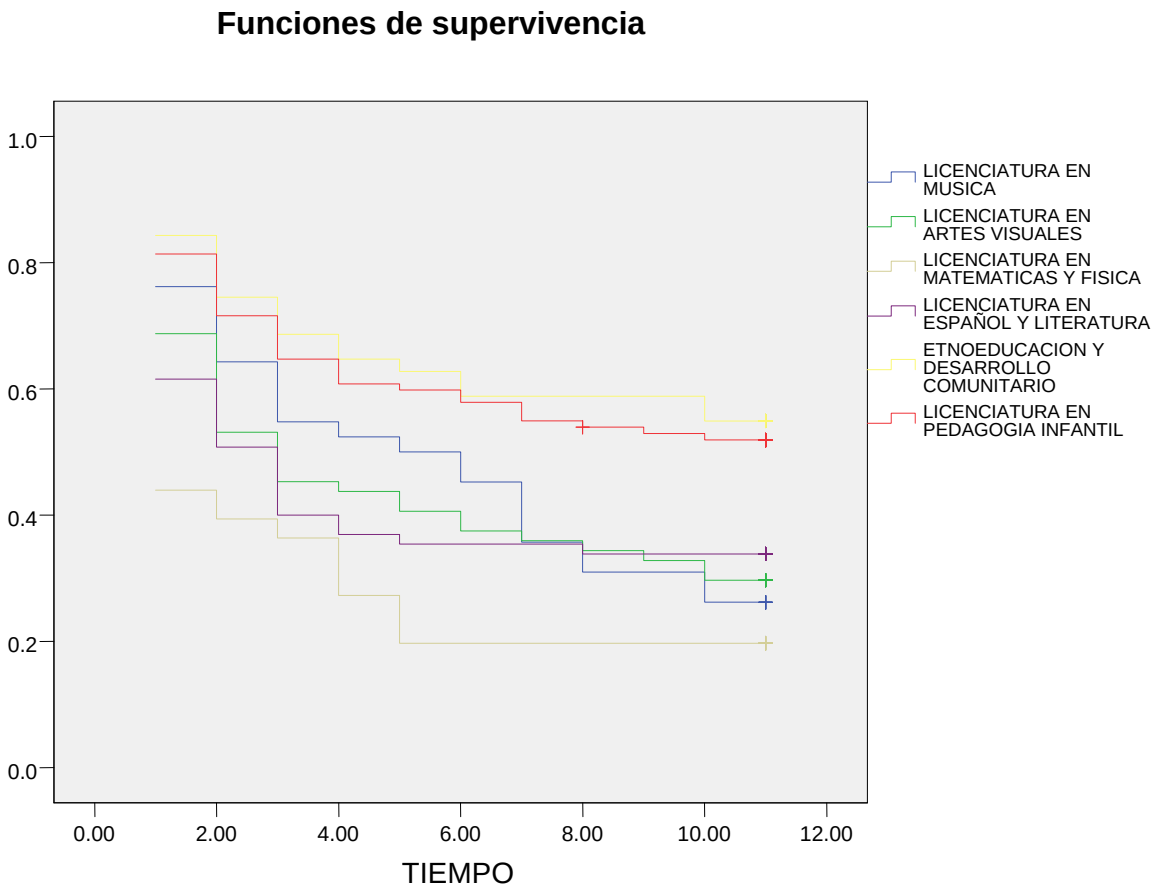

Gráfico 5 Funciones de supervivencia

En la grafica se aprecia una clara diferencia entre las funciones de supervivencia de los programas de licenciaturas, con etnoeducacion como la de mejor nivel de supervivencia.

Por Género

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por género y sus comparaciones:

Tabla 15 Resumen

Genero			Censurado	
	Nº total	Nº de eventos	Nº	Porcentaje
MUJER	240	133	107	44,6%
HOMBRE	150	111	39	26,0%
Global	390	244	146	37,4%

En la tabla anterior se presenta el resumen de casos y sus censuras, a los hombres es el que más ha desertado 74%.

Tabla 16 Medias de los tiempos de supervivencia

	Media			
			Intervalo de confianza al 95%	
Genero	Estimación	Error típico	Límite inferior	Límite superior
MUJER	6,304	,292	5,732	6,875
HOMBRE	4,847	,342	4,177	5,516
Global	5,743	,225	5,302	6,185

En la tabla 16 se muestra una clara diferencia en los tiempos de supervivencia, los intervalos de confianza para la media de supervivencia no se traslapan.

Tabla 17 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	13,021	1	,000
Breslow (Generalized Wilcoxon)	10,655	1	,001
Tarone-Ware	11,866	1	,001

Confirmando los resultados de la tabla 17, en esta se muestran como los tres estadísticos de prueba para la igualdad de las funciones de supervivencia presentan niveles de significancia muy bajos, rechazándose la hipótesis de igualdad de funciones de supervivencia.

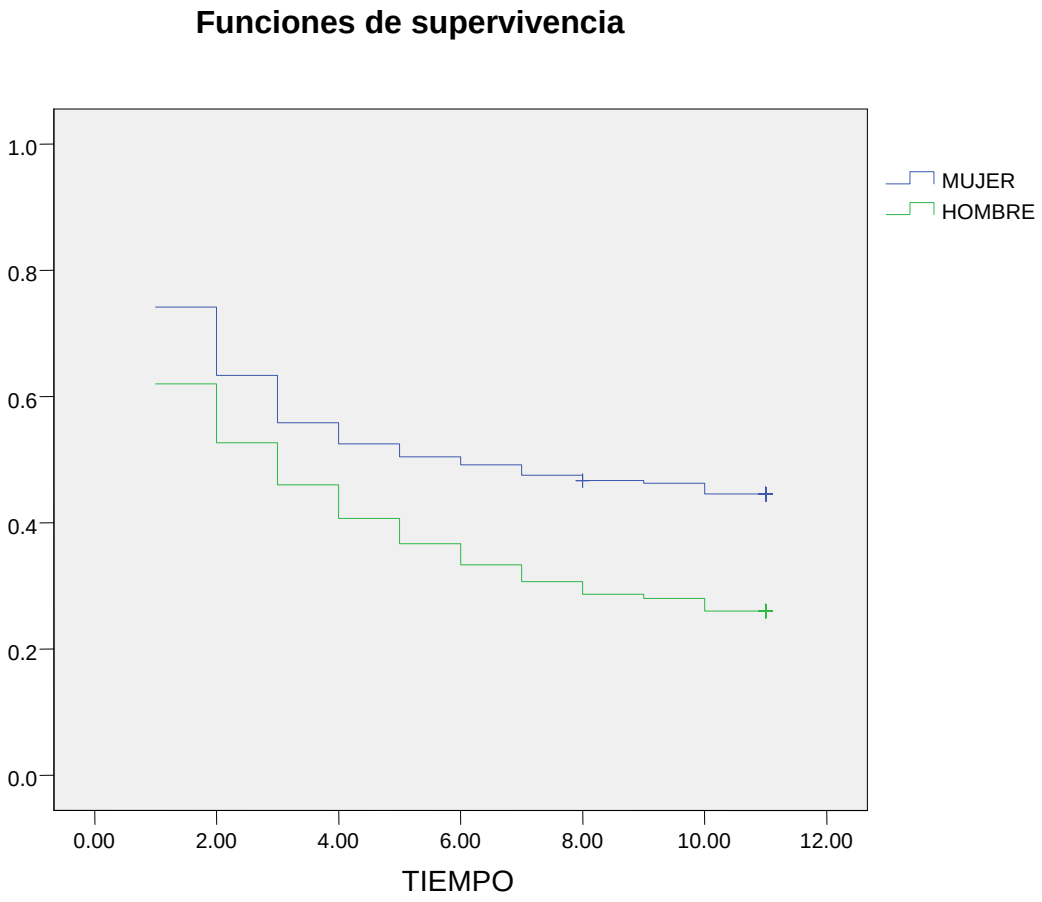

Gráfico 6 Funciones de supervivencia

En la grafica 6 se muestra claramente el mayor nivel de supervivencia que presentan las mujeres de los programas de licenciaturas frente al de los hombres.

Por Colegio

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por tipo de colegio y sus comparaciones:

Tabla 18 Resumen

Tipo de Colegio			Censurado	
	Nº total	Nº de eventos	Nº	Porcentaje
PUBLICO	284	179	105	37,0%
PRIVADO	106	65	41	38,7%
Global	390	244	146	37,4%

En la tabla 19 se muestra el resumen de casos y sus censuras, el fenómeno de deserción es ligeramente superior en los estudiantes que provienen de colegios públicos 63%.

Tabla 19 Medias de los tiempos de supervivencia

Tipo de Colegio	Media(a)			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
PUBLICO	5,735	,263	5,219	6,252
PRIVADO	5,764	,435	4,912	6,617
Global	5,743	,225	5,302	6,185

En la tabla 20 se observa como las tiempos de supervivencia para los estudiantes provenientes de colegios públicos y privados se asemejan, intervalos de confianza traslapados.

Tabla 20 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	,038	1	,846
Breslow (Generalized Wilcoxon)	,003	1	,960
Tarone-Ware	,005	1	,944

Confirmando los resultados de la tabla 20 se muestra en la tabla 21 que no existen diferencias significativas entre las dos funciones de supervivencias.

