APLICACIÓN DE UN MODELO DE SOBREVIDA PARA EL ANÁLISIS DE LA DESERCIÓN EN LOS PROGRAMAS DE LICENCIATURA EN MATEMÁTICAS Y ESTADÍSTICA Y LICENCIATURA EN TECNOLOGÍA DE LA UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA

JUAN CARLOS AGUILAR CASTRO

UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA

FACULTAD SECCIONAL DUITAMA

LICENCIATURA EN MATEMATICAS Y ESTADISTICA

DUITAMA

2016

APLICACIÓN DE UN MODELO DE SOBREVIDA PARA EL ANÁLISIS DE LA DESERCIÓN EN LOS PROGRAMAS DE LICENCIATURA EN MATEMÁTICAS Y ESTADÍSTICA Y LICENCIATURA EN TECNOLOGÍA DE LA UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA

JUAN CARLOS AGUILAR CASTRO
Código: 200810733

Proyecto de trabajo de grado en modalidad monografía

para obtener el título

de Licenciado en Matemáticas y Estadística

Directora del proyecto

MSc. Carmen Helena Cepeda Araque

UNIVERSIDAD PEDAGOGICA Y TECNOLOGICA DE COLOMBIA

FACULTAD SECCIONAL DUITAMA

LICENCIATURA EN MATEMATICAS Y ESTADISTICA

DUITAMA

2016

U.P.T.C FACULTAD SECCIONAL DUITAMA L.M. E E.E RAE No.

TÍTULO: APLICACIÓN DE UN MODELO DE SOBREVIDA PARA EL ANÁLISIS DE LA DESERCIÓN EN LOS PROGRAMAS DE LICENCIATURA EN MATEMÁTICAS Y LICENCIATURA EN TECNOLOGÍA DE LA UNIVERSIDAD PEDAGÓGICA Y TECNOLÓGICA DE COLOMBIA

AUTOR (A): AGUILAR CASTRO, Juan Carlos.

DIRECTOR (A): CEPEDA ARAQUE, Carmen Helena.

PUBLICACIÓN: Duitama. Universidad Pedagógica y Tecnológica de Colombia, 2016.

UNIDAD PATROCINANTE: Universidad Pedagógica y Tecnológica de Colombia, Facultad Seccional Duitama, Escuela de Matemáticas y Estadística.

PALABRAS CLAVES: Factores contextuales, deserción, estudiante, modelo de sobrevida.

OBJETIVO: Establecer, a través de un modelo de sobrevida, las variables determinantes en la disminución del riesgo de deserción y del aumento en los niveles de graduación de los estudiantes de los programas de Licenciatura en Matemáticas y Estadística y Licenciatura en Tecnología de la Facultad Seccional Duitama.

DESCRIPCIÓN: En el documento se presenta un estudio que tiene como eje principal modelar la deserción y aumento de graduación estudiantil en los programas de Licenciatura en Matemáticas y Estadística y Licenciatura en Tecnología de la UPTC Duitama por medio del modelo de sobrevida, en función de variables contextuales a partir de lo propuesto por el SPADIES, Universidades como; La Nacional, De los Andes, Del Rosario, De Antioquia. La muestra la constituyeron 127 estudiantes del programa de Licenciatura de matemáticas y Estadística y 128 del programa de Licenciatura en Tecnología desde el primer

semestre de 2004 hasta el segundo semestre del 2009 Lo anterior debido a que tal selección en el tiempo permitía la trazabilidad "completa" a 2015 de un estudiante que ingreso a la Universidad en 2009 y desde ahí se escogieron 5 años antes, tal como surgieren los procesos de autoevaluación.

FUENTES*:** Para el desarrollo de este proyecto se consultaron 4 libros de estadística, específicamente de análisis de modelos de sobrevida, modelación y métodos multivariados, además se consultaron 5 artículos 1 trabajo de grado (monografía) y una página web, en los cuales se describían estudios referentes a la deserción estudiantil analizada desde los modelos de sobrevida. Para el caso de la obtención del marco muestral se construyó de acuerdo a la información suministrada por el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior – SPADIES - y la Oficina de Control y Registro Académico – SIRA. Se utilizó para este caso un Muestreo Aleatorio Simple sin Reposición - M.A.S., utilizando el Software estadístico R, con el paquete "sampling Teaching"***, se seleccionó aleatoriamente el listado de los estudiantes que servirían como muestra.

CONTENIDO: El documento presenta cinco partes, empezando por una introducción seguido de una presentación del proyecto, continuando con un marco teórico en el cual se muestra antecedentes teóricos e investigaciones con los principales resultados sobre los factores que influyen en la deserción estudiantil, la conceptualización de cada una de las variables que se utilizaron para la modelación y lo relacionado con el Modelo de Sobrevida. Una cuarta parte dada por el análisis de los datos donde se aplica un análisis univariado y bivariado de las variables que posiblemente estén afectando en la deserción y luego la construcción del modelo. Finalmente se dan unas conclusiones.

METODOLOGÍA: Las fases metodológicas que permitirán la descripción y explicación de la naturaleza longitudinal del proceso de deserción, a través de un modelo de sobrevida, en los programas de Licenciatura en Matemáticas y Estadística y Licenciatura en Tecnología de la Universidad Pedagógica y Tecnológica de Colombia, Duitama, fueron:

1. Conformación del marco teórico el cual partirá de la revisión documental sobre la deserción, antecedentes investigativos y modelos de sobrevida.

2. Definición del diseño metodológico, el cual, y a partir de la disponibilidad de información en bases de datos de la Universidad, implica la selección de los variables estáticas académicas (tales como aumento de los cupos,

política de cancelaciones, desajustes del calendario académico por paros, por ejemplo) y estáticas no académicas (tales como ingreso de la familia, puntaje Icfes, sexo, nivel educativo de la madre, estado de empleo cuando presentó el Icfes, propiedad de la vivienda, edad presentación del Icfes, número de personas en la familia, estrato, nivel SISBEN). También esta fase implica determinar el momento de inicio y la longitud del proceso de monitoreo de los estudiantes.

3. Depuración de la información y ajuste para aplicación de la técnica estadística.

4. Construcción del modelo estadístico.

5. Conclusiones y elaboración del informe final del proyecto de grado.

CONCLUSIONES:

Un estudiante modal del programa de Licenciatura en Matemáticas y Estadística es de género masculino, cuyas madres cuentan en su mayoría con básica primaria, con 3 hermanos en promedio, 22 años promedio de edad, de ingresos familiares (al presentar el Icfes) bajos (0,1 o 2 SMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 67.58 sobre 100 puntos. Se encontró que el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el hecho de o no desertar del Programa. Las variables correlacionas con el tiempo hasta desertar son el número de hermanos y el puntaje en el Icfes, el valor de la correlación indicó que a medida que aumenta el número de hermanos el tiempo hasta desertar disminuye. De igual manera se tiene que a medida que aumenta el puntaje del Icfes también aumenta el tiempo hasta la deserción.

Para el caso del programa de Licenciatura en Matemáticas y Estadística, se encontró que a medida que aumenta el número de semestres la probabilidad de sobrevivir a la deserción disminuye. Es respecto a este aspecto es más rápido el decrecimiento a sobrevivir a la deserción en los primeros cuatro semestres y tiende a estabilizarse a partir del 5 semestre donde está alrededor del 46%. El cuantil 0.5 a partir de la función de sobrevida, determino que hay una probabilidad del 50% de no presentar deserción hasta el tercer semestre y hay una probabilidad del 75% de no presentar deserción hasta el segundo semestre. Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 58%, es decir que hay una probabilidad alta que deserte precozmente del Programa.

Al comparar las curvas de sobrevida para el sexo, nivel educativo de la madre, ingreso familiar y tenencia de vivienda, se encontró que, para las diferentes categorías de las variables, la sobrevivencia a la deserción tiene el mismo comportamiento. La prueba para la hipótesis nula de igualdad de las curvas de sobrevivencia respecto a si trabajaban o no a la hora de presentar el Icfes, determinó que las curvas son distintas para la condición laboral. Se puede inferir que la probabilidad de permanecer en la Licenciatura en Matemáticas y Estadística es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable se deben tomar con precaución ya que no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad.

El modelo óptimo que explica el tiempo hasta que un estudiante deserta de la Licenciatura en Matemáticas y Estadística queda determinado por la edad de ingreso al Programa, el número de hermanos (NH), el puntaje estandarizado que obtuvo en el Icfes y el sexo del estudiante. El número de hermanos incide en el riesgo a desertar, el cual aumenta a medida que el estudiante incrementa su número de hermanos. Si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un hermano más, se multiplica por 1.269 la probabilidad de desertar. Los resultados también indican que el cambio proporcional en la función de riesgo que resulta de un aumento en un punto del puntaje estandarizado del Icfes con el que un estudiante ingresa al Programa, es negativo. Lo anterior significa que el puntaje del Icfes incide en el riesgo a desertar, el cual disminuye a medida que el estudiante incrementa su puntuación. Si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un punto menos en la prueba, se multiplica por 1.0152 la probabilidad de desertar.

Un estudiante modal del programa de la Licenciatura en Tecnología es de género masculino, cuyas madres cuentan en su mayoría con básica primaria, con 3 hermanos en promedio, 20 años promedio de edad, de ingresos familiares (al presentar el Icfes) bajos (0,1 o 2 SMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 59.65 sobre 100 puntos. Se encontró que el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el estado del estudiante

Para el caso de la Licenciatura en Tecnología, se encontró que a medida que aumenta el número de semestres la probabilidad de sobrevivir a la deserción disminuye. Es más rápido el decrecimiento en los primeros seis semestres y tiende a estabilizarse a partir del 7 semestre donde está alrededor del 53%. El cuantil 0.5 a partir de la función de sobrevida, determino que hay una probabilidad del 50% de no presentar deserción hasta el noveno semestre y hay una probabilidad del 75% de no presentar deserción hasta el sexto semestre. Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 40%, es decir que hay una probabilidad alta que deserte precozmente del programa.

Al comparar las curvas de sobrevida para el sexo, nivel educativo de la madre, ingreso familiar y tenencia de vivienda, se encontró que, para las diferentes categorías de las variables, la sobrevivencia a la graduación tiene el mismo comportamiento. La prueba para la hipótesis nula de igualdad de las curvas de sobrevivencia respecto a si trabajaban o no en el momento de presentar de presentar el Icfes, determinó que las curvas son distintas para la condición laboral. Se puede inferir que la probabilidad de graduarse de la Licenciatura en Tecnología es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable se deben tomar con precaución ya que no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad.

El modelo óptimo que explica el tiempo hasta que un estudiante se gradúa de la Licenciatura en Tecnología queda determinado por la edad de ingreso al Programa y la Tenencia de vivienda al momento de presentar el Icfes

Construir a futuro nuevos modelos de sobrevida a partir de la recolección de información a través del instrumento, el cual está disponible en los Anexos C y D, y que tiene como propósito recoger información sobre aspectos importantes que no fueron abordados en este proyecto.

En este estudio se usaron modelos no paramétricos, sería conveniente en los trabajos posteriores asumir una distribución de probabilidad para las variables tiempo hasta la deserción y graduación

CONTENIDO

Pág.

LISTA DE GRÁFICAS

LISTA DE TABLAS

ANEXOS

INTRODUCCIÓN

Actualmente la deserción estudiantil es un problema a nivel nacional e inclusive a nivel de Latinoamérica, debido a que dicho problema genera a los diferentes gobiernos perdidas enormes de recursos, tanto humanos como económicos, por tal motivo en los últimos años diferentes entidades gubernamentales en asocio con la comunidad educativa han estado adelantando investigaciones en todos los aspectos relacionados al problema de deserción estudiantil.

El estudio de este fenómeno es complejo debido a sus múltiples variables que intervienen y la dificultad en si para describir el fenómeno. La deserción refleja varios problemas en varios aspectos, el estudiante a nivel personal se siente frustrado por no culminar sus estudios y por ende no poder mejorar su calidad de vida, a nivel institucional las pérdidas económicas son altas por las inversiones que hace la institución en la capacitación del talento humano necesario para que estos a su vez capaciten a los estudiantes, en lo social los impactos son nefastos; crece el desempleo, crece el ciclo de pobreza, no le aportan al desarrollo del país, no se generan nuevos conocimientos, entre otros.

En general las instituciones de educación superior y el gobierno tienen claro la problemática de la deserción, a pesar de esto no es muy común encontrar trabajos de investigación que den ideas claras sobre el tema en la Universidad Pedagógica y Tecnológica de Colombia, donde se ha abordado el tema descriptivamente.

Los objetivos del presente trabajo fueron los de establecer a través de un modelo de sobrevida las variables determinantes en la disminución del riesgo de deserción y del aumento en los niveles de graduación, establecer el marco conceptual de la deserción estudiantil en la Facultad Seccional Duitama y, dependiendo de la disponibilidad de información, la determinación de las variables académicas y no académicas que se utilizarían para el proyecto. La población analizada fueron los estudiantes de los programas de Licenciatura en Matemáticas y Estadística y la Licenciatura en Tecnología de la Facultad Seccional Duitama. La metodología usada para cumplir a cabalidad con los objetivos trazados consto de cinco etapas, conformación del marco teórico, definición del diseño metodológico, depuración de la información y ajuste para aplicación de la técnica estadística, construcción del modelo estadístico, conclusiones y elaboración del informe final del proyecto de grado.

Se encontró que para el programa de Licenciatura en Matemáticas y Estadística el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el estado del estudiante. La correlación entre las variables número de hermanos y el puntaje del Icfes, indicó que a medida que aumenta el número de hermanos y el tiempo hasta que deserte el estudiante disminuye, es decir, el

estudiante permanecerá por menos tiempo y por ende su probabilidad de desertar aumenta, caso contrario con la variable puntaje del Icfes, indico que a medida que aumenta el puntaje del Icfes también aumenta el tiempo hasta la deserción, es decir, el estudiante permanecerá por más tiempo en la universidad y por ende su probabilidad de desertar disminuye. También se encontró que la probabilidad de desertar precozmente del programa es alta. Se pudo inferir mediante las curvas de sobrevida que la probabilidad de permanecer en el programa es superior en quienes no trabajan que en los que si lo hacen finalmente el modelo óptimo que explica el tiempo hasta que deserte un estudiante queda determinado por la edad de ingreso al Programa, el número de hermanos (NH), el puntaje estandarizado que obtuvo en el Icfes y el sexo del estudiante.

Para la Licenciatura en Tecnología se encontró que las variables género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el estado del estudiante. Con base en la función de riesgo acumulada se encontró que la probabilidad que deserte precozmente del programa es alta y también se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 40%. Finalmente se encontró que el modelo óptimo que explica el tiempo hasta que un estudiante se gradúa de la Licenciatura en Tecnología queda determinado por la edad de ingreso al Programa y la Tenencia de vivienda al momento de presentar el Icfes

Se recomienda construir a futuro nuevos modelos de sobrevida a partir de la recolección de información a través de un instrumento que se derivó del presente proyecto de investigación y que es presentado en los anexos del mismo y, el cual tiene como propósito recoger información sobre aspectos importantes que no fueron abordados en este proyecto.

En este estudio se usaron modelos no paramétricos, sería conveniente en los trabajos posteriores asumir una distribución de probabilidad para las variables tiempo hasta la deserción y graduación.

1 PRESENTACIÓN DEL PROYECTO

El desarrollo de este trabajo monográfico permite, en primer lugar, al estudiante de Licenciatura en Matemáticas y Estadística poner en práctica sus perfiles ocupacional, educador y asesor estadístico en proyectos, ya que consiste en la aplicación de una técnica de modelamiento estadístico para analizar un tema relevante en educación, como lo es la de deserción. La técnica que se utilizará es un modelo de sobrevida, y resulta particularmente novedoso ya que a la fecha la deserción en los programas de la Facultad no ha sido abordada a través de ésta técnica estadística de dependencia.

En segundo lugar, el objeto de estudio que plantea esta monografía es el fenómeno de la deserción de los estudiantes constituido en un problema sobre el cual debe actuar la Universidad, y que está directamente relacionado con la eficiencia de sus recursos y el cumplimiento de sus fines misionales. De hecho, el tema que se abordará resulta de gran interés para la comunidad académica, se proporcionará a las autoridades académicas de la Facultad Seccional Duitama información sustentada y necesaria para formular posibles políticas o reformas educativas que aumenten la permanencia de los estudiantes dentro de la Institución. Es claro que todos los estudiantes que abandonan una institución crean lugares vacantes que pudieron ser ocupados por alumnos que persistieran en los estudios. La pérdida de estudiantes causa problemas financieros a las instituciones al producir inestabilidad en la fuente de sus ingresos, pero también un alto costo social que puede asociarse a la pérdida de productividad laboral derivada de la menor acumulación individual de capital humano.

En tercer lugar, tal como lo plantea el estudio denominado “Deserción estudiantil en la educación superior colombiana”, la deserción es un problema que afecta la relación del Estado con las instituciones de educación superior públicas, en el sentido del incumplimiento de las políticas y las metas sociales establecidas. Aspecto que resalta la relevancia de describir el comportamiento de la deserción en los programas de Licenciatura en Matemáticas y Estadística y la Licenciatura en Tecnología de la Facultad Seccional Duitama. Es evidente que el estudio de fenómenos como la deserción permite aportar elementos sustanciales en los procesos de autoevaluación, autorregulación y mejoramiento continuo de los Programas Académicos que se estudiarán.

La deserción estudiantil hoy en día es un problema de todas las Instituciones de Educación Superior (IES) del país, cifras obtenidas por el Ministerio de Educación Nacional de su programa “Educación de Calidad, el camino para la prosperidad, 2015”, indican que el 45% de los estudiantes que ingresan a la Educación superior, no completan sus estudios hasta graduarse, lo cual demuestra que el problema está latente y es grave, debido a que este conlleva a tres problemas propuestos por las investigaciones realizadas en “*Deserción estudiantil en la*

educación superior Colombiana", por el Ministerio de Educación Nacional, el primero de ellos es el estudiante, ya que por los factores de superación y posicionamiento económico genera frustración, el segundo es la universidad ya que pierde individuos y dinero porque invierte sus recursos en personas que abandonan sus proyectos, y por último la sociedad en donde se genera un incremento del subempleo.