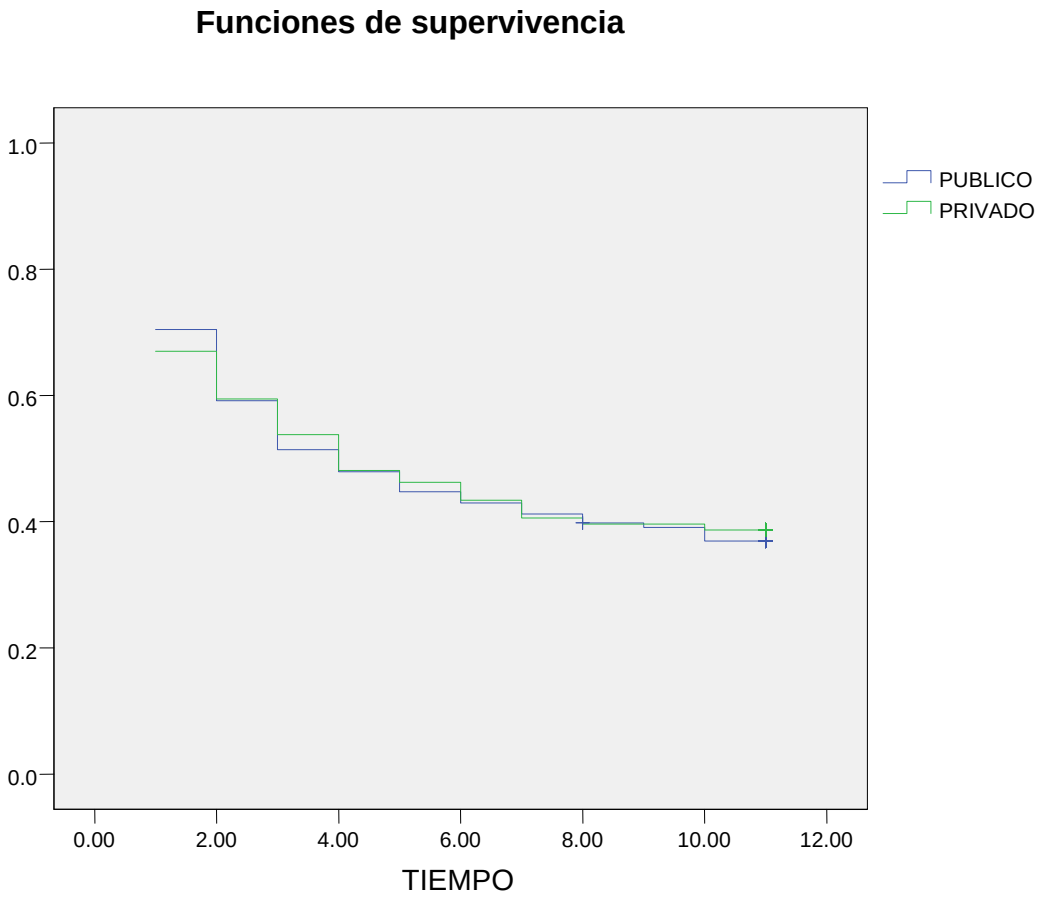

Gráfico 7 Funciones de supervivencia

En la grafica anterior se observa como las dos curvas de supervivencia están casi colineales, es decir, no existen diferencias significativas entre las dos funciones de supervivencia.

TECNOLOGIAS

Por Programa

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por programa y sus comparaciones:

Tabla 21 Resumen

Programa	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
TECNOLOGIA ELECTRICA	69	54	15	21,7%
TECNOLOGIA INDUSTRIAL	71	36	35	49,3%
TECNOLOGIA MECANICA	70	47	23	32,9%
TECNOLOGIA QUIMICA	72	38	34	47,2%
Global	282	175	107	37,9%

De la tabla 21 se tiene que la tecnología con mayor deserción es tecnología eléctrica con 78,3% de desertores seguida por tecnología mecánica con 67,1%.

Tabla 22 Medias de los tiempos de supervivencia

Programa	Media			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
TECNOLOGIA ELECTRICA	4,503	,490	3,544	5,463
TECNOLOGIA INDUSTRIAL	7,335	,469	6,416	8,254
TECNOLOGIA MECANICA	5,358	,529	4,321	6,396
TECNOLOGIA QUIMICA	6,737	,517	5,724	7,750
Global	5,996	,259	5,488	6,505

De la tabla anterior se concluye que la tecnología con mayor nivel de supervivencia es tecnología industrial seguida por tecnología química, ver intervalos de confianza.

Tabla 23 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	18,830	3	,000
Breslow (Generalized Wilcoxon)	20,675	3	,000
Tarone-Ware	19,966	3	,000

Se ve de la tabla 23 que existen diferencias significativas entre las funciones de supervivencia de los programas de tecnología. En la grafica 9 se muestra como tecnología industrial es la que tiene un mayor nivel de supervivencia contra tecnología eléctrica que es la de peor nivel de supervivencia.

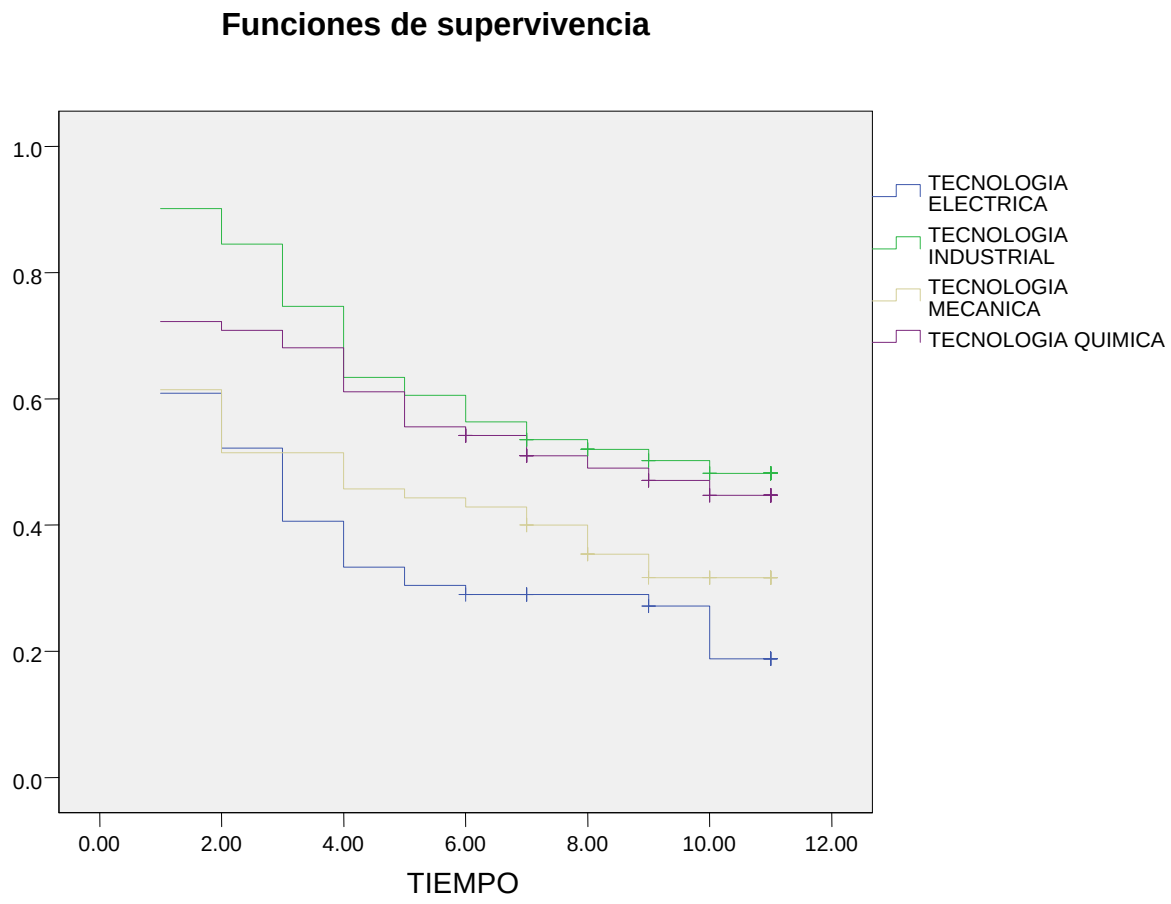

Gráfico 8 Funciones de supervivencia

Por Género

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por género y sus comparaciones:

Tabla 24 Resumen

Genero			Censurado	
	Nº total	Nº de eventos	Nº	Porcentaje
MUJER	104	43	61	58,7%
HOMBRE	178	132	46	25,8%
Global	282	175	107	37,9%

Los hombres son los que tienen la mayor cantidad de desertores dentro de las tecnologías 64,2%.

Tabla 25 Medias de los tiempos de supervivencia

sexo	Media			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
MUJER	7,637	,414	6,825	8,449
HOMBRE	5,049	,311	4,440	5,659
Global	5,996	,259	5,488	6,505

En la tabla anterior se muestra una clara distinción entre los tiempos de supervivencia entre hombres y mujeres de los programas de tecnología, las mujeres presentan un intervalo de confianza para la media de supervivencia mucho mayor que el de los hombres.

Tabla 26 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	26,692	1	,000
Breslow (Generalized Wilcoxon)	21,801	1	,000
Tarone-Ware	24,413	1	,000

En la tabla 26 se muestran los estadísticos de prueba para la igualdad de funciones de supervivencia, se presentan diferencias bastante significativas entre las funciones de supervivencia entre hombres y mujeres.

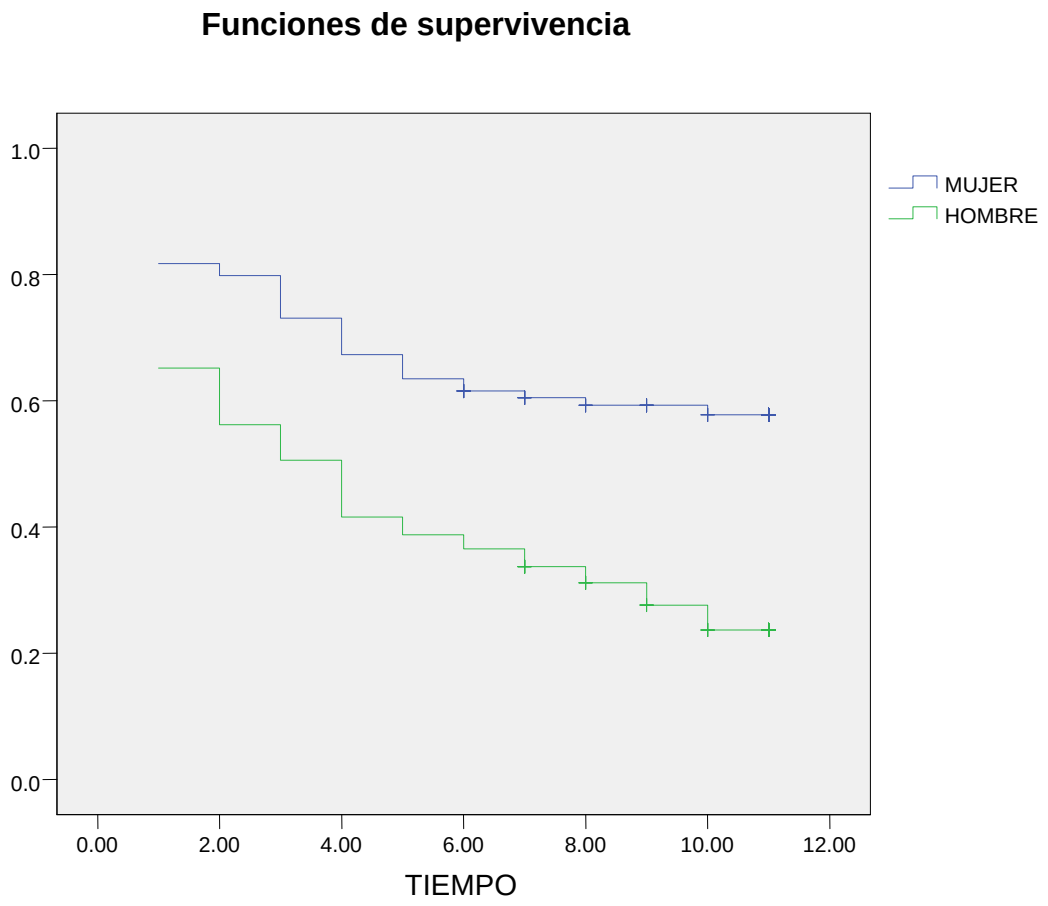

Gráfico 9 Funciones de supervivencia

En la grafica 9 se ve claramente la mejor situación de supervivencia de las mujeres con respecto de los hombres.

Por Colegio

A continuación se muestra los cuadros de salida del software SPSS 15.0 que muestran las diferentes estadísticas y curvas de supervivencia por tipo de colegio y sus comparaciones:

Tabla 27 Resumen

Tipo de Colegio	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
PUBLICO	193	119	74	38,3%
PRIVADO	89	56	33	37,1%
Global	282	175	107	37,9%

De la tabla anterior se ve que la cantidad de desertores de los estudiantes que vienen de colegios privados es ligeramente superior (62,9% privados contra 61,7 de los públicos).

Tabla 28 Medias de los tiempos de supervivencia

Tipo de Colegio	Media			
			Intervalo de confianza al 95%	
	Estimación	Error típico	Límite inferior	Límite superior
PUBLICO	6,004	,315	5,386	6,623
PRIVADO	5,973	,456	5,080	6,865
Global	5,996	,259	5,488	6,505

Confirmando los resultados de la tabla 27, la tabla 28 muestra los tiempos de supervivencias para los estudiantes provenientes de colegios públicos y privados se asemejan, intervalos de confianza muy traslapados.

Tabla 29 Comparaciones Globales

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	,026	1	,872
Breslow (Generalized Wilcoxon)	,000	1	,993
Tarone-Ware	,006	1	,936

Como se muestra en la tabla anterior, no se presentan diferencias significativas entre las funciones de supervivencia de los dos tipos de estudiantes. En la grafica 10 se aprecia las curvas de supervivencia de las dos poblaciones.

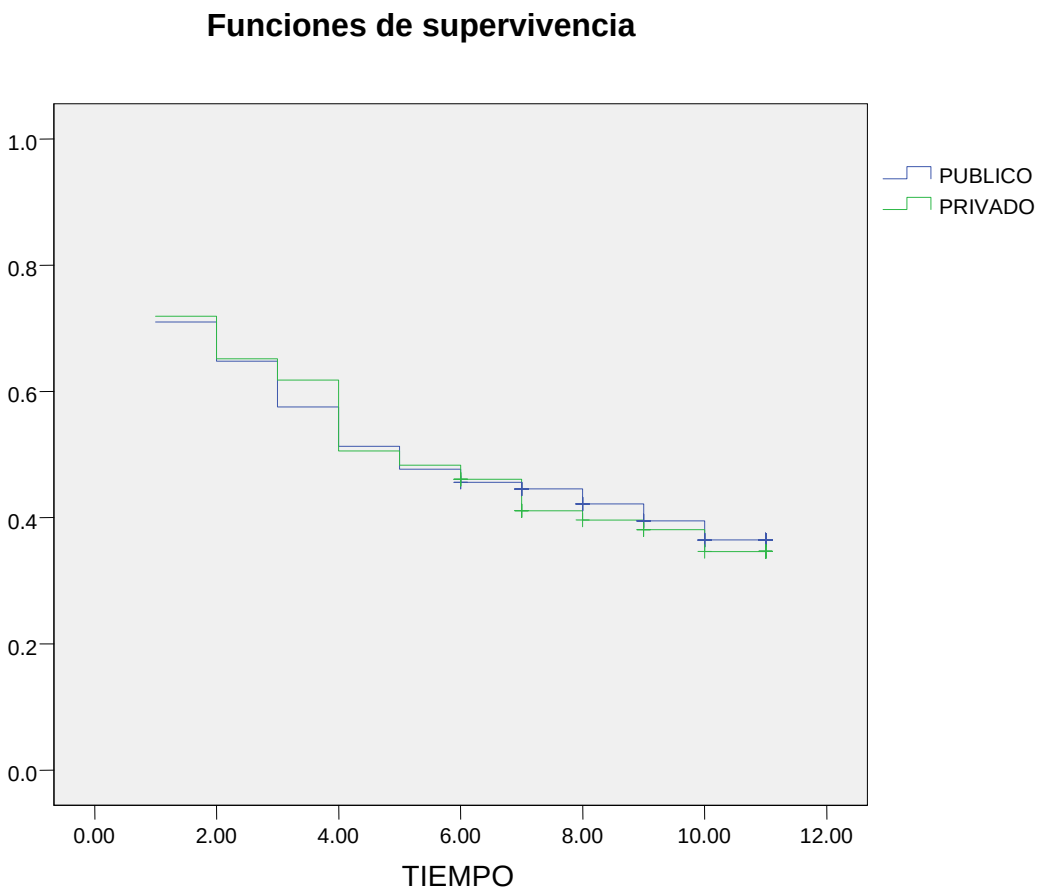

Gráfico 10 Funciones de supervivencia

7. MODELO DE REGRESION DE COX

Los modelos de Cox aquí utilizados tienen la finalidad de estimar el efecto de las variables de estudio sobre los tiempos de supervivencia de los estudiantes de los diferentes programas académicos. En la sección anterior lo que se hizo fue hacer un análisis exploratorio de la situación de supervivencia de los estudiantes de los programas de ingenierías, licenciaturas y tecnologías; en esta sección se pretende llegar un poco más lejos en el sentido de poder estimar un modelo que permita explicar la supervivencia de los estudiantes (en el sentido de la sección anterior) a partir de una serie de variables que se han podido recolectar de la matricula académica de los estudiantes.

Este capítulo se desarrolla de la siguiente manera, inicialmente se va a procede con la obtención del mejor conjunto de variables explicativas del modelo teniendo en cuenta el principio de parsimonia y utilizando la prueba de puntajes (ecuación 61). En un paso posterior se calcula un modelo de regresión de COX por tipo de programa académico con su respectiva curva de supervivencia. Por último se realiza el cálculo de un modelo de COX estratificado al interior de cada tipo de programa obteniéndose un modelo definitivo para ingenierías, licenciaturas y tecnologías.

Las variables a analizar son las disponibles en la base de datos de este trabajo (EDAD2, sexo, depto, estcivil, naturalezacol y estrato). Los cálculos fueron desarrollados en el software SAS 9.0.