Por lo anterior, el Ministerio de Educación Nacional ha liderado, en conjunto con las instituciones de educación superior, el diseño y la operación de una metodología de seguimiento de la deserción estudiantil en educación superior que se concreta en el Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES). Sin embargo hasta el año 2009, aproximadamente, el análisis de la deserción se había abordado desde un marco estático ignorando la evolución del fenómeno aún, es decir, se describía por qué un estudiante decide abandonar sus estudios pero no se explicaba el proceso de abandono, específicamente, el tiempo que ha durado el estudiante hasta el momento de desertar, por esa razón, en los últimos años el fenómeno de la deserción se ha venido estudiando como un proceso dinámico, a través de modelos como el de sobrevida, ya que permite determinar el riesgo de ocurrencia de un evento, en este caso el evento de desertar, y analizar cuándo es más probable que éste ocurra teniendo en cuenta la influencia de sus principales factores o predictores.

Por supuesto el fenómeno de la deserción no es ajeno a la Universidad Pedagógica y Tecnológica de Colombia Seccional Duitama, de hecho, la Universidad derivado del Plan de Desarrollo Institucional, Lineamiento Formación y Docencia, ha venido ejecutando el programa "Permanencia y Deserción". Programa que busca caracterizar cualitativa y cuantitativamente la deserción para definir las causas de este fenómeno y establecer acciones para minimizarla, tales como el Plan Padrino, el sistema de becas, monitores, y los programas de bienestar, con tales acciones se ha tratado de disminuir la deserción y obedeciendo a los lineamientos del programa "Permanencia y Deserción", los estudios llevados a cabo, particularmente, en la Facultad han privilegiado el diagnóstico, el problema se ha abordado desde un marco estático ignorando la evolución del evento a lo largo del tiempo. Es decir, describen tan sólo cifras del comportamiento de la deserción y en algunos casos el por qué un estudiante decide abandonar sus estudios.

En ese sentido, este proyecto busca obtener una descripción y explicación de la naturaleza longitudinal del proceso de deserción, a través de un modelo de sobrevida, ya que dicha técnica de análisis permite seguir la variable dependiente hasta que ocurra el evento de interés, en nuestro caso la deserción.

El objetivo principal del presente trabajo consiste en establecer a través de un modelo de sobrevida, las variables determinantes en la disminución del riesgo de

deserción y del aumento en los niveles de graduación de los estudiantes de los programas de Licenciatura en Matemáticas y Estadística y la Licenciatura en Tecnología de la Facultad Seccional Duitama, dentro del cual se abordaron objetivos específicos como el de establecer el marco conceptual de la deserción estudiantil en la Facultad Seccional Duitama y, dependiendo de la disponibilidad de información, determinar las variables académicas y no académicas que se utilizarían para el proyecto, identificar el momento de inicio y la longitud del proceso de monitoreo de los estudiantes de los programas académicos.

Del mismo modo, a partir de la información reportada en bases de datos de la Universidad, se evaluaron las variables académicas y no académicas relacionadas con la deserción mediante la construcción del modelo de sobrevida, aspecto que permitió cuantificar la incidencia de cada una de ellas y la determinación de las más influyentes. A partir de la información, se usó la función de riesgo del modelo de sobrevida para algunos de los estudiantes activos en I semestre de 2015. También se diseñó un mecanismo para recoger información de variables no estáticas que permita a largo plazo construir un modelo de sobrevida que las incluya.

2 MARCO REFERENCIAL

Varios estudios en los últimos años se han realizado con el fin de determinar qué factores son determinantes para la deserción estudiantil. Dichos estudios los describiremos a continuación ya que nos sirven como sustento y antecedentes teóricos para el proyecto que se presenta.

2.1 ANTECEDENTES INVESTIGATIVOS

La teoría del suicidio de Durkheim (1897), corrientes como el modelo de integración del estudiante (Tinto, 1975, Spady, 1970), el modelo de desgaste del estudiante (Bean, 1980) y el análisis costo-beneficio desde el punto de vista económico, han servido como bases conceptuales para las primeras investigaciones acerca de la permanencia de los estudiantes en la universidad, dichas teorías hacen referencia a la integración y adaptación del estudiante a la vida universitaria, es decir, entre más se adapte a lo social y académico, menor es la probabilidad de desertar, también hacen referencia a variables externas a la institución y cómo estas influyen en la probabilidad de desertar. Spady en 1970 dice que las universidades tienen sus propios valores y estructura social basándose que la deserción se toma como análoga al suicidio social.

Las anteriores investigaciones fueron realizadas longitudinal y cualitativamente abordando el estudio de forma individual y con algunos factores externos, estos estudios fueron usados para posteriores investigaciones por Adelman (1999), usando datos por cohortes del Centro Nacional de Estadística de los Estados Unidos, con los cuales creo un modelo de probabilidad lineal con el que determina los factores que explican el cambio en la probabilidad de deserción.

El estudio de la deserción tiene varias connotaciones dependiendo del tipo de investigación que se realice, por ejemplo, para Adelman la deserción es causada por los recursos académicos y la asistencia a clases. Por su parte, Robinson (1990) plantea que la deserción depende de la relación del estudiante con los profesores y sus compañeros y afirma también que la probabilidad de desertar en el primero año académico es alta.

Willet y Singer (1991) plantean que los estudiantes con menor rendimiento académico y con padres con bajo nivel educativo y menores ingresos, tienen la mayor probabilidad de desertar. En sus estudios Giovagnoli (2002) concluye que el tipo de colegio, la educación de los padres, el sexo y la situación laboral del estudiante, determinan la probabilidad de que el estudiante deserte o no. Porto y Di Gresia (2004) añaden que estos mismos estudiantes poseen una baja retención y Ainta adiciona que los estudiantes que viven con un solo padre aumentan la probabilidad de desertar. Radcliffe, Huesman y Kellog (2006) plantean que las variables que explican mejor la deserción son el rendimiento académico en el

primer semestre, la preparación académica, la realización de cursos remediales en matemáticas y vivir fuera del campus. O´toole y Wetzel (2008) concluyen en sus estudios que el momento en el que se matriculan los estudiantes, el tipo de ayuda financiera en el primer año y el estado civil, son claves a la hora de que un estudiante deserte a corto o largo plazo.

Ahora bien, los estudios en Colombia sobre las causas de deserción estudiantil eran escasos hasta hace poco tiempo y por ende las instituciones educativas de educación superior y Gobierno Nacional, no tenían herramientas para generar políticas y/o reformas efectivas para afrontar el problema.

Se puede decir que en el año 2003 se presentaron las primeras investigaciones sobre el tema, debido a que años atrás dichos estudios eran más seguimiento que estudios en sí, solo se hacían sobre algunas instituciones y en algunos casos era solo sobre los programas académicos, no estaban sistematizados y el marco conceptual no estaba adecuadamente definido, por ejemplo, no se tenía claro quién era un desertor y las estadísticas eran imprecisas en el tema. En ese año, se encuentran investigaciones que parten de revisiones exhaustivas de la literatura existente y construyendo bases conceptuales sobre la deserción.

Se encuentran investigaciones como la realizada por el Instituto Colombiano para el Fomento de la Educación Superior (Icfes) y la Universidad Nacional de Colombia (finales del 2002 y principios del 2003), donde usando los modelos de sobrevida se determina la probabilidad de que un individuo deserte o no de la educación superior bajo el contexto de ciertas variables como: genero, nivel de escolaridad de los padres, puntaje en el examen de estado, estrato socioeconómico, entre otras. Se encontró que el género, la edad, las condiciones económicas y académicas influyen en la deserción estudiantil.

También se tiene que en dos facultades de la Universidad de Antioquia (2006) se realizaron estudios sobre la deserción y graduación, concluyeron que los aspectos individuales, académicos, socioeconómicos e institucionales, determinan la probabilidad desertar o no.

La Universidad de los Andes, a través del Centro de Estudios sobre el Desarrollo Económico (CEDE), con información de las Instituciones de Educación Superior de Colombia (IES) encontró que, ser hombre, la educación de los padres, la ocupación laboral, el bajo puntaje en el examen de ingreso a la Educación Superior, estudiar en programas como ingeniería, arquitectura y ciencias de la educación, estudiar en una Universidad privada y no contar con apoyo financiero, aumentan el riesgo de desertar.

Siguiendo con los estudios, el Ministerio de Educación Nacional de Colombia (2008) con su programa del Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES), identificaron que la

probabilidad de desertar aumenta sí, es hombre, tiene más de dos hermanos, vivir en departamentos como Antioquia, Guajira, Cesar y Risaralda, menor educación de la madre, haber trabajado cuando presentó el examen de ingreso a la educación superior, menores ingresos del hogar, obtener bajos puntajes en el examen de ingreso a la educación superior, pertenecer a una universidad privada y no contar con apoyo financiero por parte de la universidad.

El programa de Economía de la Universidad del Rosario (2008) concluye en su investigación que la vinculación con el estudiante al mercado laboral y el hecho de que procedan de otras regiones aumentan las probabilidades de desertar en su Institución.

A partir de los resultados de las investigaciones como las anteriores el Gobierno Nacional y las Instituciones de educación superior (IES) proponen programas y/o políticas para disminuir el riesgo de deserción y por ende aumentar la graduación, dichas políticas son entre otras, fomentar programas de apoyo económico, facilitar las trasferencias de estudiantes entre programas, mejorar e incrementar la información que se le entrega a los estudiantes sobre los programas ofrecidos, crear programas de ayudas financieras para los estudiantes que provienen de estratos bajos y para los que provienen de otras ciudades e impulsar la orientación vocacional y profesional previa.

Las investigaciones desarrolladas también permitieron que las instituciones de educación superior cuenten con el SPADIES, el cual permite acceder a información de deserción diferenciando por regiones y departamentos, por sector (oficial y privado), por carácter institucional (universidades, instituciones universitarias, instituciones tecnológicas, instituciones técnicas profesionales), por nivel de formación, (técnico profesional, tecnológico, universitario), por áreas y núcleos del conocimiento, entre programas académico e incluso según su metodología de enseñanza (presencial y a distancia). Es tal vez la herramienta más completa con la que cuentan las IES, para contar con estadísticas claras y eficaces, que posibiliten generar nuevas políticas y contribuyan a la disminución de la deserción estudiantil en la Educación Superior en Colombia.

Pese a todos los esfuerzos realizados por las instituciones de educación superior y el Gobierno Nacional, los índices de deserción en el país siguen siendo altos, las cifras en el 2004, según el Ministerio de Educación Nacional (MEN), señalan que las tasas en Colombia están entre el 45 y 50 por ciento, estas cifras se traducen en que las políticas y/o reformas desarrolladas por las IES y las entidades gubernamentales han sido insuficientes , se pierde el fin social de la educación, particularmente en aspectos de equidad y utilización eficiente de los recursos estatales, institucionales y familiares, como consecuencia también el costo económico de la deserción es altísimo. Según cifras del Instituto Internacional para la Educación Superior en América Latina y el Caribe (IESALC), en el 2005 el costo

de la deserción fue estimado en US$11.1 billones de dólares al año para 15 países de América Latina y el Caribe.

A pesar de las investigaciones acerca del tema aún se sigue discutiendo el concepto de deserción, hay consenso en definirla como un abandono que puede ser causado por factores socioeconómicos, individuales, institucionales y académicos. Concepto que varía desde el punto de vista que se aborde la investigación bien sea individual, estatal, institucional o nacional. Autores como Tinto (1989) afirman que por ser tan complejo el estudio de la deserción, ya que involucra gran variedad de variables, se deja al investigador que defina una aproximación concepto de deserción, según los objetivos y el problema de la investigación. Aspecto que se establecerá a continuación.

2.2 DESERCIÓN ESTUDIANTIL

La deserción estudiantil tiene varios conceptos según desde el punto que se le mire y el contexto en el que se esté trabajando y esto nos lleva a revisar los diferentes conceptos y optar por el más adecuado para nuestra investigación.

De acuerdo con el Ministerio de Educación Nacional de Colombia se entiende por deserción estudiantil "como el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno". (GUZMÁN, Carolina, *et al*. Una introducción a los modelos de duración para el análisis de la deserción estudiantil. En: *Deserción estudiantil en la educación superior colombiana*. 1 ed. Bogotá: Ministerio de educación nacional, 2009. p. 22)

Higuita (2006), en su estudio sobre caracterización de la deserción estudiantil en la Universidad Nacional de Colombia sede Medellín plantea que la deserción "es el abandono de los estudios sin haberlos terminado."

Ruiz Guzmán, Muriel Durán & Gallego Franco del Ministerio de Educación Nacional de Colombia (2009), en su investigación sobre la deserción estudiantil en la educación superior de Colombia, unen las definiciones dadas por Tinto (1982) y Giovagnoli (2002) y definen a la deserción como "una situación a la que se enfrenta un estudiante cuando aspira y no logra cumplir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica".

Teniendo en cuenta lo anterior y de acuerdo al reglamento estudiantil de la Universidad (Acuerdo 130 de 1998) para el presente proyecto se considerará como desertor al estudiante que figure en el Sistema de Registro e Información

Académica SIRA como “retirado definitivamente” o “retiro definitivo”. La categoría de “retirado definitivamente” se refiere a los estudiantes que hayan perdido la calidad de estudiante según lo contemplado en el Artículo 80 del reglamento estudiantil y que se refiere a causas de tipo académico o a los estudiantes que no renovaron matrícula dentro del plazo establecido en el Artículo 42 (tres semestres consecutivos). Se asumirá que la no renovación de matrícula tiene que ver con causas no académicas. La categoría de “retiro definitivo” se refiere al estudiante que le fue aprobada solicitud de retiro definitivo del programa y que, en teoría, se refiere a causas no académicas.

La anterior definición de deserción se enmarca en lo que los autores denominan el abandono de los estudios que puede ser forzoso o voluntario. Forzoso es cuando el sistema retira al estudiante por motivos de causa académica o disciplinaria. Cuando la iniciativa de deserción va por parte del estudiante se considera retiro voluntario.

De otra parte, la deserción también presenta varias clasificaciones, las cuales se presentan a continuación.

Según el Ministerio de Educación Nacional y lo estudios realizados por Higuita (2006) la deserción puede ser con respecto al tiempo y al espacio. Respecto al tiempo puede ser temporal o definitiva. La deserción temporal ocurre cuando el estudiante ha reservado cupo y se espera su reintegro en determinado tiempo o cuando reingresa después de haber sido sancionado académicamente o haberse retirado. Estas categorías se pueden subdividir a la vez en:

Deserción precoz: individuo que habiendo sido admitido por la institución de educación superior no se matricula o se va antes de terminar el primer semestre.

Deserción temprana: individuo que abandona sus estudios entre el segundo y quinto semestre.

Deserción tardía: individuo que abandona sus estudios después del quinto semestre del Programa.

La deserción con respecto al espacio, puede ser:

“Deserción institucional: caso en el cual el estudiante abandona la institución”.

“Deserción interna o del programa académico: se refiere al alumno que decide cambiarse a otro programa que ofrece la misma institución de educación superior”, a este tipo de deserción se le conoce también como trasferencias o movilidad estudiantil.

De acuerdo a los tipos de deserción se establece que para los propósitos de este proyecto sólo se contemplará la deserción con respecto al tiempo definitivo sin

tener en cuenta el espacio ya que se cuenta con sólo la trazabilidad del estudiante en los programas objeto de estudio.

2.3 TEORÍA DE MODELOS DE SOBREVIDA

En los estudios internacionales sobre la deserción estudiantil se han aplicado diferentes metodologías con el objeto de estimar el riesgo de desertar en un punto determinado del tiempo. Se han usado por ejemplo modelos de variables dependientes discretas o cualitativas, en particular, los modelos de regresión logit, probit y el análisis discriminante. Sin embargo, el análisis de sobrevida ha venido tomando gran importancia, para el estudio de la deserción estudiantil. La razón de este crecimiento es el que permite, a diferencia de otros modelos, un análisis dinámico del fenómeno, es decir, permite hacer un seguimiento de los estudiantes desde que inician la carrera hasta que se presente la deserción, y relacionarlo con un conjunto de variables (individuales, socioeconómicas, institucionales y académicas) que puedan influir en dicho tiempo, GUZMÁN, *et al.*

La principal característica de los datos de sobrevida es la presencia de observaciones censuradas, es decir, individuos que no presentan el evento de interés durante el periodo de estudio o recolección de los datos, en nuestro caso los estudiantes que no han desertado se considerarán censurados por derecha.

En general, el análisis de sobrevida nos permite determinar la probabilidad de riesgo de ocurrencia de un evento, en nuestro caso la deserción, teniendo en cuenta una serie de variables. Permite ver la evolución del riesgo de deserción a través del tiempo y responder preguntas como ¿cuándo es más probable que presente deserción y por qué?, ¿Cuáles factores influyen en el tiempo que un estudiante transcurre en la universidad antes de retirarse sin obtener el título?. (GUZMÁN, Carolina, *et al*. Una introducción a los modelos de duración para el análisis de la deserción estudiantil. En: *Deserción estudiantil en la educación superior colombiana*. 1 ed. Bogotá: Ministerio de educación nacional, 2009. p. 42)

2.3.1 Características de los datos de sobrevida

En el análisis de sobrevida el interés se centra en un grupo o varios, con determinadas características, a los cuales, se les observará un evento determinado, llamado falla, el cual ocurre dentro de un tiempo, llamado tiempo de falla, dicha falla solo puede ocurrir máximo una vez en cualquier individuo.

El tiempo de sobrevida es el que ocurre desde el momento que entra al estudio o tiempo inicial hasta el tiempo final o tiempo que trascurre hasta la ocurrencia del evento de interés.

Para determinar el tiempo de falla de forma precisa, se requiere definir la unidad sobre la cual se registra el evento de interés, el tipo de censura y el truncamiento del tiempo, el evento de interés, un tiempo de origen que debe ser definido sin ambigüedad y sobre el cual se modela hasta la ocurrencia del primer evento que sería para el caso univariado o hasta la ocurrencia de varios eventos para el caso multivariado, una escala para medir el paso del tiempo que debe ser acorde a las necesidades del estudio y el significado de falla debe ser totalmente claro.

Tiempo de falla

El tiempo de origen debe ser definido de manera precisa para cada individuo. El tiempo de inicio no necesariamente deber ser el mismo tiempo calendario para cada individuo. En la mayoría de los estudios, se presentan entradas escalonadas; en diferentes fechas, a lo largo de un periodo de tiempo, es por eso que el tiempo de falla se estudia desde la fecha de entrada para cada individuo.

El significado puntual de falla debe ser definido de forma precisa. Por ejemplo, en algún tratamiento médico, el tiempo de falla podría significar la muerte del paciente, por una causa específica como cáncer de huesos, la primera recurrencia de una enfermedad después de un tratamiento, o el aparecimiento de una enfermedad nueva. En algunas aplicaciones hay poca o nada arbitrariedad en el significado de falla, por ejemplos en los procesos industriales, la falla puede significar, el primer momento en el cual el desempeño de un objeto con variable cuantitativa está por debajo de un nivel previamente establecido. Para el caso del proyecto que se presenta, el tiempo de falla corresponde al semestre en el que un estudiante deserta de alguno de los programas objeto de estudio o el tiempo hasta que se gradúa.