INGENIERIAS

Elección Variables

Como se había enunciado ya inicialmente se escogerá un modelo apropiado que contenga el número mínimo de variables pero mejor explicativas posible del conjunto de variables en la base (EDAD2, sexo, depto, estcivil, naturalezacol y estrato). Inicialmente se calcula la prueba de puntajes (ecuación 61) para todos los posibles modelos que contengan una variable:

Cuadro 1. Puntuaciones Variable(1)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
1	11.6791	EDAD2
1	5.0187	estcivil
1	3.7738	sexo
1	0.6135	naturalezacol
1	0.5473	depto
1	0.4413	estrato

Del cuadro 1 se observa que el modelo con la variable EDAD2 (edad) es el mejor modelo con una sola variable con una puntuación de 11,68.

Cuadro 2. Puntuaciones Variables(2)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
2	16.1098	EDAD2 estcivil
2	14.0575	EDAD2 sexo
2	12.4316	EDAD2 naturalezacol
2	11.7984	EDAD2 depto
2	11.7387	EDAD2 estrato
2	8.5550	sexo estcivil
2	5.5032	depto estcivil
2	5.3753	estrato estcivil
2	5.3427	naturalezacol estcivil
2	4.1686	sexo naturalezacol
2	4.1554	sexo depto
2	4.1534	sexo estrato
2	1.2053	naturalezacol depto
2	0.8556	estrato depto
2	0.7823	naturalezacol estrato

En el cuadro anterior se nota que los dos mejores modelos (según la prueba de puntuación) son los que contienen EDAD2(edad) y estcivil(estado civil) con una puntuación de 16,11; y el que contiene a EDAD2(edad) y sexo(genero) con una puntuación de 14,06.

Cuadro 3. Puntuaciones Variables(3)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
3	18.3399	EDAD2 sexo estcivil
3	16.5518	EDAD2 naturalezacol estcivil
3	16.2083	EDAD2 depto estcivil
3	16.1462	EDAD2 estrato estcivil
3	14.6050	EDAD2 sexo naturalezacol
3	14.1295	EDAD2 sexo depto
3	14.1097	EDAD2 sexo estrato
3	12.5726	EDAD2 naturalezacol depto
3	12.4376	EDAD2 naturalezacol estrato
3	11.8369	EDAD2 estrato depto
3	8.8901	sexo depto estcivil
3	8.8597	sexo estrato estcivil
3	8.7339	sexo naturalezacol estcivil
3	5.8595	naturalezacol depto estcivil
3	5.7477	estrato depto estcivil
3	5.5205	naturalezacol estrato estcivil
3	4.5845	sexo naturalezacol depto

Un razonamiento parecido al del párrafo anterior muestra que el modelo con tres variables, con mejor puntuación, es el que contiene a EDAD2(edad), sexo(genero) y estcivil(estado civil) con una puntuación de 18,34. Se puede concluir del cuadro 4 que una variable mas no genera una puntuación muy diferente a la ya obtenida con el modelo de las 3 variables.

Cuadro 4. Puntuación variables(4)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
4	18.6351	EDAD2 sexo naturalezacol estcivil
4	18.3973	EDAD2 sexo depto estcivil
4	18.3710	EDAD2 sexo estrato estcivil
4	16.6660	EDAD2 naturalezacol depto estcivil
4	16.5548	EDAD2 naturalezacol estrato estcivil
4	16.2300	EDAD2 estrato depto estcivil
4	14.6934	EDAD2 sexo naturalezacol depto
4	14.6070	EDAD2 sexo naturalezacol estrato
4	14.1663	EDAD2 sexo estrato depto
4	12.5930	EDAD2 naturalezacol estrato depto
4	9.1090	sexo estrato depto estcivil
4	9.0916	sexo naturalezacol depto estcivil
4	8.9182	sexo naturalezacol estrato estcivil
4	5.9467	naturalezacol estrato depto estcivil
4	4.6783	sexo naturalezacol estrato depto

Modelo de Regresión de COX

Seleccionado el modelo con las tres variables EDAD2 (edad), sexo(genero) y estcivil (estado civil) se procede a la estimación de los parámetros:

Cuadro 5. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
EDAD2	1	0.09136	0.03043	9.0128	0.0027	1.096	EDAD2
sexo	1	0.29013	0.18998	2.3323	0.1267	1.337	sexo
estcivil	1	-0.73362	0.36069	4.1368	0.0420	0.480	estcivil

En el cuadro anterior *estimador del parámetro* se refiere al peso que tiene la variable en el modelo de supervivencia, si las estimaciones son positivas quiere decir que la variable en cuestión contribuye positivamente al riesgo de desertar, cuando las estimaciones son

negativas implica que la variable en cuestión contribuye negativamente al riesgo de desertar; en este cuadro se muestran las estimaciones de los parámetros correspondientes a las variables edad, sexo y estado civil. Como se puede ver los coeficientes de la variable edad y sexo son positivos 0,09136 y 0,29013 respectivamente.. La variable de estado civil es negativa -0,73362. El Ratio del riesgo se interpreta como el riesgo relativo que tiene un individuo con respecto a otro, de estos valores se tiene que por un año adicional de un estudiante al momento de entrar a la universidad aumenta el riesgo de desertar en 9,6% (109,6%-100%). En cuanto el ratio del riesgo para la variable sexo se tiene que los hombres tienen 1,337 veces mayor riesgo de desertar que las mujeres (recordar que sexo esta codificado como 1=hombre, 0= mujer). Para estado civil se tiene los solteros tienen un poco menos de la mitad del riesgo de desertar de los que no son solteros (recordar que estado civil esta codificado 1=soltero 0=otro).

Cuadro 6. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	16.0694	3	0.0011
Puntuación	18.3399	3	0.0004
Wald	17.3891	3	0.0006

En el cuadro anterior se muestran tres pruebas de ajuste global del modelo. Chi-cuadrado es el estadístico de prueba y Pr > ChiSq es el valor p de la prueba, lo cual y teniendo en cuenta un nivel de significancia de 5% de las tres pruebas debe ser rechazada la hipótesis de la igualdad de los coeficientes a cero.

A continuación se muestra la curva de supervivencia para este modelo:

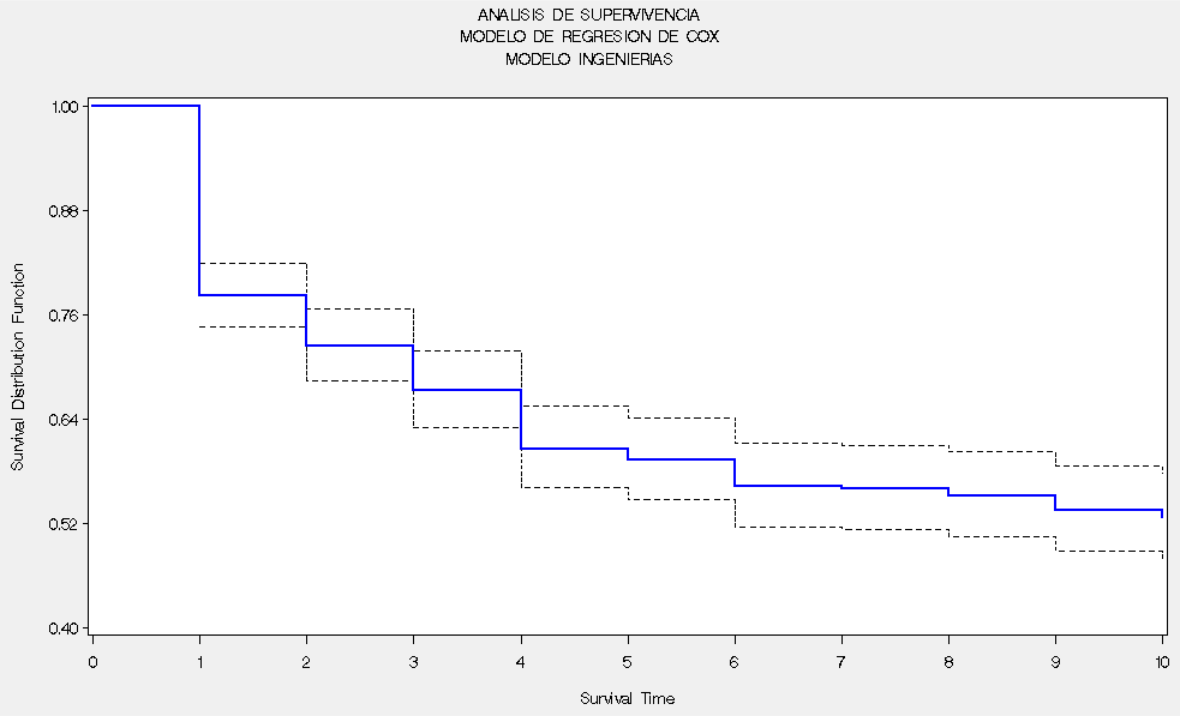

Gráfico 11 Función de Supervivencia

INGENIERÍAS ESTRATIFICADO

Realizando el mismo análisis anterior pero teniendo en cuenta las diferencias presentes entre los programas de ingenierías (cada programa es un estrato) se tiene que el modelo con las variables edad, sexo y estado civil tiene los siguientes coeficientes:

Cuadro 7. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
EDAD2	1	0.08562	0.03062	7.8193	0.0052	1.089	EDAD2
sexo	1	0.21355	0.19629	1.1836	0.2766	1.238	sexo
estcivil	1	-0.64804	0.36230	3.1994	0.0737	0.523	estcivil

Como se ve del cuadro anterior el modelo resultante es muy similar al del modelo anterior, para la variable edad el riesgo por cada año adicional al momento de entrar a la

universidad aumenta en un 8,9%, para los hombres el riego aumenta en 1,238 veces y para los solteros el riesgo disminuye a un 52,3% del riesgo de los que no son solteros.

Cuadro 8. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	12.4348	3	0.0060
Puntuación	13.9938	3	0.0029
Wald	13.2796	3	0.0041

Al igual que en el modelo anterior, en este los estadísticos muestran una buena adecuación del modelo.

En la grafica 12 se muestran las funciones de supervivencia para los programas de ingenierías. Al igual que en las estimaciones de Kaplan-Meier realizadas en la sección anterior, el programa de ingeniería industrial es el de mejor curva de supervivencia seguido por el programa de ingeniería de sistemas.

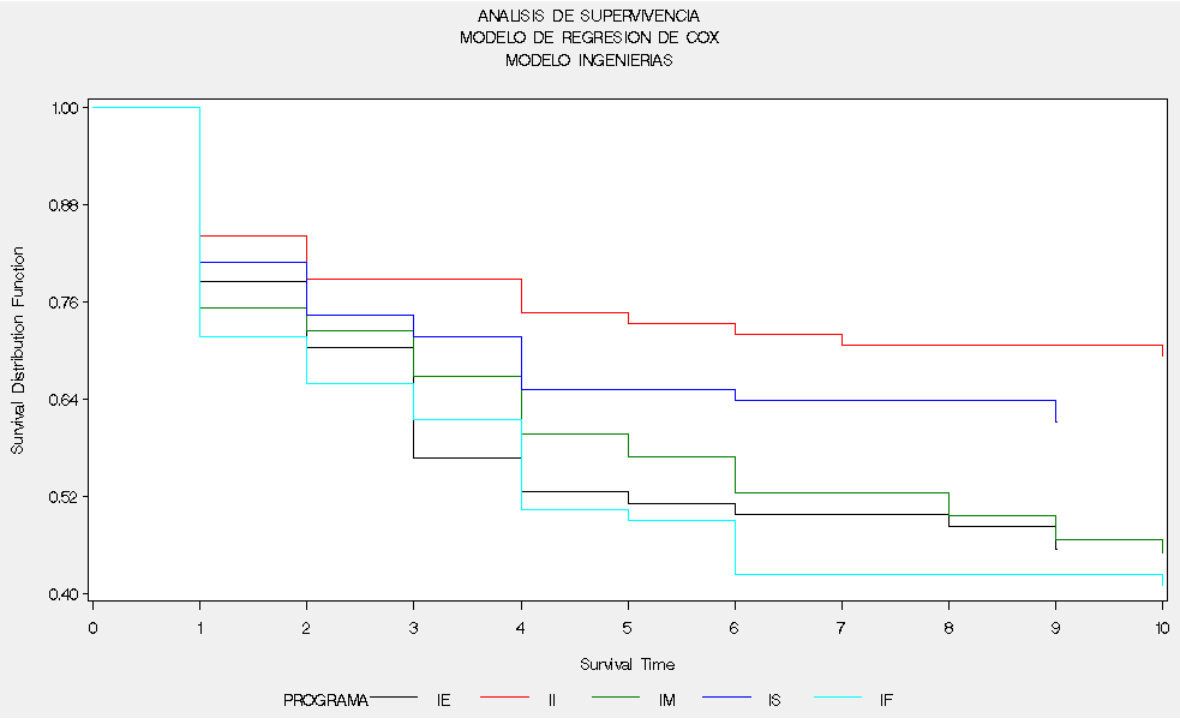

Gráfico 12 Funciones de Supervivencia

LICENCIATURAS

Elección de variables

Se calcula un modelo apropiado con el menor número de variables posible. Inicialmente se muestran los mejores modelos de una sola variable:

Cuadro 9. Puntuación variables(1)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
1	13.0208	sexo
1	3.3802	EDAD2
1	2.2863	estrato
1	2.2533	estcivil
1	0.0378	naturalezacol
1	0.0044	depto

Del cuadro anterior se ve que el mejor modelo de una sola variable es el que contiene la variable sexo con una puntuación de 13,02.

Cuadro 10. Puntuación variables(2)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
2	15.6051	sexo estrato
2	14.9580	EDAD2 sexo
2	14.6274	sexo estcivil
2	13.1459	sexo naturalezacol
2	13.0239	sexo depto
2	7.0958	EDAD2 estcivil
2	6.4735	EDAD2 estrato
2	3.9173	estrato estcivil
2	3.6391	EDAD2 naturalezacol
2	3.3839	EDAD2 depto
2	2.4232	naturalezacol estrato
2	2.3052	estrato depto
2	2.2832	depto estcivil
2	2.2828	naturalezacol estcivil
2	0.0456	naturalezacol depto

El mejor modelo de dos variables es el que contiene sexo y estrato con una puntuación de 15,61.

Cuadro 11. Puntuación variables(3)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
3	18.1777	EDAD2 sexo estrato
3	17.5571	EDAD2 sexo estcivil
3	16.6496	sexo estrato estcivil
3	15.9006	sexo naturalezacol estrato
3	15.6222	sexo estrato depto
3	15.3043	EDAD2 sexo naturalezacol
3	14.9597	EDAD2 sexo depto
3	14.7369	sexo naturalezacol estcivil
3	14.6487	sexo depto estcivil
3	13.1549	sexo naturalezacol depto
3	9.3752	EDAD2 estrato estcivil
3	7.3968	EDAD2 naturalezacol estcivil
3	7.0984	EDAD2 depto estcivil
3	7.0513	EDAD2 naturalezacol estrato
3	6.4736	EDAD2 estrato depto
3	4.0241	naturalezacol estrato estcivil
3	3.9655	estrato depto estcivil
3	3.6392	EDAD2 naturalezacol depto
3	2.4560	naturalezacol estrato depto
3	2.3200	naturalezacol depto estcivil

El mejor modelo de tres variables es el que contiene edad, sexo y estrato con una puntuación de 18,18. En el cuadro 12 se ve que una variable adicional contribuye poco al modelo.

Cuadro 12. Puntuación variables(4)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
4	20.0752	EDAD2 sexo estrato estcivil
4	18.8936	EDAD2 sexo naturalezacol estrato
4	18.1785	EDAD2 sexo estrato depto
4	17.9394	EDAD2 sexo naturalezacol estcivil
4	17.5598	EDAD2 sexo depto estcivil
4	16.9054	sexo naturalezacol estrato estcivil
4	16.6884	sexo estrato depto estcivil
4	15.9384	sexo naturalezacol estrato depto
4	15.3046	EDAD2 sexo naturalezacol depto
4	14.7703	sexo naturalezacol depto estcivil
4	9.9573	EDAD2 naturalezacol estrato estcivil
4	9.3846	EDAD2 estrato depto estcivil
4	7.4081	EDAD2 naturalezacol depto estcivil
4	7.0589	EDAD2 naturalezacol estrato depto
4	4.0909	naturalezacol estrato depto estcivil

Modelo de Regresión de COX

Seleccionado el modelo con las tres variables EDAD2(edad), sexo(genero) y estrato(estrato) se procede a la estimación de los parámetros:

Cuadro 13. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
EDAD2	1	0.02109	0.01316	2.5660	0.1092	1.021	EDAD2
sexo	1	0.49904	0.14644	11.6140	0.0007	1.647	sexo
estrato	1	0.14631	0.08184	3.1963	0.0738	1.158	estrato

En el cuadro anterior se muestran las estimaciones de los parámetros correspondientes a las variables edad, sexo y estrato. Como se puede ver los coeficientes de la variable edad y sexo son positivos 0,02109 y 0,49904 respectivamente. Del ratio del riesgo se tiene que por un año adicional de un estudiante al momento de entrar a la universidad aumenta el riesgo de desertar en 2,1% (102,1%-100%). En cuanto el ratio del riesgo para la variable sexo se tiene que los hombres tienen 1,647 veces mayor riesgo de desertar que las mujeres (recordar que sexo esta codificado como 1=hombre, 0= mujer). Para estrato se tiene que por cada aumento de estrato aumenta el riesgo en 15,8% (115,8%-100%).