Los datos de sobrevida se pueden presentar en formas que dificultan el análisis de los mismos, las características particulares que usualmente se presentan son la censura y el truncamiento. Dada la importancia de identificar estas características a continuación se presenta los tipos de censura y la diferencia con el truncamiento

Censura

Considérese una población homogénea de individuos (para el caso estudiantes), $i = 1, \dots N$, los cuales pueden experimentar el evento de interés (para el caso deserción) sin reemplazo, es decir que una vez ocurra el evento no es posible que vuelva a ocurrir (lo que significa que no se tendrá en cuenta en el estudio los casos de amnistía). Además, supóngase que dicha población es observada por un periodo de tiempo limitado a partir de $t = 0$, en el cual el individuo puede presentar o no el evento de interés, obteniendo así información incompleta en este caso, sobre la ocurrencia del evento, por lo que se considerará al individuo censurado. Este tipo de censura se le llama censura a derecha, la censura a izquierda es

cuando no es posible determinar desde qué momento el individuo puede presentar el evento de interés.

La censura de los datos repercute en la estimación y la inferencia de los parámetros estimados, ya que la información recogida sobre la ocurrencia del evento estará incompleta, alterando la función de verosimilitud empleada en la estimación de los modelos, y las propiedades de los estimadores obtenidos. En los modelos de sobrevida podemos eliminar este tipo de problemas. Para el caso de la deserción estudiantil, los estudiantes aún activos, graduados y los que terminaron sus materias constituyen individuos censurados, ya que una vez finalizado el proceso de observación de la población estos no presentaron el evento de interés, en este caso, deserción. (GUZMÁN, Carolina, *et al*. Una introducción a los modelos de duración para el análisis de la deserción estudiantil. En: *Deserción estudiantil en la educación superior colombiana*. 1 ed. Bogotá: Ministerio de educación nacional, 2009. p. 42)

Censura a la derecha

Un primer tipo de censura es a la derecha y de tipo I, la cual se presenta cuando el evento de interés ocurre antes de un tiempo predeterminado, independientemente del tamaño de muestra. En la censura a derecha es conveniente usar notación específica, como:

$X = Tiempo\ de\ vida\ para\ un\ individuo\ especifico\ bajo\ estudio$

$C_r = Tiempo\ fijo\ de\ censura$

$X´s = Variables\ aleatorias\ independientes\ e\ identicamente\ distribuidas$

$f(x) = Función\ de\ densidad$

Así, el tiempo de vida exacto de un individuo puede ser conocido si solo y si $X \leq C_r$ Si $X > C_r$, el individuo es un sobreviviente y su tiempo de vida es censurado C_r

Las parejas de variables se pueden representar por (T, δ) donde δ es una variable indicadora definida como sigue:

$$\delta = \begin{cases} 1, si\ el\ tiempo\ de\ vida\ de\ X\ corresponde\ a\ un\ evento \\ 0, si\ el\ tiempo\ de\ \ X\ es\ censurado \end{cases}$$

Se tiene que T es igual a X si el tiempo de vida observado es menor o igual a C_r si es censurado, $T = \min(X, C_r)$. Por construcción cada T para cada individuo es una variable aleatoria.

Cuando los individuos tienen diferentes tiempos de censura, fijados previamente, esta forma de cesura se denomina: censura tipo I progresiva, dicha censura se

amplía cuando los individuos entran al estudio de forma escalonada, es decir, en diferentes tiempos, y el punto terminal de estudio predeterminado por el investigador es el mismo para todos. En este caso, el tiempo de censura para cada sujeto es conocido en el momento en que entra al estudio, de manera que cada individuo tiene fijo y especificado su propio tiempo de censura. A este tipo de censura ha sido denominado Censura de Tipo I generalizado. Veamos un ejemplo para este Tipo de censura, Censura Tipo I generalizada para 4 individuos:

$T_1 = X_1, \quad Tiempo\ de\ falla\ para\ el\ primer\ individuo\ (\delta_1 = 1)$

$T_2 = C_{r2}, \quad Tiempo\ censurado\ por\ derecha\ para\ el\ segundo\ individuo\ (\delta_2 = 0)$

$T_3 = X_3, \quad Tiempo\ de\ falla\ para\ el\ tercer\ individuo\ (\delta_3 = 1)$

$T_4 = C_{r4}, \quad Tiempo\ censurado\ por\ derecha\ para\ el\ cuarto\ individuo\ (\delta_4 = 0)$

Un segundo Tipo de censura por derecha es la Censura tipo II, en la cual hay dependencia del tamaño de la muestra (denotado por n) y las fallas que se observen. Aquí todos los individuos son puestos en estudio al mismo tiempo y se da el término de este cuando r es un número entero positivo determinado previamente por el investigador, tal que $r < n$. La notación conveniente para este tipo de censura se presenta como sigue.

Sean $T_1, T_2, \dots, T_n$ los tiempos de falla de los n individuos y sean $T_{(1)}, T_{(2)}, \dots, T_{(n)}$ sus respectivas estadísticas de orden. Entonces el final del estudio queda de forma aleatoria por $T_{(r)}$, la r- ésima estadística de orden. Por tanto $(n - r)$ observaciones serán censuradas al tiempo dado por la r- ésima falla, la cual no se sabe cuando ocurrirá. De modo que esto marca una diferencia importante entre la Censura de Tipo I y la Censura Tipo II,

$$Censura = \begin{cases} Tipo\ I \quad C\ determinístico \\ Tipo\ II \quad C\ aleatorio \end{cases}$$

Otro tipo de Censura es la Tipo III o también llamada Censura aleatoria, la cual surge cuando los sujetos salen del estudio sin presentar la falla por razones no controladas por el investigador. Por ejemplo, en un estudio donde el evento de interés es la muerte por alguna razón ajena a la de interés o si el investigador pierde acceso al sujeto y este sale del estudio. (GODOY, Ángel. *Introducción al análisis de supervivencia con R*. Distrito Federal de México. 2009, p. 10 -16. Trabajo de grado de Actuario. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística). Nótese que para este proyecto solo se abordara la censura a derecha tipo I, debido a que si el estudiante presenta el evento antes del tiempo determinado este será censurado.

Censura por la izquierda

Un tiempo de vida X asociado con un individuo especifico en el estudio, se considera censurado por la izquierda, si este es menor que un tiempo de censura C_1. Esto es, que el evento de interés le ha ocurrido al sujeto en estudio, antes que el sujeto haya sido observado por el tiempo C_1. El dato proveniente de una muestra censurada por la izquierda puede ser representado por las variables (T, ϵ) donde $T = X$ si el tiempo de vida es observado o $T = C_1$ si es censurado y ϵ indica cuando el tiempo de vida exacto es observado $(\epsilon = 1)$ o no $(\epsilon = 0)$.

Algunas veces, si la censura por la izquierda ocurre en el estudio, la censura por la derecha puede ocurrir también y los tiempos de vida son considerados *doblemente censurados*. De nuevo, los datos pueden ser representados por las variables (T, δ) donde $T = \max[\min(X, C_r), C_1]$ es el tiempo de estudio y δ es una variable indicadora definida como sigue: (GODOY, Ángel. *Introducción al análisis de supervivencia con R*. Distrito Federal de México. 2009, p. 10 -16. Trabajo de gado de Actuario. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística).

$$\boldsymbol{\delta} = \begin{cases} 1, & si\ T\ es\ el\ tiempo\ de\ ocurrencia\ del\ evento \\ 0, & si\ T\ es\ el\ tiempo\ censurado\ por\ la\ derecha \\ -1, & si\ T\ es\ el\ tiempo\ censurado\ por\ la\ izquierda \end{cases}$$

Censura por intervalo

Este es un tipo de censura más general que ocurre cuando el tiempo de vida se sabe que ocurre solamente dentro de un intervalo. Este tipo de censura se presenta cuando se tiene un estudio longitudinal donde el seguimiento del estado de los sujetos se realiza periódicamente y por tanto, la falla solo puede conocerse entre dos periodos de revisión, generando un intervalo de la forma (L_i, R_i) para cada sujeto en el estudio.

Truncamiento

Una segunda característica que puede presentarse en algunos estudios de sobrevida, son los datos truncados. El truncamiento es definido como una condición que presentan ciertos sujetos en el estudio y el investigador no puede considerar su existencia. Cuando los datos presentan truncamiento, solamente individuos a los que les ocurre algún evento de interés o la censura, son considerados en el análisis por el investigador.

Truncamiento por la izquierda

Este ocurre cuando los sujetos entran al estudio a un tiempo particular (no necesariamente el origen del evento de interés), y son observados desde este "tiempo retrasado de entrada", hasta que el evento ocurra o hasta que el evento es

censurado. Si Y es el momento de ocurrencia del evento que trunca a los sujetos en estudio, entonces para muestras truncadas por la izquierda, solo los individuos tales que $X \geq Y$ serán considerados. El tipo más común de truncamiento por la izquierda ocurre cuando los sujetos entran al estudio a un tiempo aleatorio y son observados por este "tiempo retrasado de entrada", hasta que el evento ocurre o hasta que el sujeto es censurado por la derecha. En este caso, todos los sujetos que presenten el evento de interés antes del "tiempo retrasado de entrada", no serán considerados por el experimento. Note que esto es opuesto a la censura por la izquierda, donde se tiene información parcial de individuos que presentan el evento de interés antes de su tiempo de entrada al estudio, para truncamiento por la izquierda, estos individuos no serán considerados para ser incluidos en el estudio.

Truncamiento por la derecha

Este ocurre cuando solo individuos que han presentado el evento son incluidos en la muestra y ningún sujeto que haya presentado aún el evento será considerado. Un ejemplo de muestras que presentan truncamiento por la derecha, son los estudios de mortalidad basados en registros de muerte.

2.3.2 Especificación del modelo de sobrevida

En el análisis de sobrevida el interés se centra en un grupo o varios, con determinadas características, a los cuales, se les observará un evento determinado, llamado falla, el cual ocurre dentro de un tiempo, llamado tiempo de falla, dicha falla solo puede ocurrir máximo una vez en cualquier individuo.

El tiempo de sobrevida es el que ocurre desde el momento que entra al estudio o tiempo inicial hasta el tiempo final o tiempo que trascurre hasta la ocurrencia del evento de interés.

Para determinar el tiempo de falla de forma precisa, se requiere definir la unidad sobre la cual se registra el evento de interés, el tipo de censura y el truncamiento del tiempo, un tiempo de origen que debe ser definido sin ambigüedad y sobre el cual se modela hasta la ocurrencia del primer evento que sería para el caso univariado o hasta la ocurrencia de varios eventos para el caso multivariado, una escala para medir el paso del tiempo que debe ser acorde a las necesidades del estudio y el significado de falla debe ser totalmente claro.

Función de sobrevida

La variable aleatoria no negativa T que representa el tiempo de falla y por lo general especificada en el análisis de sobrevida por su función de sobrevida o la función de falla (o riesgo). Estas dos funciones, y funciones relacionadas, que se

utilizan ampliamente en el análisis de datos de sobrevida se presentan a continuación.

Puesto que la población se está monitoreando en el tiempo hasta que se presente o no el evento, la variable de interés en los modelos de duración es el tiempo de duración hasta la ocurrencia de éste, es decir, el tiempo que ha durado el estudiante hasta el momento de desertar. Así, considérese a *T* como una variable aleatoria continua no negativa, la cual denota los tiempos de duración de los individuos de la población. Esta variable con función de densidad de probabilidad $f(t)$ y función de distribución acumulada $F(t) = P(T \leq t)$, también conocida en la literatura de análisis de sobrevida como ***función de falla***, puede caracterizarse de manera única por su función de sobrevida, denotada como *S*(*t*) y dada por: (GUZMÁN, Carolina, *et al*. Una introducción a los modelos de duración para el análisis de la deserción estudiantil. En: *Deserción estudiantil en la educación superior colombiana*. 1 ed. Bogotá: Ministerio de educación nacional, 2009. p. 42).

$$S(t) = 1 - F(t) = 1 - P(T \leq t) = P(T > t) \qquad (1)$$

La cual expresa la probabilidad de que el evento de interés ocurra en un tiempo mayor a t .Función de riesgo. Otra función de interés para el 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+Δt),* dado que no ha ocurrido hasta el tiempo t, que se puede escribir como:

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

Desarrollando la probabilidad condicionada de (2) se obtiene

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t \leq T < t + \Delta t)}{P(T > t)} \qquad (3)$$

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

$$h(t) = \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \qquad (4)$$

De la definición de derivada se tiene que (4) se convierte en:

$$h(t) = \frac{f(t)}{s(t)} \qquad (5)$$

Se tiene entonces una expresión que relaciona la función de sobrevida y la función de riesgo. Si se integra (5), se obtiene entonces la función de riesgo acumulada:

$$H(t) = \int_0^t h(s)ds = \int_0^t \frac{f(s)}{S(s)}ds \qquad (6)$$

Que se puede escribir como $H(t) = \int_0^t \frac{f(s)}{1-F(s)}ds$, haciendo la sustitución $u = 1 - F(s)$, y $du = -f(s)ds$, se tiene que (6) se convierte en: (REBELLON BARRERA, Mauricio. *Análisis de supervivencia aplicado al problema de la deserción estudiantil en la Universidad Tecnológica de Pereira*. Pereira. 2008, p. 3 – 9. Trabajo de grado en, Magister en Investigación de Operativa y Estadística. Universidad Tecnológica de Pereira. Facultad de Ciencias. Departamento de Ingeniera Industrial).

$$H(t) = \int \frac{-du}{u} \qquad (7)$$

De donde resulta que;

$$H(t) = -\ln S(t) \; ; \; S(t) = e^{-H(t)}$$

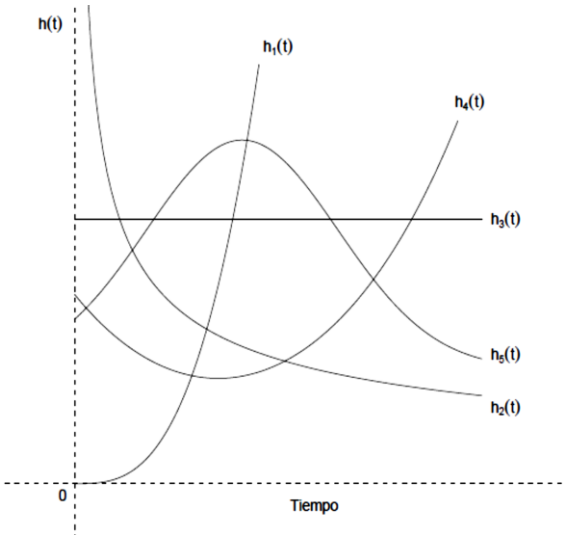
derivando,

$$h(t) = -\frac{d(\ln S(t))}{dt} = -\frac{1}{S(t)}\frac{dS(t)}{dt}$$

usando la ecuación 5

$$f(t) = -dS(t)/dt$$

Para ilustrar las funciones de riesgo se presentan algunos escenarios en la gráfica 1, por ejemplo, pacientes con leucemia que no responden al tratamiento tienen una tasa de riesgo creciente $h_1(t)$. $h_2(t)$, es una función de riesgo decreciente que puede indicar el riesgo de soldados heridos por bala que fueron sometidos a cirugía. El peligro principal es la operación misma y este peligro decrece si la cirugía es exitosa. Una función de riesgo constante como en $h_3(t)$es el riesgo de individuos saludables entre 18 y 40 años de edad cuyos riesgos principales de muerte son los accidentes. La llamada curva de tubo de baño $h_4(t)$describe el proceso de vida humana, durante el periodo inicial el riesgo es alto (alta mortalidad infantil), subsecuentemente el riesgo permanece aproximadamente constante hasta un cierto tiempo, después del cual crece debido a fallas por deterioro. (GODOY, Ángel. *Introducción al análisis de supervivencia con R*. Distrito Federal de México. 2009, p. 25 -26. Trabajo de gado de Actuario. México. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística).



Gráfica 1: Gráficas funciones de riesgo
COLOSIMO, Enricon A. y GIOLO Ruíz, Suely

Finalmente, pacientes con tuberculosis tienen riesgos que se incrementan inicialmente, luego decrecen después de tratamiento. Este incremento y luego decremento se muestra en la función de riesgo $h_5(t)$.

En el caso discreto. Sea T una variable aleatoria discreta que toma valores t_j con $j = 1,2,\ldots$ La función de riesgo se define para los valores t_j y proporciona la probabilidad condicional de falla al tiempo $t = t_j$, dado que el individuo estaba vivo antes de t_j, por lo tanto, se tiene que:

$$h(t_j) = P(T = t_j | \mathrm{T} > t_j)$$

$$= \frac{P(T = t_j)}{P(T \geq t_j)}$$

$$= \frac{f(t_j)}{S(t_j - 1)}$$

Donde $t_j - 1$ corresponde a un instante antes de t_j y, por tanto

$P(t \geq t_j) = 1 - P(T < t_j) = S(t_j-) \neq S(t_j)$, en el caso discreto

Nótese que:

$$f(t_j) = S(t_{j-1}) - S(t_j)$$

Por tanto:

$$h(t_j) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \qquad (8)$$

Despejando a $S(t_j)$ en (8) tenemos:

$$S(t_j) = [1 - h(t_j)]S(t_{j-1}).$$

Donde se puede ver que:

$$S(t_1) = [1 - h_1]S(0) = [1 - h_1]$$

$$S(t_2) = [1 - h(t_2)]S(t_1) = [1 - h(t_2)[1 - h(t_1)]$$

Por tanto, se tiene que:

$$S(t) = \prod_{t_j \le t} \left(1 - h(t_j)\right) \qquad (9)$$

Y consecuentemente de (9) se tiene que:

$$S(t) = \prod_{t_j \le t} \frac{S(t_j)}{S(t_{j-1})}$$

(GODOY, Ángel. *Introducción al análisis de* sobrevida *con R*. Distrito Federal de México. 2009, p. 26 -27. Trabajo de gado de Actuario. México. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística).