Cuadro 13. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	17.6575	3	0.0005
Puntuación	18.1777	3	0.0004
Wald	17.8809	3	0.0005

La hipótesis de que los betas sean iguales a cero, debe ser rechazada a un nivel de significancia del 5%. A continuación se muestra la curva de supervivencia para este modelo:

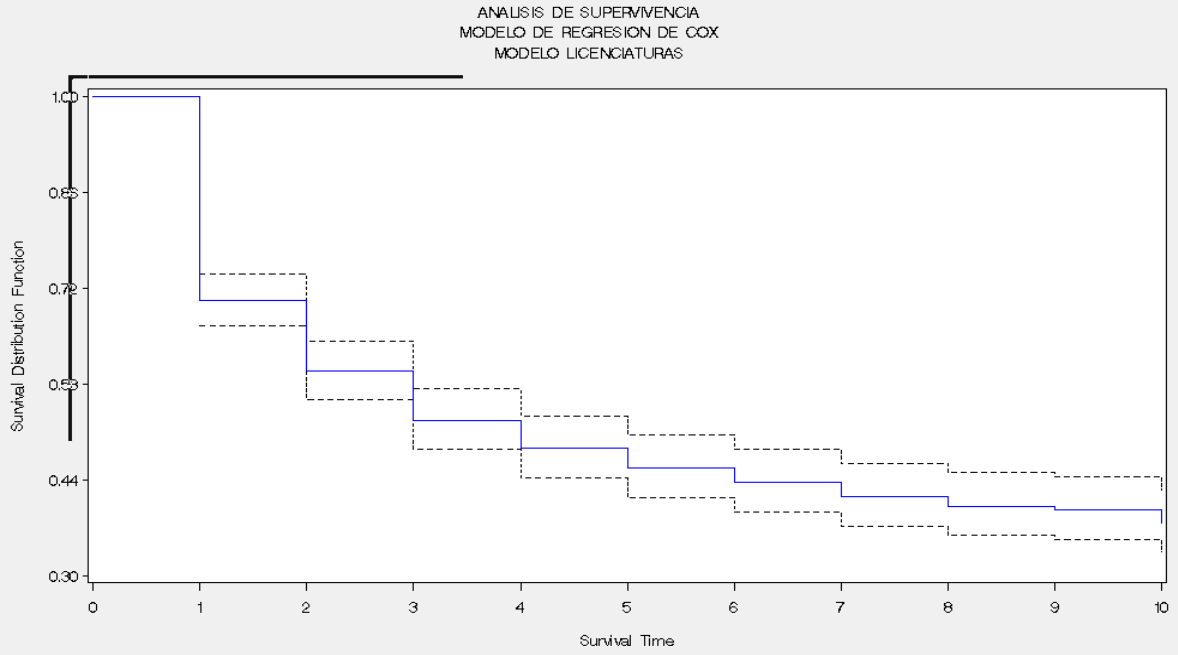

Gráfico 13 Funciones de Supervivencia

LICENCIATURAS ESTRATIFICADO

Realizando el mismo análisis anterior pero teniendo en cuenta las diferencias presentes entre los programas de licenciaturas se tiene que el modelo con las variables edad, sexo y estrato tiene los siguientes coeficientes:

Cuadro 14. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
EDAD2	1	0.03115	0.01463	4.5325	0.0333	1.032	EDAD2
sexo	1	0.17820	0.17113	1.0844	0.2977	1.195	sexo
estrato	1	0.15996	0.08422	3.6078	0.0575	1.173	estrato

Para la variable edad el riesgo por cada año adicional al momento de entrar a la universidad aumenta en un 3,2%, para los hombres el riego aumenta en 1,195 veces y para el aumento de estrato se tiene un aumento del riesgo del 17,3%.

Cuadro 15. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	8.7279	3	0.0331
Puntuación	9.0147	3	0.0291
Wald	8.9150	3	0.0304

La hipótesis de coeficientes iguales a cero sigue siendo rechazada como en el modelo anterior pero con p-valores más grandes. En la grafica siguiente se muestran las curvas de supervivencia.

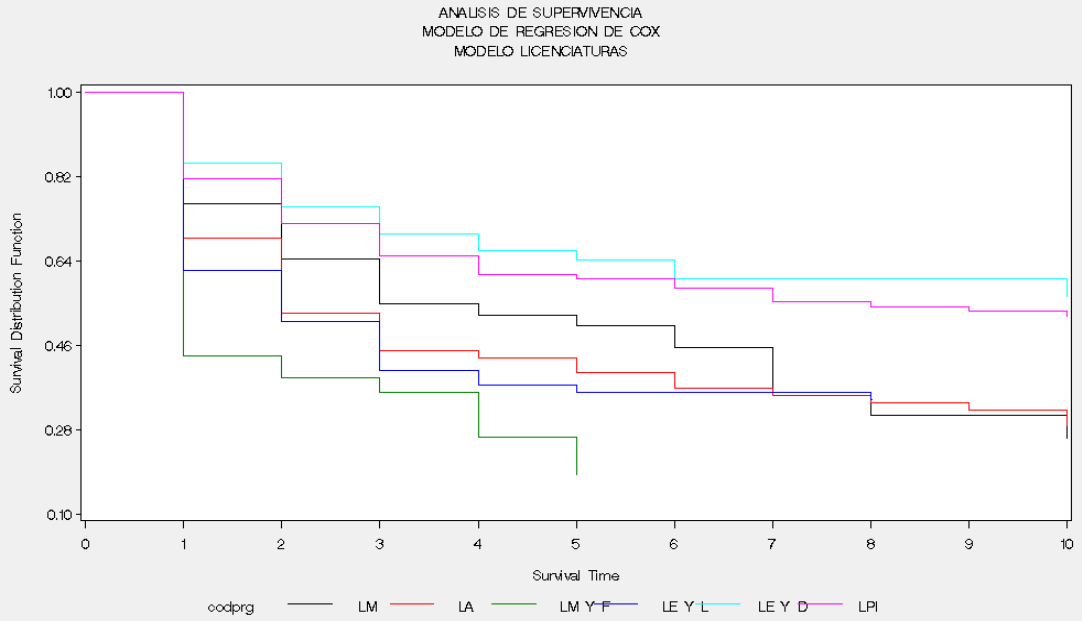

Gráfico 14 Funciones de Supervivencia

TECNOLOGIAS

Elección de variables

Se calcula un modelo apropiado con el menor número de variables posible. Inicialmente se muestran los mejores modelos de una sola variable:

Cuadro 16. Puntuación variables(1)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
1	26.6919	sexo
1	9.3474	depto
1	8.5542	estcivil
1	6.9097	EDAD2
1	1.2511	estrato
1	0.0258	naturalezacol

El mejor modelo de una sola variable es claramente el modelo con la variable sexo con una puntuación de 26,69.

Cuadro 17. Puntuación variables(2)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
2	33.6177	sexo depto
2	31.0733	sexo estcivil
2	27.7893	EDAD2 sexo
2	27.1163	sexo estrato
2	26.7969	sexo naturalezacol
2	17.5590	depto estcivil
2	16.0128	EDAD2 depto
2	14.9783	EDAD2 estcivil
2	9.9418	estrato depto
2	9.3637	naturalezacol depto
2	9.3131	estrato estcivil
2	8.6404	naturalezacol estcivil
2	7.4710	EDAD2 estrato
2	6.9402	EDAD2 naturalezacol
2	1.4570	naturalezacol estrato

El mejor modelo de dos variables como se ve en el cuadro 17 es el que contiene la variable sexo y depto con una puntuación de 33,62.

Cuadro 18. Puntuación variables(3)

Número de variables	Chi-cuadrado de puntuación	Variables incluidas en el modelo
3	37.9378	sexo depto estcivil
3	34.7816	EDAD2 sexo depto
3	33.7531	sexo estrato depto
3	33.6224	sexo naturalezacol depto
3	32.2676	EDAD2 sexo estcivil
3	31.3258	sexo estrato estcivil
3	31.2410	sexo naturalezacol estcivil
3	28.0683	EDAD2 sexo estrato
3	27.8909	EDAD2 sexo naturalezacol
3	27.3671	sexo naturalezacol estrato
3	23.7611	EDAD2 depto estcivil
3	17.8423	estrato depto estcivil
3	17.5591	naturalezacol depto estcivil
3	16.1783	EDAD2 estrato depto
3	16.0249	EDAD2 naturalezacol depto
3	15.2469	EDAD2 estrato estcivil

El modelo de tres variables que mejor se ajusta según el criterio de puntuación chi-cuadrado es el que contiene las variables sexo, depto y estcivil con una puntuación de 37,94. Ver cuadro anterior.

Modelo de Regresión de COX

Seleccionado el modelo con las tres variables sexo(genero), depto(departamento), y estcivil(estado civil) se procede a la estimación de los parámetros:

Cuadro 19. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
sexo	1	0.85425	0.19120	19.9608	<.0001	2.350	sexo
depto	1	-0.61453	0.23999	6.5571	0.0104	0.541	depto
estcivil	1	0.76085	0.34539	4.8525	0.0276	2.140	estcivil

En el cuadro anterior se muestran las estimaciones de los parámetros correspondientes a las variables sexo, depto y estcivil. Como se puede ver los coeficientes de la variable sexo, y estcivil son positivos 0,85425 y 0,76085 respectivamente, el coeficiente asociado a la variable depto es negativo –0,61453. Del ratio del riesgo se tiene que los hombres tienen 2,35 veces más riesgo de desertar que las mujeres, los solteros tienen 2,14 veces más riesgo de desertar que los no solteros. Los estudiantes provenientes de Risaralda tienen 45,9%(1-0,541) menor riesgo de desertar que de los que no son de Risaralda.

Cuadro 20. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	39.8579	3	<.0001
Puntuación	37.9378	3	<.0001
Wald	35.4436	3	<.0001

Como se aprecia en el cuadro anterior, la hipótesis de que los coeficientes asociados a las variables sean iguales a cero debe ser rechazada. En la grafica siguiente se muestran la curva de supervivencia.

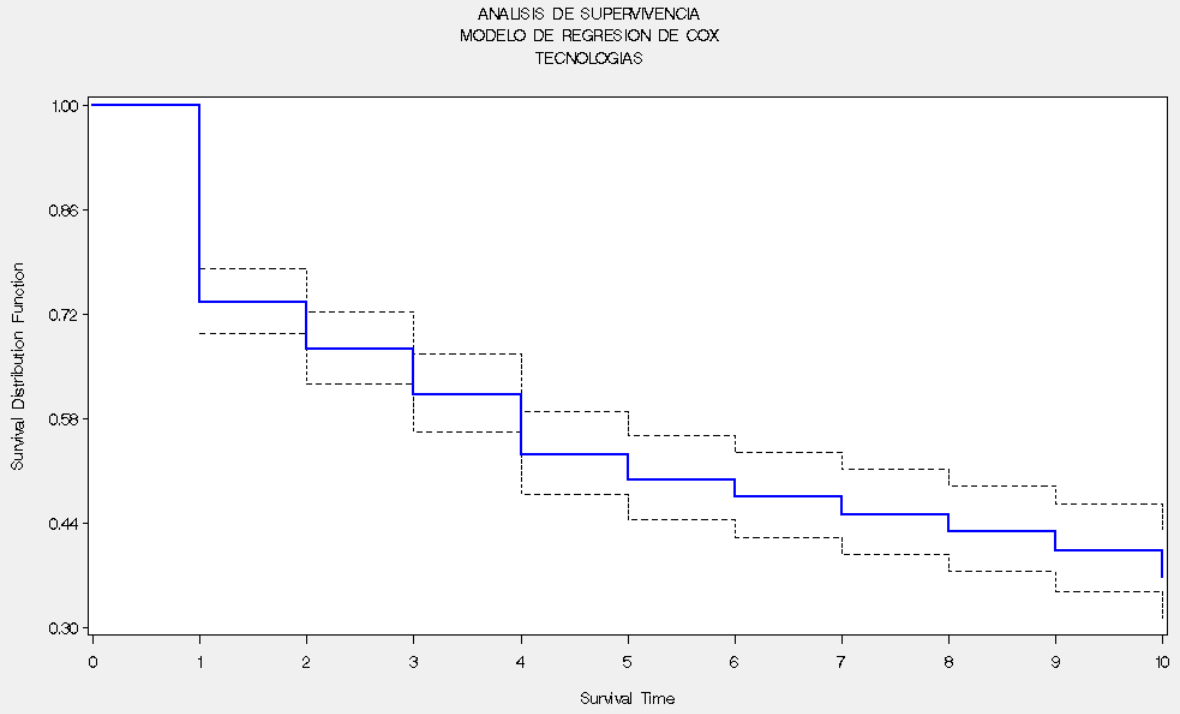

Gráfico 15 Función de Supervivencia

TECNOLOGIAS ESTRATIFICADO

Teniendo en cuenta las diferencias presentes entre los programas de tecnologías se tiene que el modelo con las variables sexo, depto y estcivil tiene los siguientes coeficientes:

Cuadro 21. Estimaciones de los parámetros

Análisis del estimador de máxima verosimilitud

Variable	DF	Estimador del parámetro	Error estándar	Chi-cuadrado	Pr > ChiSq	Ratio del riesgo	Etiqueta de la variable
sexo	1	0.75347	0.22182	11.5380	0.0007	2.124	sexo
depto	1	-0.67597	0.25025	7.2964	0.0069	0.509	depto
estcivil	1	0.70455	0.35709	3.8928	0.0485	2.023	estcivil

El análisis estratificado presenta el mismo patrón de comportamiento del no estratificado con algunas diferencias en los coeficientes y en los ratios de riesgo.

Cuadro 22. Estadísticos de ajuste

Probar hipótesis nula global: BETA=0

Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	27.0986	3	<.0001
Puntuación	26.9684	3	<.0001
Wald	25.6928	3	<.0001

Al igual que en el modelo no estratificado la hipótesis de coeficientes nulos debe ser rechazada.

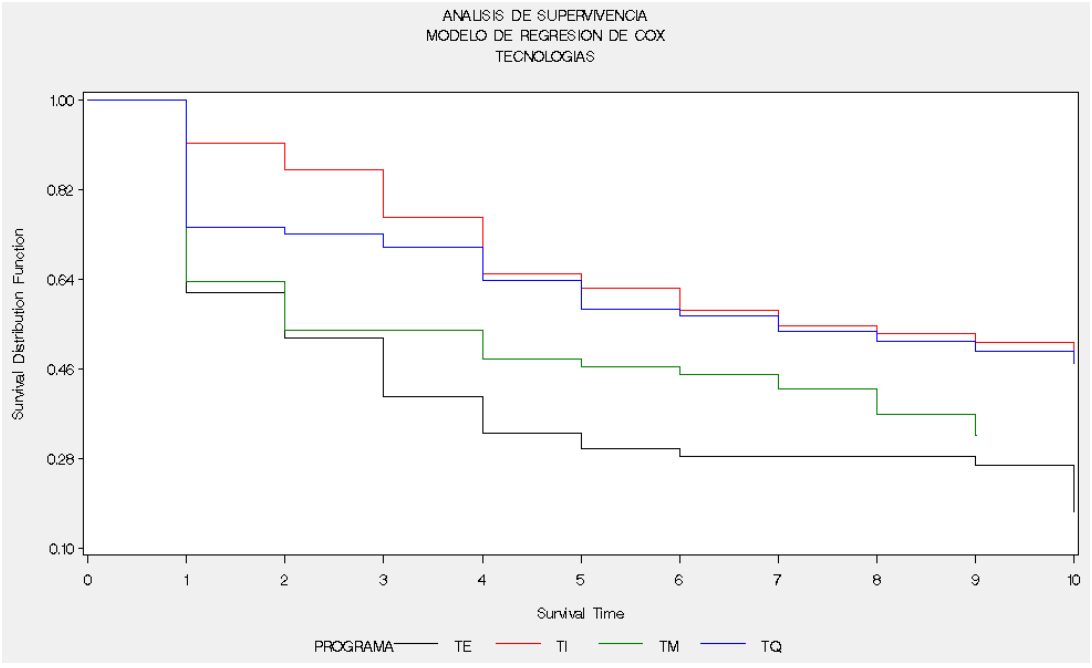

Gráfico 16 Funciones de Supervivencia

8. CONCLUSIONES Y RECOMENDACIONES

8.1 CONCLUSIONES

Modelo Kaplan-Meier

Ingenierías

- Dentro de los programas de ingenierías, se nota una clara superioridad en términos de supervivencia universitaria de los programas de ingeniería industrial y de ingeniería de sistemas.

- Los programas de ingeniería mecánica y eléctrica tienen niveles de supervivencia similares, el programa de ingeniería física tiene el peor nivel de supervivencia.

- En los programas de ingenierías las mujeres tienen mejor nivel de supervivencia que los hombres.

- Dentro de las ingenierías los estudiantes que provienen de colegios privados tienen una ligera superioridad en cuanto al nivel de supervivencia de los estudiantes que provienen de colegios públicos.

Licenciaturas

- Los programas de licenciatura en etnoeducion y desarrollo comunitario y licenciatura en pedagogía infantil son los programas de mejor nivel de supervivencia.

- El programa de licenciatura en matemáticas y física es el de mayor nivel de deserción.

- Al igual que en los programas de ingenierías en las licenciaturas los hombres presentan un mayor nivel de deserción.

- No existen diferencias significativas entre los estudiantes provenientes de colegios públicos y privados.

Tecnologías

- El programa de tecnología industrial es el de mayor nivel de supervivencia, mientras que el de tecnología eléctrica es el de peor nivel de supervivencia.

- En los programas de tecnologías al igual que en los de ingenierías y licenciatura las mujeres presentan un mayor nivel de supervivencia respecto de los hombres.

- No existen diferencias significativas entre los estudiantes provenientes de colegios públicos y privados.

Modelo de COX

Ingenierías estratificadas

El modelo seleccionado para los programas de ingenierías es el que contiene las variables edad, sexo y estado civil. Ver siguiente tabla:

VARIABLE	COEFICIENTE	RATIO DE RIESGO
edad	0,08562	1,089
sexo	0,21355	1,238
estado civil	-0,64840	0,523

El riesgo por un año adicional al momento de entrar a la universidad aumenta en un 8,9%, para los hombres el riesgo aumenta en 1,238 veces y para los solteros el riesgo disminuye a un 52,3% del riesgo de los que no son solteros.

Licenciaturas estatificadas

El modelo seleccionado para los programas de licenciaturas es el que contiene las variables edad, sexo y estrato. Ver siguiente tabla:

VARIABLE	COEFICIENTE	RATIO DE RIESGO
edad	0,03115	1,032
sexo	0,17820	1,195
estrato	0,15996	1,173

El riesgo por un año adicional al momento de entrar a la universidad aumenta en un 3,2%, para los hombres el riesgo aumenta en 1,195 veces y para el aumento de estrato se tiene un aumento del riesgo del 17,3%.