Función de riesgo acumulada

La función de riesgo acumulado es denotada por $H(t)$ en el caso continuo corresponde a

$$H(t) = \int_0^t h(u)du$$

y en el caso discreto,

$$H(t) = \sum_{t_j \le t} h(t_j)$$

Pero esta definición tiene un inconveniente con la relación:

$$S(t) = \exp[-H(t)]$$

pues esta definición en el caso discreto no es cierta, aunque es utilizada como una aproximación, sucede que

$$S(t) = \exp\{-H(t)\} = e^{h(t_1)}e^{h(t_2)} \dots e^{h(t_j)}, \quad con\ t_j \leq t,$$

Lo cual no corresponde con la relación entre $S(t)$ y $h(t)$ de la ecuación 9 en el caso discreto. Por este motivo se prefiere definir a la función de riesgo acumulado en el caso discreto como

$$H(t) = -\sum_{t_j \leq t} \ln[1 - h(t_j)],$$

Expresión que está bien definida dado que $0 < h(t_j) < 1$, pues

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

y para los valores t_j donde $S(t_j)$ tiene sentido en el caso discreto, sucede que

$$S(t_j) > S(t_{j+1})$$

De tal modo que

$$S(t) = \exp\{-H(t)\}$$

$$= \exp\left\{\sum_{t_j \leq t} \ln[1 - h(t_j)]\right\} = \prod_{t_j \leq t} \left(1 - h(t_j)\right).$$

Lo cual concuerda con la relación entre $S(t)$ y $h(t)$ de la ecuación 9 en el caso discreto.

En ambos casos, tanto el discreto como el continuo, esta función como su nombre lo indica, acumula el riesgo al paso del tiempo. De tal manera que corresponde a una función no decreciente y de acuerdo a su forma de incrementarse, se podrá tener información del comportamiento del riesgo a lo largo del tiempo, lo cual es una ventaja en el análisis de sobrevida.

Función de vida media residual

La otra función básica en el análisis de sobrevida es la función de *vida media residual* al tiempo $t_0$denotado como $mrl(t_0)$ (por el nombre en inglés *mean residual life*). Para los sujetos de edad t_0, esta función mide la esperanza de tiempo de vida restante, o el tiempo esperado antes de la ocurrencia del evento de interés. Por tanto, queda definida por:

$$mrl(t_0) = E(T - t_0 \,|T > t_0)$$

Para el caso continuo, por definición de esperanza condicional se tiene que

$$E(T - t_0 \,|T > t_0) = \int_0^{\infty} (t - t_0)\, dP(T \leq t | T > t_0)$$

$$\int_{t_0}^{\infty} (t - t_0)\, d\,\frac{S(t_0) - S(t)}{S(t_0)}$$

Por lo cual la función de *Vida media residual* al tiempo t queda definida por

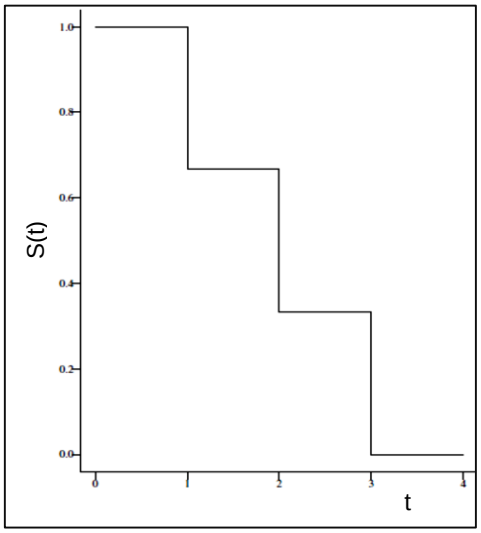
$$mrl(t) = E(T - t_0 | T > t_0) = \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}$$

Se puede apreciar que la vida media residual es el área bajo la curva de sobrevida a la derecha de t dividida entre $S(t)$. De tal modo que la vida media $\mu = E(T) = E(T - 0/T > 0) = mrl(0)$, es el área total de la curva de sobrevida. (Díaz, Guillermo., (agosto 2015). *Análisis estadístico de datos "tiempo para un evento" univariados y multivariados*), III Encuentro Internacional de Matemáticas, Estadística y Educación Matemática, Duitama, Colombia).

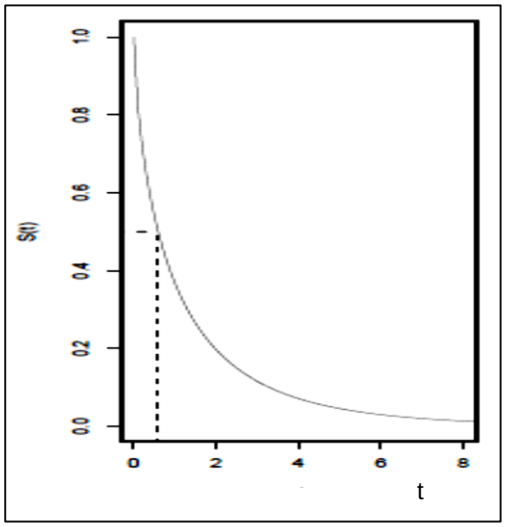
Función de densidad

Como cualquier variable aleatoria, el tiempo de sobrevida T tiene una función de densidad de probabilidad.

Función que evalúa la probabilidad de que en una unidad el evento ocurra después de un cierto tiempo t o equivalentemente, el evento no se ha presentado sobre la unidad antes del tiempo t.



Gráfica 3: Función de sobrevida para para el caso discreto
Díaz, Guillermo



Gráfica 2: Función de sobrevida para el caso continuo
Díaz, Guillermo

Para el caso continuo se tiene que:

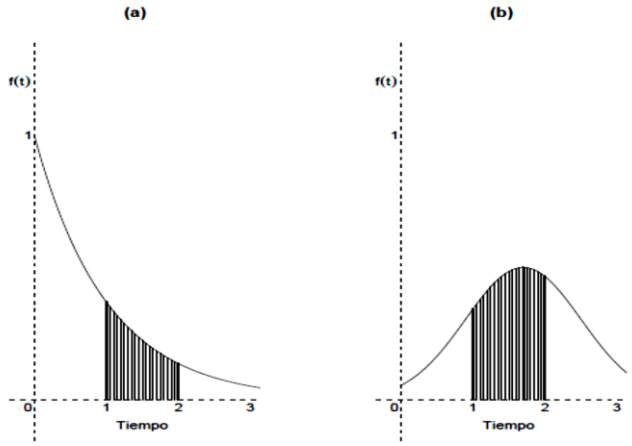
$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)du$$

A partir de la cual se determina por el Teorema fundamental del cálculo que:

$$f(t) = -\frac{dS(t)}{dt}$$

De la anterior relación, $f(t)dt$ puede ser, aunque de manera aproximada, la probabilidad de que un evento pueda ocurrir al tiempo t y $f(t)$ es una función no negativa con área bajo $f(t)$ igual a uno.

De la función de densidad puede ser encontrada la proporción de individuos que cae en cualquier intervalo de tiempo y el pico de frecuencia más alto de fallas. La curva de densidad en la gráfica 4(a) da un patrón de alta tasa de fallas al principio del estudio y una tasa decreciente de fallas cuando se incrementa el tiempo. En la figura gráfica 4(b), el pico de frecuencia alta de fallas ocurre a aproximadamente 1.7 unidades de tiempo. La proporción de individuos que cae entre 1 y 2 unidades de tiempo es igual al área sombreada que aparece en las figuras. (GODOY, Ángel. *Introducción al análisis de supervivencia con R*. Distrito Federal de México. 2009, p. 21 -25. Trabajo de gado de Actuario. México. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística).



Gráfica 4: Curvas de densidad
COLOSIMO, Enricon A. y GIOLO Ruíz, Suely

En el caso discreto, se tiene que:

$$f(t_j) = P(T = t_j)$$

Cuando se trata de describir la variable "tiempo hasta el fallo" las distribuciones, exponencial, weibull, gamma generalizada, log-normal entre otras resultan ser adecuadas para modelarlo.

Es importante mencionar que la elección de la distribución de probabilidad resulta bastante importante ya que como genera estimadores diferentes a las mismas cantidades desconocidas, la utilización de un modelo inadecuado generará errores graves en los estimadores de estas cantidades. La elección del modelo probabilístico adecuado para describir el tiempo de falla, debe, hacerse entonces con mucho cuidado.

2.3.3 Estimación paramétrica del modelo de sobrevida.

Este tipo de estimación requiere suponer un modelo probabilístico para la variable tiempo hasta el evento. Los modelos probabilísticos, por ejemplo, exponencial, weibull, gamma, y log normal, entre otros, son caracterizados por cantidades desconocidas, denominadas parámetros. El modelo gamma generalizado se caracteriza por tres parámetros, los modelos weibull, log-normal y gamma por dos y, exponencial por un parámetro. Estas cantidades conforman de manera general los modelos probabilísticos. Sin embargo, para cada estudio de tiempos de falla, los parámetros deben ser estimados a partir de observaciones de la muestra para que el modelo que se determine logre responder a las preguntas de interés.

Existen algunos métodos para estimar los parámetros de los modelos de sobrevida. Tal vez el más conocido sea el método de mínimos cuadrados. Sin embargo, no es el más adecuado para estimar parámetros censurados. El método

de máxima verosimilitud surge como una opción apropiada para este tipo de datos. Este incorpora las censuras, y es relativamente más simple de entender y tiene propiedades óptimas para muestras grandes.

Suponga que, inicialmente que tenemos una muestra de observaciones $t_1, \ldots\ldots, t_n$ de una cierta población de interés con todos los datos censurados y que la población tiene como función de densidad $f(t)$. Por ejemplo, sea $f(t) = \left(\frac{1}{\alpha}\right) exp\left(\frac{-t}{\alpha}\right)$, significa que las observaciones vienen de una distribución exponencial con parámetro α a ser estimado. La función de verosimilitud para un parámetro genérico θ de una población esta expresada por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta)$$

La dependencia de f en θ, ahora se muestra como L en función de θ.

En esta expresión, θ representa un único parámetro de un conjunto de parámetros. En el modelo Log-normal, por ejemplo, $\theta = (\mu, \sigma)$.

La función de verosimilitud $L(\theta)$ muestra la distribución de cada observación no censurada, en función de su densidad. Estas observaciones no censuradas nos informan que el tiempo de falla es mayor que el tiempo de censura observado y, por tanto, contribuye para que $L(\theta)$ es una función de sobrevida $S(t)$. Las observaciones pueden ser individuos de dos conjuntos, así r serían las primeras observaciones no censuradas $(1, 2, \ldots\ldots, r)$, y $(n - r)$ las segundas observaciones censuradas $(r + 1, r + 2, \ldots\ldots, n)$. La función de verosimilitud asume la siguiente forma:

$$L(\theta) = \prod_{i=1}^{r} f(t_i; \theta) \prod_{i=r+1}^{n} S(t_i; \theta)$$

Lo que es equivalente a:

$$L(\theta) = \prod_{i=1}^{n} [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} [\lambda(t_i; \theta)]^{\delta_i} S(t_i; \theta)$$

Donde δ_i, es la variable indicadora de falla o censura, la expresión obtenida anteriormente es válida para los mecanismos de censura a derecha y a izquierda y también sirve para los mecanismos de censura tipo aleatorio. Siempre es

conveniente, trabajar con el logaritmo de la función de verosimilitud ya que los estimadores de máxima verosimilitud son valores de θ que maximizan a $L(\theta)$ o equivalente a $\log(L(\theta))$. Estos son encontrados resolviendo el sistema de ecuaciones:(COLOSIMO, Enricon A. y GIOLO Ruíz, Suely. Conceptos básicos y ejemplos, técnicas no paremétricas y modelos probabilísticos. En: *Análisis de sobrevida aplicada.* 1 Ed.1997. P. *61).*

$$U(\theta) = \frac{\partial log L(\theta)}{\partial \theta} = 0$$

2.3.4 Estimación no paramétrica del modelo.

Este tipo de estimación no requiere suponer un modelo probabilístico para la variable tiempo hasta el evento. En las observaciones donde hay datos censurados, los métodos convencionales como el análisis descriptivo como, la media, la desviación estándar y técnicas gráficas como, el histograma, box-plot, entre otras, genera problemas que constituyen la división del tiempo en un cierto número de intervalos y luego las de fallas en número de ocurrencias en cada intervalo y por lo tanto cuando existe no es posible construir un histograma, pues no se conoce la frecuencia exacta asociada a cada intervalo.

Existen test no-paramétricos para la comparación de dos o más funciones de sobrevida, tal como el de Kaplan Meier.

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

El estimador de Kaplan Meier utiliza toda la información disponible, casos censurados y no censurados, para realizar la estimación de la función de sobrevida. El estimador en cualquier instante de tiempo es obtenido de la multiplicación de una secuencia de probabilidades condicionales de sobrevida 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 $\frac{(n-\mathrm{d})}{n}$, donde n es el número de casos en riesgo y d es el número de “muertes” observadas.

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

n_i =Número de sujetos en riesgo en el instante t_i

d_i =Número de muertes en el instante t_i

c_i = Número de censuras en el intervalo $[t_1, t_{i+1}]$

Se puede ver:

$$n_1 = n$$

$$t_0 = 0$$

$$d_0 = 0$$

$$c_0 = 0$$

$$t_{k+1} = \infty$$

$$n_{i+1} = n_i - d_i - c_i \text{ para } i = 1,2,3, \dots k-1$$

La expresión general del estimador de Kaplan Meier se puede presentar después de tener estas consideraciones preliminares. Considérese formalmente:

$t_1 < t_2 \dots \dots t_k$, los k tiempos distintos de falla

d_i ó número de faltas en $t_i, i = 1, \dots, k,$ y

n_i ó número de individuos en riesgo en t_i, es decir, los individuos que no han fallado y no han sido censuradó hasta el momento inmediatamente anterior a t_i.

El estimador de Kaplan Meier de la función de sobrevida; está definido por:

$$\widehat{S_{KM}}(t) = \prod_{j=1}^{k} \left(\frac{n_{(j)} - d_j}{n_{(j)}} \right)$$

$$= \prod_{j=1}^{k} \left(1 - \frac{d_j}{n_{(j)}} \right)$$

$$= \prod_{j=1}^{k} \left(1 - q_{(j)} \right)$$

Donde $q_{(j)}$ es la probabilidad que a una unidad le ocurra el evento en el periodo $[t_{(j-1)}; t_{(j)})$ dado que antes del tiempo $t_{(j-1)}$ no le ha ocurrido; es decir,

$$q_{(j)} = P[t_{(j-1)} \in [t_{(j-1)}; t_{(j)}) | T \geq t_{(j-1)}]$$

Si no hay tiempos censurados en el conjunto de datos, $n_{(j)} - d_{(j)} = n_{(j+1)}, j = 1,2, \ldots, k$ y por tanto se tiene que,

$$\widehat{S_{KM}}(t) = \prod_{j=1}^{k} \left(\frac{n_{(j)} - d_j}{n_{(j)}} \right)$$

$$= \prod_{j=1}^{k} \left(\frac{n_{(j+1)}}{n_{(j)}} \right) = \frac{n_{(2)}}{n_{(1)}} X \frac{n_{(3)}}{n_{(2)}} X \ldots \frac{n_{(k+1)}}{n_{(k)}}$$

$$= \frac{n_{(k+1)}}{n_{(1)}}$$

donde $\widehat{S_{KM}}(t) = 1$ para $t < t_1$ y $\widehat{S_{KM}}(t) = 0$ para $t > t_1$.

Ahora $n_{(1)}$ es el número de unidades expuestas al evento justo antes del primer tiempo para el evento, el cual es el número de unidades en la muestra y $n_{(k+1)}$ es el número de unidades tales que para el evento es mayor o igual a $t_{(k+1)}$. Así, en ausencia de censura, $\widehat{S_{KM}}(t)$ es simplemente la función de sobrevida empírica definida anteriormente.

La función de verosimilitud para el estimador de Kaplan Meier se expresa como:

$$L(S(.)) = \prod_{i=1}^{k} \left\{ [S(t_i) - S(t_i + 0)]^{d_i} \prod_{\ell=1}^{m_i} S(t_{i\ell} + 0) \right\}$$

Ahora bien, como uno de los propósitos del proyecto es construir las funciones de sobrevida, respecto a la deserción, para cada uno de los programas de Licenciatura, resulta también útil el poder compararlas. Tal comparación se puede hacer de manera visual a manera exploratoria o utilizando la estadística Log-Rank.

Esta estadística es apropiada cuando la razón de las funciones de riesgo de los grupos a compararse es aproximadamente constante.

La estadística Log-Rank es la diferencia entre el número observado de eventos (fallas) en cada grupo y una cantidad que, para varios propósitos, puede ser considerada como el correspondiente número esperados de eventos (fallas) bajo la hipótesis nula, es decir, que no hay diferencia entre los eventos observados sobre las unidades en los dos grupos.

$H_0: S_1(t) = S_2(t)$; Igualdad de dos funciones de sobrevida.

Suponga que al momento t_j ocurren d_j eventos y n_j unidades expuestas al evento (en riesgo) en un tiempo inmediatamente inferior a t_j de la muestra combinada y, respectivamente, d_{jg} y n_{jg} en la muestra $g = 1,2$ y $j = 1, \dots, k$. En cada tiempo al evento t_j, los datos pueden ser dispuestos en una tabla de contingencia $2X2$. (Díaz, Guillermo., (agosto 2015). *Análisis estadístico de datos "tiempo para un evento" univariados y multivariados*), II Encuentro Internacional de Matemáticas, Estadística y Educación Matemática, Duitama, Colombia).

Tabla 1: Tabla de contingencia 2X2 para la estadística de Log-Rank

Evento	Grupos		Total
	1	2	
Ocurre	d_{j1}	d_{j2}	d_j
No ocurre	$n_{j1} - d_{j1}$	$n_{j2} - d_{j2}$	$n_j - d_j$
Total	n_{j1}	n_{j2}	n_j

Una estadística aproximada para verificar la igualdad de las dos funciones De sobrevida puede hacerse mediante la estadística

$$LR = \frac{\sum_{j=1}^{k}\left[d_{j2} - \mu_{dj2}\right]^2}{\sum_{j=1}^{k}\sigma_{dj2}^2}$$

La cual, para muestras de tamaño grande, tiene distribución Ji-cuadrado con 1 grado de libertad.

También es importante en el proyecto determinar las variables que afectan el tiempo de falla (deserción) en cada uno de los programas de estudio. Razón por la cual a continuación se presenta el modelo de regresión de Cox.

2.4 TEORÍA DEL MODELO DE REGRESIÓN DE COX

Se quiere comparar la sobrevida de dos grupos de estudiantes, el primer grupo hace parte del programa de acompañamiento de tutores y el otro grupo sin el acompañamiento de los tutores, 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 de tutores.