Tecnologías estratificadas

El modelo seleccionado para los programas de licenciaturas es el que contiene las variables sexo, departamento y estado civil. Ver siguiente tabla:

VARIABLE	COEFICIENTE	RATIO DE RIESGO
sexo	0,75347	2,124
depto	-0,67597	0,509
estado civil	0,70455	2,023

Para los hombres el riesgo aumenta en 2,124 veces, para los estudiantes que vienen de Risaralda el riesgo disminuye a un 50,9% con respecto de los que vienen de otras partes y para los solteros el riesgo es 2,023 veces el riesgo de los que no son solteros.

De los resultados proporcionados por el modelo de regresión de Cox se puede concluir que el genero, el estado civil, el origen y en menor medida la edad son las variables que mas inciden en el riesgo de desertar. Así se tiene que el estudiante desertor de la universidad (según las variables utilizadas en este modelo) es un estudiante masculino, soltero, viene de otra región y con mayoría de edad.

8.2 RECOMENDACIONES

- Aunque la información utilizada para realizar los cálculos de los diferentes modelos de regresión es necesaria, no suficiente para realizar análisis más detallados de las variables que involucra el fenómeno de la deserción estudiantil. Se recomienda mejorar el sistema de captura de información de los estudiantes de primer semestre de la universidad.

- Es de interés para este tipo de análisis manejar otras variables que puedan incidir en el riesgo de desertar como son: Escolaridad de los padres, el Icfes, ingresos familiares, orientación vocacional, comprensión lectora y matematica, entre otras.

- Los modelos de regresión utilizados en este trabajo corresponden a modelos semiparametricos, se sugiere para trabajos posteriores probar y construir modelos parametricos en los cuales se tenga mayor certeza de la estructura funcional del modelo.

- En este trabajo el evento deserción ocurre cuando el estudiante sale del sistema (universidad) por cualquier motivo diferente a la graduación, es decir, solo se presentan censuras cuando el estudiante se gradúa, esto conlleva a considerar desertor a cualquier estudiante que salga de la universidad sin importar si lo hace de manera voluntaria o no. Se recomienda construir una definición y metodología de cálculo para la deserción más apropiada que permita diferenciar a los estudiantes realmente desertores de los no desertores (bajo promedio, muerte, entre otras).

- Para trabajos posteriores se recomienda manejar covariables dependientes del tiempo que permitan capturar variaciones temporales de algunas covariables.

- Se recomienda para trabajos posteriores realizar análisis mas juiciosos sobre los residuos y la teoría de martingalas involucrada para hacer mejores validaciones de los supuestos y del ajuste del modelo.

- Se recomienda en los programas de maestría en la universidad con énfasis en estadística incluir en los currículos cursos de análisis de supervivencia

- Se recomienda masificar las metodologías de análisis de supervivencia (tablas de vida, Kaplan-Meier, regresión de COX) a las instancias de la universidad a las que le compete el tema de la deserción estudiantil, ya que es una herramienta muy poderosa, fácil de usar y de interpretar para construir políticas antidesercion.

- Para trabajos posteriores se recomienda realizar análisis más desagregados, construir modelos para cada programa académico sería lo ideal.

- Es necesario brindar un tratamiento especial a aquellos estudiantes de otras regiones que presentan mayor riesgo de desertar, así como investigar más a fondo las necesidades, comportamiento y capacidades de los estudiantes de mayoría de edad que permita diseñar políticas antidesercion eficaces y eficientes que beneficien esta población vulnerable.

9. BIBLIOGRAFIA

Bean, J. P (1980). “Student attrition, intensions and confidence”. Research in Higher Education 17. pp291-320.

Borges, R. (2002). “Análisis de supervivencia aplicado a un caso de diálisis renal: diálisis peritoneal en el hospital clínico universitario de Caracas y hemodiálisis en el hospital de clínicas Caracas, 1980-2000”. Tesis de maestría en estadística aplicada de la universidad de los Andes Mérida-Venezuela.

Castaño, E. Gallon, S. Gomez, K. Vasquez, J. (2006). “Análisis de los factores asociados a la deserción y graduación estudiantil universitaria”. Lecturas de Economía. Universidad de Antioquia.

Carvajal, P. Trejos, A. Caro, C. (2004). “Identificar las causas de deserción en la Universidad Tecnológica de Pereira usando la técnica multivariada análisis de correspondencias”. Proyecto de Investigación. Universidad Tecnológica de Pereira.

Cramer, H. (1968). “Métodos Matemáticos de Estadística”. Aguilar.

Demaris, A. (2004). “Regression with Social Data Modeling Continuous and Limited Response Variables”. Wiley.

Hosmer, D.W. y Lemeshow, S. (1999). “Applied Survival Analysis: Regression Modeling of Time to Event Data”. N.Y.: John Wiley & Sons, Inc.

Hosmer, D.W. y Lemeshow, S. (2000). “Applied Logistic Regression”. John Wiley & Sons, Inc.

Giovagnoli, P.I. (2002). “Determinantes de la deserción y graduación universitaria: Una aplicación utilizando modelos de duración”. Tesis de maestría en economía de la UNLP dirigida por el Dr. Alberto Porto.

Landau, S. Everitt, B. (2004). “A handbook of statistical analyses using SPSS”. Chapman and Hall/CRC.

Robinson, R. (1990).”Underestanding the gap between entry and exit: a cohort analysis of african american students persistence”. Journal of Negro Educational. Vol. 59.

Spady, W. (1970). “Dropouts from higher education: an interdisciplinary review and synthesis”. Intechange 1. pp.64-85

Tinto, V (1975). “Dropout from higher education: A theoretical synthesis of recent research.” Review of Educational Research 45. pp. 89-125

ANEXO 1

BASE DE DATOS

El análisis fue realizado sobre una base de datos que contenía la información referente a todos los matriculados por primera vez en los programas de pregrado de la Universidad y pertenecientes a la cohorte del primer semestre del 2003. El periodo de análisis es de 11 semestres (2003-1 hasta 2008_1).

La base de datos que se trabajo está compuesta por 1.264 registros de todos los programas académicos y con variables como se muestra a continuación:

Variables	**Descripción**
Documento	Es el código que identifica al estudiante
Codprg	Corresponde al programa al que el estudiante pertenece
Nombre	Nombre del estudiante
EDAD2	Edad del estudiante
Sexo	Genero del estudiante (1=Hombre; 0=Mujer)
Naturalezacol	Naturaleza jurídica del colegio (1=Privado; 0=Publico)
Estrato	Estrato socioeconómico
Depto	Departamento de origen (1=Risaralda; 0=Resto)
Estcivil	Estado civil del estudiante (1=Soltero; 0=Otro)
FECHAGRADO	Es la fecha en la que obtuvo el titulo
TIEMPO	Tiempo de permanencia en la institución
DESERCION	Se refiere al estado de deserción (1=Deserto; 0=Censura**)**

Esta base de datos fue desagregada y dividida en tres partes, una conteniendo los estudiantes de Ingenierías, otra los estudiantes de Tecnologías y otra los estudiantes de Licenciatura.

ANEXO 2

CODIGO DE COMPUTACION SPSS

```
KM
  TIEMPO  BY codprg  /STATUS=DESERCION(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED
  /SAVE SURVIVAL .

KM
  TIEMPO  BY sexo  /STATUS=DESERCION(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED
  /SAVE SURVIVAL

KM
  TIEMPO  BY naturalezacol  /STATUS=DESERCION(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED
  /SAVE SURVIVAL
```

ANEXO 3

CODIGO DE COMPUTACION SAS

```
1    ods preferences;
2    ods listing;
3    title1;
4    options ps=60;
5          title;
6          footnote;
7    title1 "ANALISIS DE SUPERVIVENCIA";
8    title2 "MODELO DE REGRESION DE COX";
9    proc printto new print=_proj_.PROJ19.RSLT8536.OUTPUT
label="Analysis";
10   run;

11   *** Proportional Hazards Models *** ;
12   options pageno=1;
13   proc phreg data=TMPLIB.datos_tesis OUTEST=_PROJ_.ESTIM25 COVOUT;
14      strata CODPRG;
15      model TIEMPO * DESERCION (0) = EDAD2 SEXO NATURALEZACOL ESTRATO
DEPTO
16      ESTCIVIL / ties=discrete selection=score;
17      baseline out=work._surv survival=_surviv_ upper=_sdfucl_
lower=_sdflcl_;
18   run;

18 !      quit;
19   proc printto; run;

20   options ps=54;
21   goptions reset=all device=WIN;
22   &_gnodisp
23   ** Survival plot **;
24         title;
25         footnote;
26   title1 "ANALISIS DE SUPERVIVENCIA";
27   title2 "MODELO DE REGRESION DE COX";
28   goptions ftext=SWISS ctext=BLACK htext=1 cells;
29   proc gplot data=work._surv gout=WORK._APPG_;
30      label TIEMPO = 'Survival Time';
31      axis2 minor=none major=(number=6)
32         label=(angle=90 'Survival Distribution Function');
33      symbol1 i=stepj l=1 width=1;
34      symbol2 i=stepj l=2 width=1;
35      symbol3 i=stepj l=3 width=1;
36      plot _surviv_ * TIEMPO=codprg /
37           description="SDF of TIEMPO by codprg"
38             frame cframe=CXF7E1C2 caxis=BLACK
39             vaxis=axis2 hminor=0 name='SDF';
40      run;

41   quit;

42   goptions ftext= ctext= htext= reset=symbol;
43   proc delete data=work._surv;run;
```

```
44   &_gdisp
45   goptions reset=all device=WIN;
46   quit;
```