Una manera usual de incluir el efecto de las variables predictoras sobre la función de riesgo es a través de la adopción de especificaciones de la familia de modelos de riesgo proporcional, es decir, adoptar modelos con la propiedad que la razón entre las funciones de riesgo de cualquier par de individuos i y j es constante en el tiempo y son proporcionales una de otra.

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

La presentación inicialmente se realizará sobre el supuesto de una sola covariable y después se generalizará a p variables. Retomando el caso de los dos grupos de estudiantes se tendrá que para el grupo que recibe el tratamiento (acompañamiento de tutores), será el grupo 0 y los que no reciben el tratamiento (el no acompañamiento de tutores), será el grupo 1. Así la Función de falla del primer grupo será representada por $\lambda_0(t)$ y para el segundo grupo $\lambda_1(t)$. Asumiendo proporcionalidad entre estas funciones tenemos que: (REBELLON BARRERA, Mauricio. *Análisis de* sobrevida *aplicado al problema de la deserción estudiantil en la Universidad Tecnológica de Pereira*, Magister en Investigación de Operativa y Estadística. Pereira: Universidad Tecnológica de Pereira. Facultad Ingeniera Industrial. 2008. p. 19).

$$\frac{\lambda_1(t)}{\lambda_0(t)} = K,$$

donde K es la razón de tasas de fallas o riesgo relativo, constante para todo tiempo t de acompañamiento del estudio y no depende del tiempo. X Es una variable indicadora de grupo, en dónde.

$$X = \begin{cases} 0, & grupo\ 0, \\ 1, & grupo\ 1. \end{cases}$$

Si $K = \exp\{\beta x\}$, se tiene que:

$$\lambda_1(t) = \lambda_0(t)\exp(\beta x) \qquad (10)$$

Es decir,

$$\lambda(t) = \begin{cases} \lambda_1(t) = \lambda_0(t)\exp(\beta) & si\ x = 1 \\ \lambda_0(t) & si\ x = 0 \end{cases}$$

De (10) es un modelo de Cox para una única covariable.

De forma general, considere p covariables de modo que X es un vector con componentes $X = (X_1, X_2, \ldots\ldots X_p)$, sea λ_0(t) la función de riesgo para una unidad para la cual los valores de todas las variables explicativas que forman el vector X son cero, la función $\lambda_0(t)$ es llamada función de riesgo inicial o función de riesgo base de tal manera que el modelo lineal múltiple de Cox corresponde a:

$$\lambda_i(t) = \lambda_0(t)g(X^T\beta) \qquad (11)$$

En donde g es una función que debe ser especificada, tal que $g(0) = 1$. Este modelo es compuesto por el producto de dos componentes, uno no-paramétrico y el otro paramétrico. El componente no-paramétrico, $\lambda_0(t)$ se especifica por una función no-negativa del tiempo. Es usualmente llamada función de base, pues $\lambda(t) = \lambda_0(t)$ cuando $X = 0$. El componente paramétrico y que se usa frecuentemente es la siguiente forma multiplicativa.

$$g(X^T\beta) = \exp(X^T\beta) = \exp\left(\beta_1 X_1 + \cdots + \beta_p X_p\right)$$

En donde β es el vector de parámetros asociados a las covariables. Esta forma garantiza que $\lambda_i(t)$ será siempre positiva. Observemos que la constante β_0, presente en los modelos paramétricos, no aparece en el componente mostrado en (11). Esto ocurre debido a la presencia del componente no-paramétrico en el modelo que absorbe este término constante.

Este modelo es también llamado modelo de riesgos proporcionales, pues la razón de las tasas de falla de dos diferentes individuos en el mismo tiempo. La razón de las funciones de tasa de falla para dos individuos diferentes, i e j, será:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t)\exp\{X_i^{\mathrm{T}}\beta\}}{\lambda_0(t)\exp\{X_j^{\mathrm{T}}\beta\}} = \exp\{X_i^T\beta - X_j^T\beta\},$$

La cual no depende del tiempo. La suposición básica para uso de un modelo de regresión de Cox es, por tanto, que las tasas de falla sean proporcionales. (COLOSIMO, Enricon A. y GIOLO Ruíz, Suely. Conceptos básicos y ejemplos, técnicas no paremétricas y modelos probabilísticos. En: *Análisis de sobrevida aplicada.* 1 Ed., 1997. P. *112-113).*

2.4.1 Estimación de parámetros del 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 maximicen 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 parcial en el caso de una única covariable se puede escribir como:(REBELLON BARRERA, Mauricio. *Análisis de supervivencia aplicado al problema de la deserción estudiantil en la Universidad Tecnológica de Pereira*, Magister en Investigación de Operativa y Estadística. Pereira: Universidad Tecnológica de Pereira. Facultad Ingeniera Industrial. 2008. P.21).

$$L_p = \prod_{i=1}^{K} \frac{\exp\{X_i\beta\}}{\sum_{j\epsilon R(t_i)} \exp\{X_j\beta\}},$$

donde el producto se hace sobre K tiempos de sobrevida distintos y X_i corresponde al 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 a todos los individuos con tiempos de sobrevida o censura superiores a t_i.

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

$$\mathrm{Ln}(L_p) = \sum_{i=1}^{k} \left\{ X_i\beta - \ln\left[\sum_{j\epsilon R(t_i)} \exp(X_j\beta) \right] \right\}$$

Derivando con respecto a β se obtiene:

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{k} \left\{ x_{(i)} - \frac{\sum_{j\epsilon R(t_i)} x_j \exp(X_j\beta)}{\sum_{j\epsilon R(t_i)} \exp\{X_j\beta\}} \right\}. \qquad (12)$$

La anterior ecuación se debe igualar a cero y resolverse, pero como se puede notar la ecuación, debido a su complejidad, debe ser resuelta a través de un método numérico como el de Newton-Raphson con algún criterio para finalizar las iteraciones:

$$\sum_{i=1}^{k}\left\{x_{(i)}-\frac{\sum_{j\epsilon R(t_i)}x_j\exp(X_j\beta)}{\sum_{j\epsilon R(t_i)}\exp\{X_j\beta\}}\right\}=0$$

También se tiene que el estimador de la varianza del estimador del parámetro es obtenido de la segunda derivada. Tomando la derivada de (12) se tiene:

$$\frac{\partial^2 LnL(\beta)}{\partial\beta^2}=-\sum_{i=1}^{k}\left\{\frac{\left[\sum_{j\epsilon R(t_i)}\exp(X_j\beta)\right]\left[\sum_{j\epsilon R(t_i)}x_j{}^2\exp(X_j\beta)\right]-\left[\sum_{j\epsilon R(t_i)}x_j\exp(X_j\beta)\right]^2}{\left[\sum_{j\epsilon R(t_i)}\exp(X_j\beta)\right]^2}\right\}$$

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

$$I(\beta)=\frac{\partial^2 L_p(\beta)}{\partial\beta^2}\qquad(13)$$

El estimador de la varianza del coeficiente estimado es el inverso de (13) evaluado para $\hat{\beta}$ y se calcula:

$$\widehat{var}(\hat{\beta})=I^{-1}(\hat{\beta})$$

La desviación estándar $\widehat{SE}(\hat{\beta})$, se calcula como la raíz cuadrada de la varianza dada en la ecuación anterior. Los intervalos de confianza se construyen de la siguiente manera:

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

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 verosimilitudes 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:

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

dónde:

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

Y n_i denota el número de individuos en riesgo en el tiempo de sobrevida 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 Wald se calcula como la razón entre el coeficiente estimado y la desviación estándar de ese coeficiente como se muestra a continuación:

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

Bajo la hipótesis nula el coeficiente es igual a cero, con 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 a continuación.

$$\hat{Z} = \frac{1}{\sqrt{I(\beta)}} \frac{\partial L_p}{\partial \beta}, \qquad \beta = 0$$

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

A continuación, se presenta el caso de varias covariables. Sean p covariables medidas para el individuo i que deban ser incluidas en el modelo de regresión y denotadas por $X_i^T = (X_{i1}, X_{i2}, \dots\dots, X_{ip})$. 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 parámetros, quedando así.

$$h(t, X, \beta) = h_0(t)\exp(X^T\beta)$$

La función de verosimilitud parcial se construye de la misma manera como se construyó el modelo univariado, quedado así.

$$L_p = \prod_{i=1}^{K} \frac{\exp\{X_i^T\beta\}}{\sum_{j\epsilon R(t_i)} \exp\{X_j^T\beta\}}$$

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

$$\frac{\partial L_p(\beta)}{\partial \beta_m} = \sum_{i=1}^{k} \left\{ x_{(im)} - \frac{\sum_{j \epsilon R(t_i)} x_{jm} \exp(X_j^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_j^T \beta\}} \right\} \quad m = 1, \dots p$$

Entonces el estimador de máxima verosimilitud parcial se denota por $\widehat{\beta}' = (\widehat{\beta_1}, \widehat{\beta_1}, \dots, \widehat{\beta_m})$.

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

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

Donde los elementos de la diagonal vienen dados por

$$I(\beta) = \frac{\partial^2 L_p(\beta)}{\partial \beta_m^2} = -\sum_{i=1}^{k} \sum_{j \epsilon R(t_i)} \frac{\exp(X_i^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_j^T \beta\}} \left[x_{jm} - \sum_{j \epsilon R(t_i} \frac{\exp(X_i^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_j^T \beta\}} x_{jm} \right]^2$$

Los elementos fuera de la diagonal vienen dados por:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta_1 \partial \beta_m} = -\sum_{i=1}^{k} \sum_{j \epsilon R(t_i)} \frac{\exp(X_i^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_i^T \beta\}} \left[x_{ji} - \sum_{j \epsilon R(t_i} \frac{\exp(X_i^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_i^T \beta\}} x_{ji} \right] \left[x_{jm} - \sum_{j \epsilon R(t_i} \frac{\exp(X_i^T \beta)}{\sum_{j \epsilon R(t_i)} \exp\{X_i^T \beta\}} x_{jm} \right]$$

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

$$\widehat{var}\left(\hat{\beta}\right) = I^{-1}(\hat{\beta})$$

Las pruebas expuestas 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 el 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:

$$G = 2\left\{L_p\left(\hat{\beta}\right) - L_p(0)\right\}$$

Donde $L_p(0)$ representa la verosimilitud del modelo con cero variables. 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 libertad (un grado por cada variable en el modelo).

Para calcular los estadísticos de Wald y de puntajes implican cálculos matriciales. Si se denota el vector de primeras derivadas parciales de la función parcial de verosimilitud evaluado en 0 como $U(0) = U(\beta)/_{\beta=0}$, 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 cero y matriz de covarianzas dada por la matriz de información evaluada en el vector $\vec{0}, I(\vec{0}) = I(\beta)/_{\beta=0}$.

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

$$U'(\vec{0}) = [I(\vec{0})]^{-1}](0)$$

Bajo la hipótesis nula que todos los coeficientes son iguales a cero la expresión anterior se distribuye chi-cuadrado con p grados de libertad.

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

$$\hat{\beta}' I(\hat{\beta})\hat{\beta}$$

Bajo la hipótesis nula de que todos los coeficientes son iguales a cero la expresión anterior se distribuye chi-cuadrado con p grados de libertad. (REBELLON BARRERA, Mauricio. *Análisis de* sobrevida *aplicado al problema de la deserción estudiantil en la Universidad Tecnológica de Pereira*, Magister en Investigación de Operativa y Estadística. Pereira: Universidad Tecnológica de Pereira. Facultad Ingeniera Industrial. 2008. p. 21 – 25).

2.4.2 Selección del modelo

Cuando se tienen todas las variables a ser incluidas en el modelo, se debe proceder a obtener el modelo más reducido que siga explicando los datos. Para ello se puede recurrir a métodos de selección paso a paso, bien mediante inclusión, hacia adelante, o por eliminación, hacia atrás. Estos métodos consisten en construir sucesivos modelos de manera que uno difiera del precedente en una sola variable e ir comparando los resultados de cada versión con los de la anterior.

A continuación, se describen dos métodos para llegar a determinar el modelo óptimo en términos de parsimonia.(Díaz, Guillermo., (agosto 2015). *Análisis estadístico de datos "tiempo para un evento" univariados y multivariados*, II Encuentro Internacional de Matemáticas, Estadística y Educación Matemática, Tunja, Colombia).

Hacia adelante

1. Inicia con un modelo vacío (sólo la constante).

2. Se ajusta un modelo y se calcula el p-valor del contraste de razón de verosimilitud que resulta de incluir cada variable por separado.

3. Se selecciona el modelo con el p-valor más significativo.

4. Se ajusta de nuevo un modelo con la(s) variable(s) seleccionada(s) y se calcula el p-valor de añadir cada variable no seleccionada anteriormente por separado.

5. Se selecciona el modelo con el más significativo.

6. Se repite 4 –- 5 hasta que no queden variables significativas para incluir.

Hacia atrás

1. Se inicia con un modelo con todas las variables candidatas.

2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.

3. Se selecciona para eliminar la menos significativa.

4. Se repite 2 – 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

Los métodos anteriores pueden utilizar como indicador de comparación AIC - Criterio de Información de Akaike, índice que evalúa el ajuste del modelo a los datos. Cuanto más pequeño es el AIC mejor es el ajuste. El AIC es muy útil para comparar modelos similares con distintos grados de complejidad o modelos iguales (mismas variables) pero con funciones de enlace distintas. El criterio está en función de la verosimilitud de las observaciones.

2.4.3 Evaluación del modelo

El modelo de regresión de Cox es bastante flexible debido a que presenta un componente no-paramétrico. Por esto no se ajusta a cualquier situación como cualquier otro modelo estadístico, requiere el uso de técnicas para evaluar su adecuación. La evaluación consiste en verificar, entre otros aspectos, el supuesto de riesgos proporcionales.

Supuesto de riesgos proporcionales

Para verificar el supuesto de riesgos proporcionales en el modelo de Cox, se puede utilizar un gráfico. El cual consiste en dividir los datos en m estratos, usualmente de acuerdo con alguna covariable. Por ejemplo, dividir los datos en dos estratos de acuerdo a una variable de interés, enseguida se debe estimar la función de tasa de falla de base acumulada para cada estrato usando la expresión:

$$\widehat{\Lambda}_0(t_1) = \sum_{j:t_{j\leq t}} \frac{d_j}{\sum_{i\in Rj} \exp\{X_i^T \hat{\beta})} \qquad (14)$$

Si la suposición es válida, para curvas de logaritmo de $\widehat{\Lambda}_0(t)$ versus t, o $\log(t)$, se deben presentar diferencias aproximadamente constantes de tiempo. Curvas no paralelas significan desvíos de su posición de riesgos proporcionales.

Además de los riesgos proporcionales, hay interés en examinar otros aspectos del modelo de Cox, entre ellos; verificar el ajuste final del modelo, determinar la mejor forma funcional para explicar la influencia de una variable, en presencia de otras covariables, verificar la presencia posible de individuos atípicos (outliers).

Para revisar los casos mencionados, existen técnicas para el modelo de regresión de Cox, que se basan, esencialmente, los mismos tipos de residuos definidos para los modelos paramétricos, tales como los residuos de Cox-Snell, Martingale, residuos de deviance, entre otros, los cuales se resumen a continuación.

Residuos de Cox-Snell

Para el modelo de Cox, los residuos de Cox-Snell se definen por:

$$\widehat{e_i} = \widehat{\Lambda}_0(t_i) exp\left\{\sum_{k=1}^{p} x_{ip}\widehat{\beta_k}\right\}, \qquad i = 1, \dots, n$$

como $\widehat{\Lambda}_0(t_i)$ estimado por (14), si el modelo está bien ajustado, $\widehat{e_i}$ es considerado como una muestra censurada de una distribución exponencial y, el gráfico de, por ejemplo, $\widehat{\Lambda}_0(\widehat{e_i})$ versus $\widehat{e_i}$, debe ser aproximadamente una escalera. Al igual que en los modelos paramétricos, los residuos de Cox-Snell son útiles para aproximar el ajuste global del modelo de Cox. En consecuencia $\Lambda_T(t) = -\log S_T(t)$ tiene una distribución exponencial. (COLOSIMO, Enricon A. y GIOLO Ruíz, Suely. Conceptos básicos y ejemplos, técnicas no paremétricas y modelos probabilísticos. En: *Análisis de sobrevida aplicada.* 1 Ed., 1997. P. *122-123)*

Residuos de martingala

Éstos son una transformación de los denominados residuos de Cox-Snell. Tienen una distribución asimétrica y su esperanza debería ser asintóticamente 0. Son

útiles para indicar si con las covariables del modelo hemos predicho bien los tiempos de sobrevida. Sirven para analizar transformaciones de las covariables que mejoren dichos residuos. Su fórmula es:

$$R_{Mi} = \delta_i - \widehat{\Lambda}_0(t_i) exp\left\{\sum_{k=1}^{p} x_{ip}\widehat{\beta_k}\right\} = \delta_i - \widehat{e_i}, \qquad i = 1, \dots, n$$

Para $i = 1, \dots., n$, siendo $\delta_i = 1$ el indicador de si el sujeto no está censurado y $\delta_i = 0$ si el sujeto está censurado. Dónde $R_i = -log\hat{S}(t_i, X_i)$ son los denominados residuos de CoxSnell extendidos. Los residuos de Cox-Snell se utilizan del siguiente modo: si el modelo de riesgos proporcionales estimado es adecuado el plot de dichos residuos y su correspondiente curva $KM, \hat{S}(R)$, aparecerían como una recta de 45 grados.

Residuos basados en deviances

Son los análogos para datos censurados de los residuos de desviance de un modelo lineal generalizado de la familia exponencial de McCullagh y Nelder. Estos residuos sí se distribuyen de forma simétrica alrededor del cero y sirven para analizar el ajuste del modelo para cada sujeto no censurado y por lo tanto detectar observaciones "atípicas". Su fórmula es: (BOJ, Eva del Val. Evaluación de la hipotesis de riesgos Proporcionales. En: *El modelo de regresion de Cox..* 1 Ed., 2015. P. *35)*

$$R_{Di} = sing(R_{Mi})\sqrt{2[-R_{Mi} - \delta_i log(\delta_i - R_{Mi})]}$$

3 DISEÑO METODOLÓGICO

El diseño metodológico para desarrollar el proyecto, se efectuó en dos fases; en primer lugar, se estableció el marco que sustenta la recolección de la información, luego se describen las variables objeto de estudio.

3.1 DISEÑO MUESTRAL

La recolección de la información se sustenta en:

Población

Constituida por el total de estudiantes que ingresan a la Licenciatura en Matemáticas y Estadística y la Licenciatura en Tecnología desde del primer semestre de 2004 hasta el segundo semestre del 2009. Lo anterior debido a que tal selección en el tiempo permitía la trazabilidad "completa" a 2015 de un estudiante que ingreso a la Universidad en 2009 y desde ahí se escogió 5 años antes, tal como surgieren los procesos de autoevaluación.

De esta población se seleccionó una muestra a través de un diseño aleatorio simple sin reemplazo. Lo anterior ya que es razonable suponer homogeneidad frente al fenómeno deserción que se observará. Al aproximar las características poblacionales mediante estimadores basados en la muestra se comete un error, error que mide la representatividad de dicha muestra (Pérez, 2000), razón por la cual se prefijó un error de muestreo mínimo para determinar el tamaño de muestra.

A continuación, se precisa el tipo de muestreo, marco muestral, unidades de muestreo, error de muestreo y nivel de confianza.

Tipo de Muestreo

Se utilizó para este caso un Muestreo Aleatorio Simple sin Reposición - M.A.S. El M.A.S. consiste en la selección de un subconjunto de elementos de la población en forma aleatoria sin reposición, los cuales tienen la misma probabilidad de ser seleccionados y el orden no interviene. Debido a que el procedimiento de selección es con probabilidades iguales, todas las muestras son equiprobables, es decir, el M.A.S es un método de selección de muestras en las cuales las unidades se eligen individual y directamente por medio de un proceso aleatorio en el que cada unidad no seleccionada tiene la misma oportunidad de ser elegida que todas las otras unidades en cada extracción de la muestra. (Lininger & Warwick, 1978). (ACERON, Deisy. Modelamiento estadístico en el rendimiento académico de los estudiantes los programas de administración de la Universidad Pedagógica y Tecnológica de Colombia – Duitama, segundo semestre de 2012., Licenciada en

Matemáticas y Estadística. Duitama: Universidad Pedagógica y Tecnológica de Colombia. Facultad Seccional Duitama, Departamento de Educación. 2012. P.33).

Unidad de muestreo y marco muestral

Para el caso que compete, la unidad elemental de muestreo corresponde al estudiante que pertenece a un programa de Licenciatura de la UPTC – Duitama, cuyo marco muestral se construyó de acuerdo a la información suministrada por El Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior – SPADIES - y la Oficina de Control y Registro Académico – SIRA -, esta población para el programa de Licenciatura en Matemáticas y Estadística está conformada por 240 estudiantes y para la Licenciatura en Tecnología está conformada por 245 estudiantes. Marco muestral que coincide con las poblaciones de estudio.

Tamaño y selección de la muestra

Los parámetros para calcular el tamaño de la muestra se presentan en la tabla 2:

Tabla 2: Parámetros para determinar tamaño de muestra

N	Total de estudiantes, de un programa de Licenciatura de la UPTC Duitama, desde el primer semestre del 2004 hasta el segundo semestre del 2009 N = 245 y 240
e	Error máximo aceptable, e= 0.06
$1-\alpha$	Nivel de confiabilidad, 0.95
n	Tamaño muestral corregido por el tamaño de la población
Z	Valor correspondiente de la curva normal con una confiabilidad del 95% (1.96)

A partir de la expresión $\mathrm{n} = \frac{\mathrm{NZ^2PQ}}{\mathrm{e^2(N-1)+Z^2PQ}}$ se estima el tamaño de la muestra según el supuesto de la población finita, obteniendo:

Muestra para el programa de Licenciatura en Matemáticas y Estadística

$$\mathrm{n} = \frac{240 * (1.96)^2 * 0.5 * 0.05}{(0.06)^2 * 239 + (1.96)^2 * 0.5 * 0.5} = 126{,}59051 \approx 127$$

Muestra para la Licenciatura en Tecnología

$$\mathrm{n} = \frac{245 * (1.96)^2 * 0.5 * 0.05}{(0.06)^2 * 244 + (1.96)^2 * 0.5 * 0.5} = 127{,}9628 \approx 128$$

Usando el Software estadístico R, con el paquete "***sampling Teaching"***, se seleccionó aleatoriamente el listado de los estudiantes que servirían como muestra para el presente trabajo. En el anexo A se presenta el listado de los seleccionados en la muestra.

3.2 RECOLECCIÓN DE LOS DATOS

Para recoger las mediciones de las variables objeto de estudio en los estudiantes de la muestra se recurrió a los registros que suministra el programa SAPDIES (Sistema para la Prevención y Análisis de la deserción en las Instituciones de Educación Superior), y se completa la medición con la información suministrada por la Oficina de Registro y Control Académico de Tunja, particularmente en su base de datos, Sistema de Información y Registro Académico – SIRA.

El Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior –SPADIES– es una iniciativa del Ministerio de Educación Nacional que incluye una plataforma (base de datos) en la que se consolida y ordena información que permite hacer seguimiento a las condiciones académicas y socioeconómicas de los estudiantes que han ingresado a la educación superior en Colombia.

El SPADIES permite conocer el estado y evolución de la caracterización y del rendimiento académico de los estudiantes, lo cual es útil para: establecer los factores determinantes de la deserción, estimar el riesgo de deserción de cada estudiante, diseñar y mejorar las acciones de apoyo a los estudiantes, orientado a fomentar su permanencia y graduación.

El SPADIES hace parte del Sistema Nacional de Información de la Educación Superior —SNIES— y puede entenderse como un módulo particular de este último aplicado al seguimiento especializado de un fenómeno de especial interés del sector como lo es la deserción estudiantil. Ministerio de Educacion Nacional de Colombia. (30 de Octubre de 2007). *Ministerio de Educacion Nacional, Republica de Colombia.* Obtenido de http://www.mineducacion.gov.co/1621/article-133057.html).

3.3 TÉCNICAS PARA EL ANÁLISIS DE DATOS

Teniendo en cuenta los objetivos del proyecto, se realiza el análisis para cada uno de los programas de estudio, para ello se emplea la técnica estadística de modelamiento y algunas técnicas de estadística descriptiva univariada y bivariada. También se utilizó el programa R con sus diferentes librerías. R es un lenguaje y entorno para computación y gráficos estadísticos. R ofrece una amplia variedad de técnicas gráficas estadísticas (análisis de series de tiempo lineal y no lineal de modelado, pruebas estadísticas clásicas, clasificación, agrupación, etc.) y, es altamente extensible. Uno de los puntos fuertes de R es la calidad de producción de las gráficas y como sus diseños pueden ser producidos, incluyendo símbolos y fórmulas matemáticas para publicaciones posteriores. R está disponible como software libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. Se compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS. R es un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica. Incluye un manejo eficaz de los datos y la instalación de almacenamiento, un conjunto de operadores para los cálculos de matrices particulares, una integrada colección de herramientas intermedias para el análisis de datos, instalaciones gráficas para el análisis y visualización de datos, ya sea en pantalla o en la versión impresa, y un lenguaje de programación bien desarrollado, simple y eficaz que incluye condicionales, bucles, funciones recursivas definidos por el usuario y facilidades de entrada y salida.

R puede ser extendido (fácilmente) a través de paquetes. Hay alrededor de ocho paquetes suministrados con la distribución R y muchos más están disponibles a través de la familia CRAN de sitios de Internet que cubren una amplia gama de estadísticas modernas. (The R Project for Statistical Computing. The R journal (1998). Kurt, Hornik. Obtenido de: https://www.r-project.org/).

Los paquetes usados para el análisis de los datos del proyecto de investigación fueron: sampling Teacheing, R-Commander, Epi, Survival, R-Commander.Survival.

3.4 VARIABLES OBJETO DE ESTUDIO

A continuación, se hará la descripción de las variables que son objeto de medición en este proyecto. Indicando el tipo de variable y su escala de medición.

Tabla 3: Lista de Variables que Influyen en el Rendimiento Académico

Nombre de la variable	Escala de medición
Ingreso familiar al presentar el examen de estado Variable Cualitativa	Ordinal
Género Variable Cualitativa	Nominal
Numero de hermanos Variable Cuantitativa	Discreta
Edad al presentar el examen de estado Variable Cuantitativa	Continua
Vivienda propia Variable Cualitativa	Nominal
Trabajaba al presentar el examen de estado Variable Cualitativa	Nominal
Nivel educativo de la madre Variable Cualitativa	Ordinal
Puntaje en el examen de estado Variable Cuantitativa	Continua
Estado académico del estudiante Variable Cualitativa	Nominal
Tiempo hasta la graduación Variable cuantitativa	Discreta
Tiempo hasta la deserción Variable cuantitativa	Discreta

4 ANÁLISIS LICENCIATURA EN MATEMÁTICAS Y ESTADÍSTICA

A continuación, se realiza un análisis descriptivo de las variables consideradas, serán resumidas en porcentajes y algunos estadísticos, como las medias, la mediana, mínimo y máximo, en el apartado gráfico se presentan algunos diagramas de barras para las variables.

La descripción se realiza a través de R-Comander. Todos los detalles de las sentencias en R figuran en el Anexo B al final del documento.

4.1 DESCRIPCIÓN DE LOS DATOS

La tabla 4 muestra el resumen numérico y descriptivo de las variables objeto de estudio para el programa de Licenciatura en Matemáticas y Estadística.

Tabla 4 Resumen descriptivo de las variables de estudio del programa de LME

Variable	Unidades de medición /codificación	Resumen numérico
Género (Sexo)	F = femenino M = masculino	40 (31.5%) 87 (68.5%)
Nivel educativo de la madre (NEM)	Básica Primaria = BP Básica Secundaria = BS Técnico = TEC Universitario = UN Sin Clasificar = SC	61 (48.03%) 36 (28.35%) 4 (3.15%) 1 (0.79%) 25 (19.69%)
Numero de Hermanos (NH)		Media =2.6 Desviación estándar =1.123208 Coeficiente de variación=0.4309983
Edad ingreso al programa (EdadIngreso)	En Años	Media = 21.98425 Desviación estándar = 4.599835 Coeficiente de Variación = 0.2092332
Ingreso Familiar al Presentar el Examen de Estado (Ingresofami)	Bajo Medio Alto NA Bajo: ingresos de 0,1,2 SMLV; Medio=3,4,5 SMLV, Alto: >5 SMLV	Bajo = 86 (67.72%) Medio = 15 (11.81%) NA = 26 (20.47%)
Trabajaba al presentar el examen de estado (Trabajo)	Si = S No = N Sin clasificar = SC	S = 8 (6.3%) N = 95 (74.8%) SC = 24 (18.9%)
Vivienda Propia (tenenciavivi)	Si = Si No = No Sin Clasificar = SC	Si = 90 (70.31%) No = 28 (21.87%) SC = 10 (7,81%)
Puntaje en el Examen de Estado (Puntaje)	De 1 – 100 puntos	Media = 67.58 Min – Max = 3 - 100 Mediana = 74
Estado del estudiante	Matriculado = MT Desertor = DE Graduado = GR Terminación Académica = TA	17 (13.6%) 77 (60%) 28 (22.4%) 5 (4.0%)
Tiempo Hasta la Deserción (Tdeserción)	Semestres Académicos (16 semanas)	Media = 5.929 Min – Max = 0 - 21 Mediana = 3
En la consideración de estas medidas se debe tener en cuenta que, debido a las censuras en los datos de esta variable, las estadísticas de resumen pueden no tener las propiedades estadísticas deseadas tales como insesgamiento. Por ejemplo, la media muestral ya no es un estimador insesgado de la media poblacional (del tiempo de sobrevida).		
Estado del estudiante Dicotomizado	Desertor = D, No Desertor = ND NA	D = 75 (59%) ND = 50 (39%) NA = 2(2%)
Motivo de deserción Académica	Perdió cupo Articulo 80 Literal B (Quien teniendo un promedio aritmético acumulado inferior a tres cero (3.0), obtenga un promedio aritmético semestral inferior a dos	LB = 7 (5.6%)

	cero (2.0)) = LB Perdió cupo Articulo 80 Literal C (Quien pierda una asignatura que curse en calidad de repitente siendo su promedio acumulado inferior a tres cero (3.0). En el caso en que el promedio aritmético acumulado sea igual o superior a tres cero (3.0), la podrá cursar por tercera y última vez) = LC Perdió cupo Articulo 80 Literal D (Quien pierda en un mismo periodo académico dos asignaturas que se cursen en calidad de repitente) = LD Perdió cupo Articulo 80 Literal E (Quien pierda una asignatura por tercera vez) = LE	LC = 4 (3.2%) LD = 3 (2.4%) LE = 2 (1.6%)
Motivo de deserción no Académica	No matriculado = NM Retiro definitivo = RD Sin clasificar = SC	NM = 1 (0.8%) RD = 58 (46.4%) SC = 2 (1.57%)
Tipo de deserción	Precoz = P Temprana = Te Tardía = Ta SC = Sin clasificar	P = 55 (71.43%) Te = 16 (20.78%) Ta =4 (5.2%) SC =2 (2.6%)

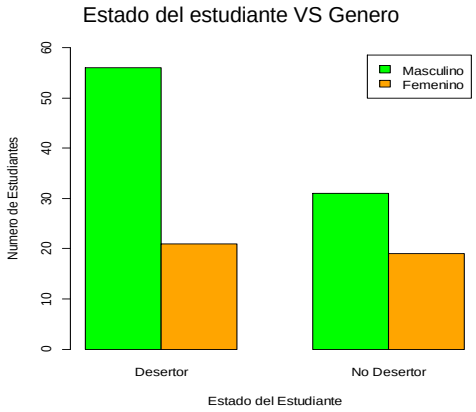
Con los resultados anteriores podemos ver que, de los estudiantes seleccionados en la muestra, la mayoría son hombres (68.5%). La edad promedio al momento de ingresar al Programa es de 22 años. El 70.082% tenían vivienda propia y el 74.8% no trabajaban al momento de presentar el examen de Estado. Se puede ver que el 48.03% de los estudiantes tenían madres con nivel educativo de básica primaria. El 33.07% de los estudiantes tenían dos hermanos al momento de presentar el examen de Estado.

El puntaje medio del examen de estado fue de 67.58, en escala de 0 a 100. El 54.33% tenía como ingreso familiar al momento de presentar el examen de Estado era de 1.2 Salarios Mínimos Mensuales Legales Vigentes (SMMLV). Para la muestra observada, se tiene que la deserción precoz acumulada es la que más se presenta y tiene una proporción de 71.4%, seguida de la temprana acumulada con el 20.78%. De los estudiantes que ingresaron entre el 2004 y 2009 se han graduado el 22.05%. En resumen, se puede decir que un estudiante modal del programa de Licenciatura es de género masculino, cuyas madres cuentan en su mayoría con básica primaria, con 3 hermanos en promedio, 22 años promedio de edad, de ingresos familiares (al presentar el Icfes) bajos (0,1 o 2 SMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 67.58 sobre 100 puntos.

A continuación, se presenta una descripción bivariada de las variables objeto de estudio. En primer lugar, se pretende observar si las variables categóricas

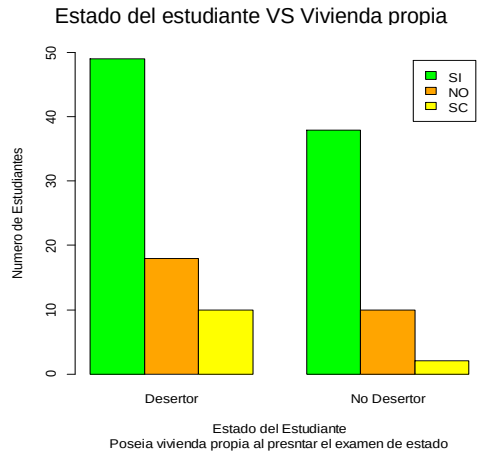
estudiadas se asocian con el estado del estudiante (desertor o no desertor), para lo cual se utilizarán pruebas de independencia y de asociación. Lo anterior se efectúa con el fin de advertir que variables están asociadas al hecho de desertar.

Tabla 5: Resumen bivariado de unas variables de interés del programa de LME

Variables		Estado del estudiante		Estadístico		Gráfica
		N° Desertor	N° de No desertor	Prueba Chi-cuadrado	Valor p	
Genero	M	56	31	1.6168, df = 1	0.2035	
	F	21	19			

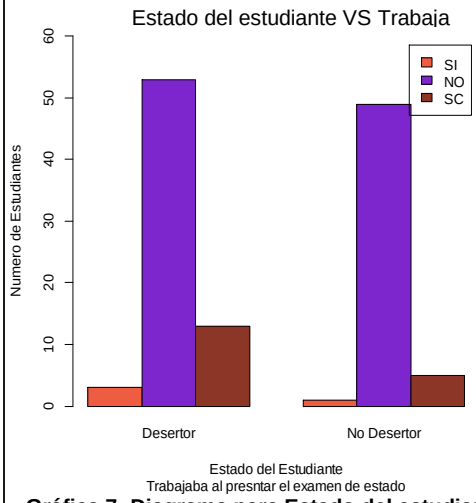
Gráfica 5: Diagrama para Estado del estudiante VS Genero

Se determina que el “Estado del Estudiante” y “Género” son independientes a un nivel del 5%.

Variable		Estado del estudiante		Estadístico		Gráfica
		N° Desertor	N° de No desertor	Prueba Chi-cuadrado	Valor p	
Vivienda Propia	Si	49	38	3.4245, df = 2	0.1805	
	No	18	10			
	SC	10	2			

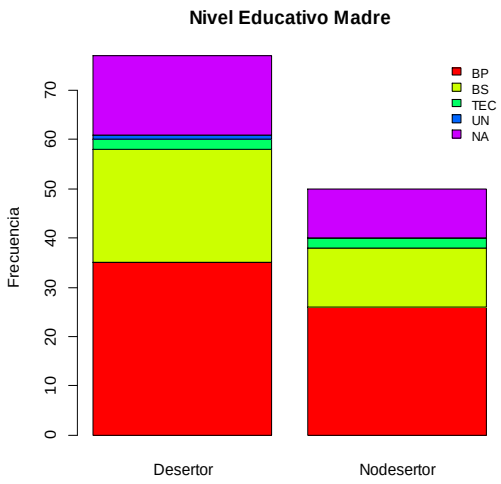
Gráfica 6: Diagrama para Estado del estudiante y Vivienda Propia

Se determina que el “Estado del Estudiante” y “Vivienda propia” son independientes a un nivel del 5%.

Variable		Estado del estudiante		Estadístico		Gráfica
		N° Desertor	N° de No desertor	Prueba Chi-cuadrado	Valor p	
Trabajaba al Presentar el Examen de Estado	Si	8	0	5.853, df = 2	0.05358	
	No	53	41			
	SC	14	9			

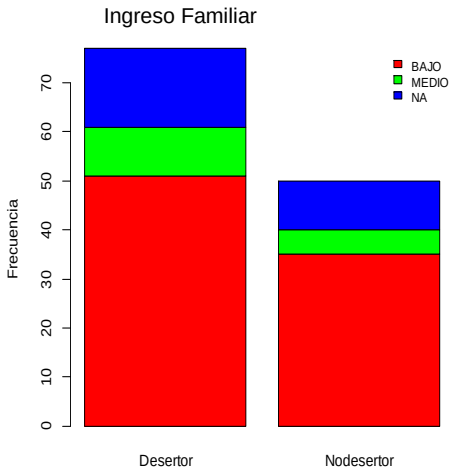
Gráfica 7: Diagrama para Estado del estudiante y Trabajaba

Se determina que el “Estado del Estudiante” y “Trabajaba al presentar el examen de estado” son independientes a un nivel del 5%.

Variable		Estado del estudiante		Estadístico		Gráfica
		N° Desertor	N° de No desertor	Prueba Chi-cuadrado	Valor p	
Nivel educativo madre	BP	35	26	1.4971 df = 4	0.8271	
	BS	23	12			
	TEC	2	2			
	UN	1	0			
	NA	16	10			

Gráfica 8: Diagrama para Estado del estudiante y Nivel educativo de la madre

Se determina que el “Estado del Estudiante” y el “Nivel educativo de la madre” son independientes a un nivel del 5%.

Variable		Estado del estudiante		Estadístico		Gráfica
		Desertor	No desertor	Prueba Chi-cuadrado	Valor p	
Ingreso familiar	BAJO	51	35	0.3015, df = 2,	0.8601	
	MEDIO	10	5			
	NA	16	10			

Gráfica 9: Diagrama para Estado del estudiante e Ingreso familiar

Se determina que el “Estado del Estudiante” y el “ingreso familiar” son independientes a un nivel del 5%.

A partir de la información anterior se puede afirmar que el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el hecho de desertar del Programa.

A continuación, se presentan los resultados del análisis de correlación del tiempo hasta que deserta un estudiante con las variables cuantitativas del estudio.

Tabla 6: Test de Correlación de Pearson para LME

Variable	Tiempo hasta Desertar		
	Coeficiente de Correlación	Estadístico t	p - valor
Edad de Ingreso al programa	-0.0980621	t = -1.0928, grados de libertad = 123	0.2766
Número de hermanos	-0.1908388	t = -1.9048, grados de libertad = 96	0.05979
Puntaje Icfes	0.1924692	t = 2.1215, grados de libertad = 117	0.03599

Se observa que las variables correlacionas con el tiempo hasta desertar son el número de hermanos y el puntaje en el Icfes, el valor de la correlación nos indica que a medida que aumenta el número de hermanos el tiempo hasta desertar

disminuye. De igual manera se tiene que a medida que aumenta el puntaje del Icfes también aumenta el tiempo hasta la deserción.

4.2 MODELO DE SOBREVIDA PARA LA DESERCIÓN

Como el propósito del estudio es identificar los factores que están relacionados al tiempo de la deserción de un estudiante de Licenciatura en Matemáticas y Estadística. En primer lugar, se presentará la aplicación de la metodología de Kaplan Meier para estimar la función de sobrevida, con la cual se puede analizar la evolución de la probabilidad de deserción con su respectivo intervalo de confianza. Posteriormente, se construirá el modelo de regresión de Cox para estimar el efecto de las variables de estudio sobre los tiempos de sobrevida a la deserción. Recordemos las características del estudio:

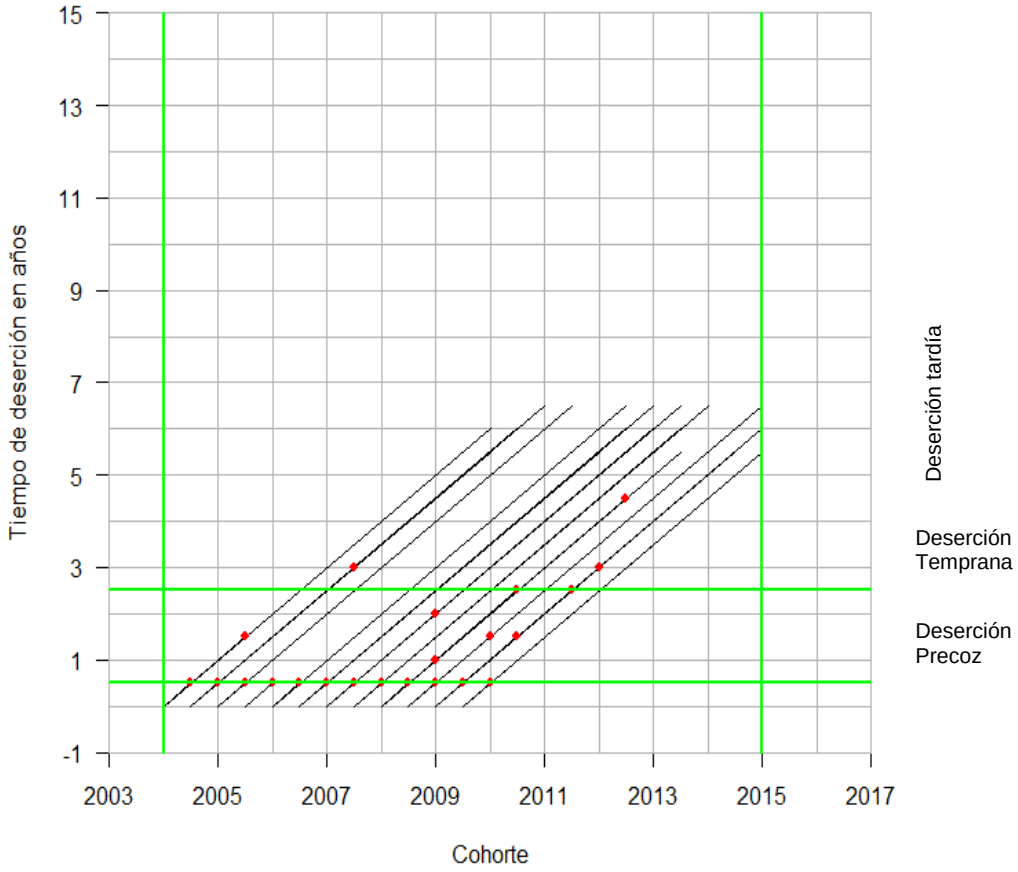
Tabla 7: Estructura del modelo de sobrevida para LME

Objeto de estudio o unidad sobre la cual se registra el evento: estudiante que haya ingresado a la Licenciatura en Matemáticas y Estadística (LME) entre el 2004 y 2009.
Evento de interés o falla*: Abandono del programa por parte del estudiante (deserción).*
***La variable respuesta "Tiempo hasta la deserción"*:** Tiempo hasta que el estudiante se va del programa ya sea porque lo abandona por causas académicas o no académicas. Cuya escala de medición es de razón de tipo discreto, ya que se mide en número de semestres cursados hasta que presenta el fenómeno de estudio.
Tiempo de origen del evento*:* primera matrícula del estudiante en el programa entre el primer semestre del 2004 y del 2009
Tiempo inicial del estudio*:* primer semestre del 2004
Tiempo final del estudio*:* primer semestre del 2015
Tipo de censura*:* Tipo I a derecha. Los individuos entran al estudio en diferentes tiempos, es decir, diferentes cohortes, y el punto final del estudio es el mismo para todos. En este caso, el tiempo de censura para cada estudiante es conocido en el momento que ingreso al estudio, de manera que cada individuo tiene fijo y especificado su tiempo de censura. Se considera como censura al estudiante graduado o que continúa estudiando.
Tiempo de censura*:* tiempo promedio de graduación en la LME, 13 semestres.

A continuación, se presenta la representación de los estudiantes del estudio mediante el diagrama de Lexis, cuya sintaxis R aparece en el Anexo B. Diagrama que refleja el tiempo calendario en el eje horizontal y la longitud del tiempo de

vida, representada por una línea a 45°. El tiempo que un individuo pasa en el estudio es representado por la altura del rayo en el eje vertical.

Gráfica 10. Diagrama de Lexis - Deserción



La gráfica 10 evidencia que los estudiantes bajo estudio no tienen el mismo tiempo de origen, se señala con los puntos rojos aquellos que presentaron el evento y el tiempo en que sucedió (expresado en años). Nótese que la mayoría de estudiantes desertan precozmente, seguido de la deserción temprana, unos pocos se registran con deserción tardía.

A continuación, se presenta el análisis del tiempo hasta la deserción, teniendo en cuenta toda la información disponible, es decir tanto los datos censurados como los no censurados. Las probabilidades de sobrevida en cada intervalo, así como la función de sobrevida se calculan con el estimador de Kaplan Meier, teniendo en cuenta que no se asumirá modelo probabilístico para el tiempo hasta la deserción y que se cuenta con datos censurados a derecha.

4.2.1 Función de sobrevida

El estimador de Kaplan Meier para la función de sobrevida es obtenido en el paquete estadístico R mediante la función survfit, en el anexo B se muestra la sintaxis R. En el gráfico 11 y la tabla 8 se pueden observar las probabilidades de sobrevida en cada intervalo. Time representa el tiempo para el que se presenta la información, n risk indica la cardinalidad del conjunto en riesgo o del número de individuos que continua en el estudio al tiempo correspondiente, n event corresponde al número de fallas que se presentan entre cada tiempo, survival indica el valor que toma la función de sobrevida estimada por el método de Kaplan Meier en el tiempo correspondiente, std err corresponde al error estándar estimado para la función de sobrevida en el tiempo respectivo y lower y upper 95% CI denotan el límite inferior y superior respectivamente del intervalo de confianza al 95% para la función de sobrevida.

Gráfica 11: Función de sobrevida estimada

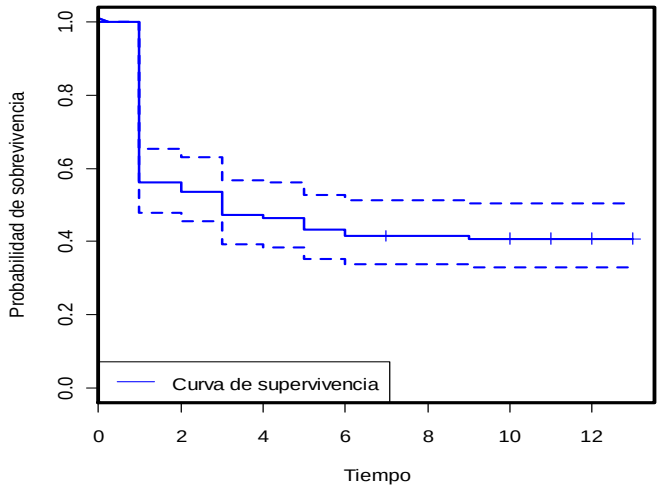


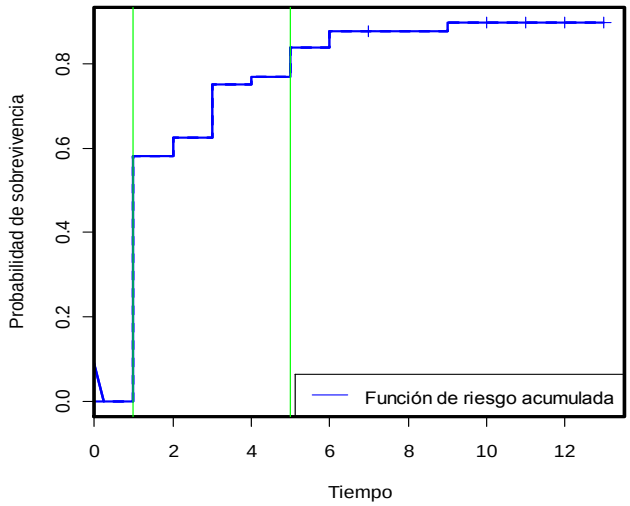
Tabla 8: Estimaciones función de sobrevida para LME

					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	125	55	0.56	0.0444	0.479	0.654
2	70	3	0.536	0.0446	0.455	0.631
3	67	8	0.472	0.0447	0.392	0.568
4	59	1	0.464	0.0446	0.384	0.56
5	58	4	0.432	0.0443	0.353	0.528
6	54	2	0.416	0.0441	0.338	0.512
9	51	1	0.408	0.044	0.33	0.504

Se observa que a medida que aumenta el número de semestres la probabilidad de sobrevida disminuye. Nótese que es más rápido el decrecimiento en los primeros cuatro semestres y tiende a estabilizarse a partir del 5 semestre donde está alrededor del 42%. Se deja de presentar información para los tiempos mayores que 9 ya que es el tiempo mayor de deserción para el estudio, pero con la función de sobrevida estimada se puede calcular la misma para cualquier tiempo mayor que cero. Se calculó el cuantil 0.5 a partir de la función de sobrevida, ver en el anexo B la sintaxis R, encontrándose que hay una probabilidad del 50% de "estar vivo" o "no presentar deserción" hasta el tercer semestre y hay una probabilidad del 75% de "estar vivo" o "no presentar deserción" hasta el segundo semestre.

La función de riesgo, también conocida como la tasa instantánea de mortalidad, describe la forma en que cambia la tasa instantánea de la deserción al paso del tiempo. La función de riesgo acumulado, permite tener información del comportamiento del riesgo a lo largo del tiempo. A continuación, se presenta la correspondiente al programa de Licenciatura, para efectos de interpretación se indica en la gráfica el semestre 1 (deserción precoz) y semestre 5 (deserción temprana).

Gráfica 12: Función de riesgo acumulada - deserción



Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 58%, es decir que hay una probabilidad alta que deserte precozmente del Programa. Vale la pena retomar en este punto el estudio descriptivo para la Licenciatura en Matemáticas y Estadística efectuado por Siauchó y Torres, 2013, quienes determinaron que el fenómeno de la deserción en el Programa se ubicaba principalmente en la categoría de "deserción temprana". De igual manera se observa que la función de riesgo presenta los mayores cambios en la deserción temprana, por ejemplo, se puede apreciar, que se tiene un punto crítico en tercer semestre donde la probabilidad de que, habiendo llegado a segundo deserte en el tercer semestre es del 13%.

Teniendo en cuenta que el fenómeno de la deserción no se está observando de manera aislada, se hace necesario un análisis de la función de sobrevida separado por poblaciones de estudio, es decir para las diferentes categorías que toma la variable cualitativa, lo cual se presenta a continuación. En primer lugar, veremos la función de sobrevida por sexo:

Gráfica 13: Función de sobrevida por sexo

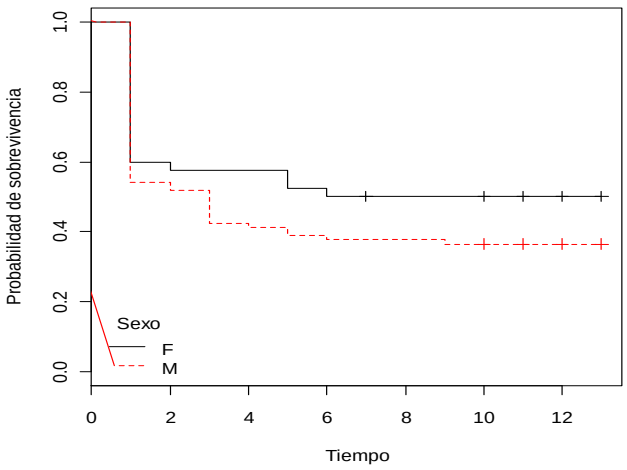


Tabla 9: Análisis de la función de sobrevida por sexo

Sexo=F					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	40	16	0.6	0.0775	0.466	0.773
2	24	1	0.575	0.0782	0.441	0.751
5	23	2	0.525	0.079	0.391	0.705
6	21	1	0.5	0.0791	0.367	0.682
Sexo=M					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	85	39	0.541	0.054	0.445	0.658
2	46	2	0.518	0.0542	0.422	0.636
3	44	8	0.424	0.0536	0.33	0.543
4	36	1	0.412	0.0534	0.319	0.531
5	35	2	0.388	0.0529	0.297	0.507
6	33	1	0.376	0.0526	0.286	0.495
9	32	1	0.365	0.0522	0.275	0.483

Al comparar las curvas de sobrevida de los hombres y las mujeres se obtiene un valor de chi-cuadrado = 1.9 con 1 grado de libertad y un valor p de 0.172, estadística que permite afirmar que no se puede rechazar la hipótesis nula de igualdad de las curvas de sobrevida. Por lo tanto, se puede afirmar que la sobrevida (a la deserción) de hombres y mujeres tiene el mismo comportamiento.

Tal como se presenta en el Anexo B se efectuaron las pruebas de hipótesis para nivel educativo de la madre, ingreso familiar y tenencia de vivienda, encontrándose que para las diferentes categorías de las variables la sobrevida (a la deserción) tiene el mismo comportamiento.

De igual manera se efectuó la prueba para la hipótesis nula de igualdad de las curvas de sobrevida respecto a si trabajaban o no a la hora de presentar el Icfes, encontrándose que, con chi-cuadrado = 9.4 con 2 grados de libertad y un valor p de 0.00928, podemos afirmar que las curvas de sobrevida son distintas para la condición laboral. A continuación, se presenta la información que sustenta la afirmación.

Gráfica 14: Función de sobrevida por ocupación al presentar el Icfes

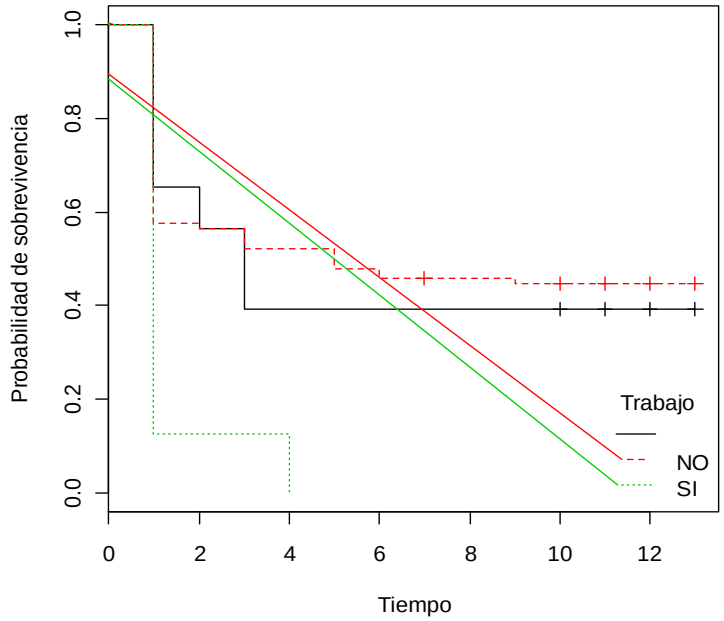


Tabla 10: Análisis de la función de sobrevida para "Trabajo"

Trabajo					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	23	8	0.652	0.0993	0.484	0.879
2	15	2	0.565	0.1034	0.395	0.809
3	13	4	0.391	0.1018	0.235	0.651
Trabajo=NO					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	94	40	0.574	0.051	0.483	0.684
2	54	1	0.564	0.0511	0.472	0.674
3	53	4	0.521	0.0515	0.429	0.633
5	49	4	0.479	0.0515	0.388	0.591
6	45	2	0.457	0.0514	0.367	0.57
9	42	1	0.447	0.0513	0.357	0.559
Trabajo=SI					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	8	7	0.125	0.117	0.02	0.782
4	1	1	0	NaN	NA	NA

Se puede ver en una prueba de hipótesis que la probabilidad de permanecer en el Programa es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable y su consideración en el modelo de Cox (se presenta más adelante) se deben tomar con precaución ya que, en primer lugar, no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad. En segundo lugar, de los 127 estudiantes del estudio tan sólo 8 presentaron la condición de estar trabajando a la hora de presentar el Icfes.

4.2.2 Factores relacionados con el riesgo de deserción

Con el objetivo de medir los efectos de las variables consideradas en el estudio como explicativas del fenómeno de la deserción (edad de ingreso al Programa, nivel educativo de la madre, número de hermanos, tenencia o no de vivienda, ocupación a la hora de presentar el Icfes, ingreso familiar, puntaje Icfes estandarizado), a continuación, se presentan las estimaciones del modelo de

riesgo proporcional utilizando el modelo semiparamétrico de Cox. Los resultados del modelo inicial se presentan en la tabla 11

Tabla 11: Análisis del modelo semiparamétrico de Cox para LME

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
EdadIngreso	5.22E-02	1.05E+00	5.12E-02	1.019	0.30831	
Ingresofami[T.Medio]	3.55E-01	1.43E+00	4.59E-01	0.772	0.43983	
NEM[T.BP]	1.39E+01	1.08E+06	1.32E+03	0.011	0.99161	
NEM[T.BS]	1.43E+01	1.54E+06	1.32E+03	0.011	0.99139	
NEM[T.TEC]	1.37E+01	8.48E+05	1.32E+03	0.01	0.99175	
NH	2.09E-01	1.23E+00	1.85E-01	1.131	0.25815	
Puntaje	-1.55E-02	9.85E-01	5.82E-03	-2.656	0.00791	**
Sexo[T.M]	4.12E-01	1.51E+00	3.27E-01	1.261	0.20741	
tenenciavivi[T.NO]	3.95E-01	1.49E+00	9.04E-01	0.437	0.66208	
tenenciavivi[T.SI]	7.65E-01	2.15E+00	8.87E-01	0.862	0.38849	
Trabajo[T.SI]	7.92E-01	2.21E+00	6.60E-01	1.2	0.2303	

Seguidamente se procede a seleccionar el modelo más parsimonioso, utilizando el método hacia adelante y usando como criterio el AIC (Criterio de Información de Akaike), encontrándose que el modelo óptimo queda determinado por la edad de ingreso al Programa, el número de hermanos (NH), el puntaje estandarizado que obtuvo en el Icfes y el sexo del estudiante. La tabla 12 resume el modelo óptimo:

Tabla 12: Resumen del modelo óptimo para LME

	Coef	exp(coef)	se(coef)	Z	Pr(>|z|)	
EdadIngreso	7.04E-02	1.07E+00	4.53E-02	1.554	0.12022	
NH	2.38E-01	1.27E+00	1.21E-01	1.97	0.04883	*
Puntaje	-1.51E-02	9.85E-01	5.45E-03	-2.773	0.00556	**
Sexo[T.M]	4.76E-01	1.61E+00	3.07E-01	1.548	0.1216	
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'		

	exp(coef)	exp(-coef)	Límite inferior 0.95	Límite superior 0.95
EdadIngreso	1.073	0.932	0.9818	1.1726
NH	1.269	0.7881	1.0012	1.6081
Puntaje	0.985	1.0152	0.9745	0.9956
Sexo[T.M]	1.609	0.6216	0.8812	2.9374

En cuanto a las características observadas en los estudiantes se tiene que tan sólo el número de hermanos y el puntaje Icfes son significativas a la hora de explicar el riesgo de deserción. Los resultados indican que el cambio proporcional

en la función de riesgo que resulta de un aumento en un hermano del individuo que ingresa al Programa es positivo. Lo anterior significa que el número de hermanos incide en el riesgo a desertar, el cual aumenta a medida que el estudiante incrementa su número de hermanos. La estimación del riesgo relativo para el número de hermanos es de 1.269 con un intervalo de confianza al 95% de (1.0012, 1.6081) que indica que, si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un hermano más, se multiplica por 1.269 la probabilidad de desertar.

Los resultados también indican que el cambio proporcional en la función de riesgo que resulta de un aumento en un punto del puntaje estandarizado del Icfes con el que un estudiante ingresa al Programa, es negativo. Lo anterior significa que el puntaje del Icfes incide en el riesgo a desertar, el cual disminuye a medida que el estudiante incrementa su puntuación, en otras palabras, se tiene que ante mayor sea el puntaje del Icfes con el que ingresan al Programa menor será el riesgo de desertar. La estimación del riesgo relativo para el número de hermanos es de 0.985 con un intervalo de confianza al 95% de (0.9745, 0.9956), que indica que, si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un punto menos en la prueba, se multiplica por 1.0152 la probabilidad de desertar. Este es un resultado consistente en el entendido que la prueba es una buena medida de las habilidades y conocimientos, los cuales son fundamentales en los primeros semestres del Programa ya que éstos influyen el en resultado académico y consecuentemente en su riesgo de desertar.

Se observa que el modelo es aceptable para cualquiera de los tres criterios y que las variables seleccionadas para el estudio tan sólo explican el 14.7% de la variabilidad en el tiempo de deserción.

Tabla 13: Estadísticos del modelo

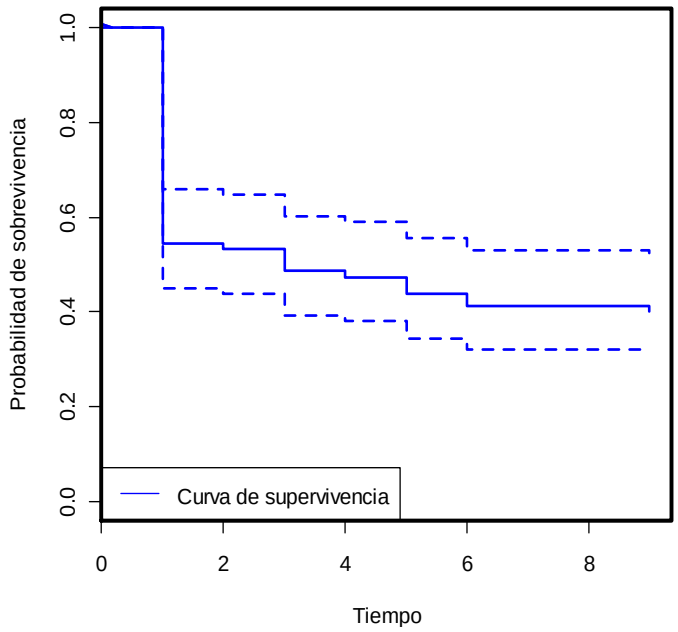
R-cuadrado= 0.147 (Máximo posible= 0.994)			
Prueba de verosimilitud	15.21	con 4 grados de libertad	p=0.004288
Prueba de Wald	15.97	con 4 grados de libertad	p=0.003065
Prueba de puntaje (Log Rank)	16.48	con 4 grados de libertad	p=0.002434

Una vez estimado el modelo, se desea saber cuándo es más probable que ocurrirá la deserción. A continuación, se presenta la estimación “no paramétrica” de la función de sobrevida y la función de riesgo.

Tabla 14: Estimación no paramétrica de la función de sobrevida y de riesgo

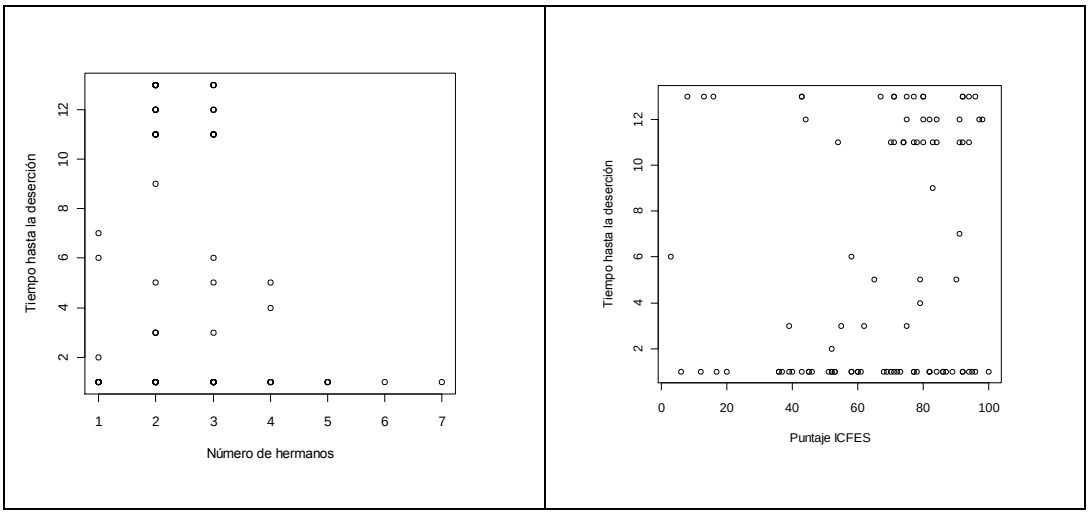
					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	96	46	0.544	0.0531	0.449	0.659
2	50	1	0.532	0.0533	0.438	0.648
3	49	4	0.486	0.0535	0.391	0.603
4	45	1	0.474	0.0535	0.38	0.591
5	44	3	0.437	0.0533	0.344	0.555
6	41	2	0.413	0.0529	0.321	0.531
9	38	1	0.4	0.0527	0.309	0.518

Gráfica 15: Función de sobrevida del modelo óptimo para LME



Se observa que la curva de sobrevida que incluye las variables consideradas en el modelo no presenta mayores variaciones con la estimada de Kaplan Meier presentada en la gráfica 11. Esto se debe a que el número de hermanos y el puntaje Icfes de ingreso (las variables que resultaron significativas) influyen de manera muy particular en el tiempo hasta la deserción. Como se puede ver en la gráfica 15 los estudiantes desertan en primer semestre independiente del número de hermanos y los estudiantes con puntaje Icfes inferior a 20 puntos pueden durar mucho tiempo en el Programa.

Gráfica 16: "dfbetas" para número de hermanos y Puntaje Icfes para LME



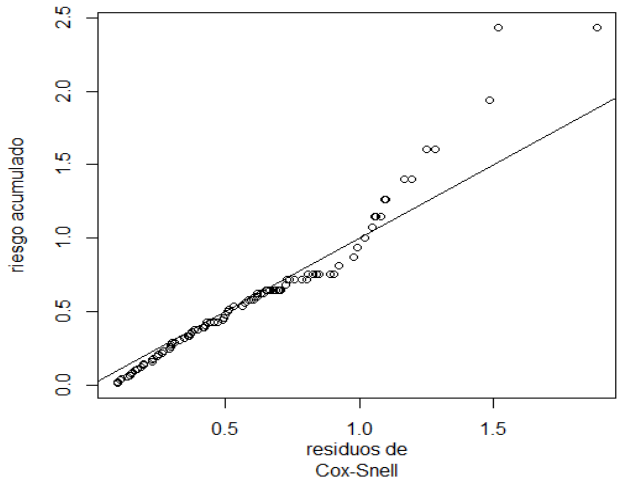
4.2.3 Evaluación del modelo

Una vez se ha seleccionado el modelo más parsimonioso procedemos a evaluarlo. Las pruebas y los diagnósticos gráficos para evaluar el modelo de riesgos proporcionales se pueden basar en los residuales de Cox-Snell, los de Martingala, los de deviance, los de score y los de Schoenfeld. Residuales que se pueden utilizar para evaluar el ajuste global del modelo; la forma funcional apropiada de los predictores continuos; identificar los sujetos que están pobremente predichos por el modelo; identificar los puntos de influencia y verificar el supuesto de riesgo proporcional, (MONTOYA, L., Antonio. *Comparación de dos modelos de regresión en fiabilidad*. Granada, España. 2011, p. 31 Trabajo de gado de Master en Estadística Aplicada. De Granada de España. Facultad de Ciencias. Departamento de Estadística e Investigación Operativa)

Evaluación del ajuste global del modelo

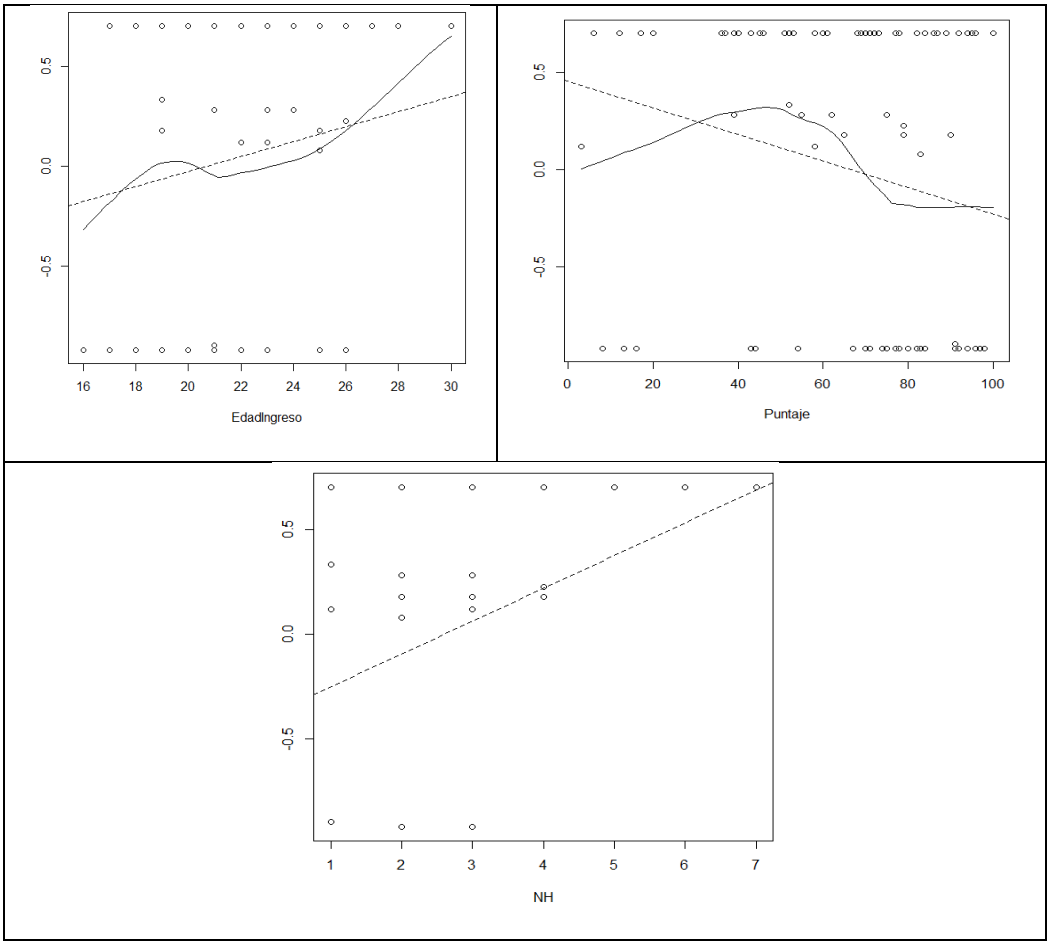
Después de ajustar el modelo, tenemos que calcular los residuos de Cox-Snell con el fin de evaluar el ajuste del modelo de riesgos proporcionales. Si el modelo es correcto y la estimación de los $\beta's$ son cercanas a los valores reales, entonces los residuales deberían corresponder con una muestra censurada de observaciones de una distribución exponencial. Al aplicar el estimador de Nelson-Aalen de la tasa de riesgo acumulado de los residuos de Cox-Snell y gráficarlos versus los residuales de Cox-Snell, si una distribución exponencial ajusta a los datos, entonces, este estimador debería aproximadamente describir una línea de pendiente igual a 1. A partir de la gráfica 17 se puede afirmar que el modelo propuesto ajusta bien a los datos.

Gráfica 17: Residuos de Cox Snell para LME



Evaluación de la forma funcional de las variables explicativas continuas

Gráfica 18: Residuales Martingala



A partir de la información de la gráfica anterior podemos afirmar que las variables número de hermanos, edad de ingreso al programa, puntaje Icfes tienen una relación lineal con el tiempo de deserción. En otras palabras, se evidencia que la función de enlace entre el tiempo de deserción y cada una de las variables explicativas es lineal.

Comprobación de la hipótesis de riesgos proporcionales por cada variable explicativa

Ahora estamos interesados en evaluar la hipótesis de riesgos proporcionales para cada variable explicativa del modelo, es decir se observará si para cada variable explicativa, el riesgo de desertar puede variar con el tiempo, en el sentido de que el correspondiente coeficiente β puede no ser constante, es decir que $\beta(t)$ no depende del tiempo. La función cox.zph de R calcula la prueba de riesgos proporcionales para cada variable explicativa a partir de la correlación entre los residuales estandarizados de Schoenfeld (MONTOYA, L., Antonio. *Comparación de dos modelos de regresión en fiabilidad*. Granada, España. 2011, p. 31 Trabajo de gado de Master en Estadística Aplicada. De Granada de España. Facultad de Ciencias. Departamento de Estadística e Investigación Operativa)

Tabla 15: Prueba de riesgos proporcionales

	rho	Chi-cuadrado	p-valor
EdadIngreso	0.2998	5.722	0.0168
NH	-0.0532	0.227	0.6334
Puntaje	-0.1757	1.744	0.1866
Sexo[T.M]	0.042	0.114	0.7352
GLOBAL	NA	6.286	0.1788

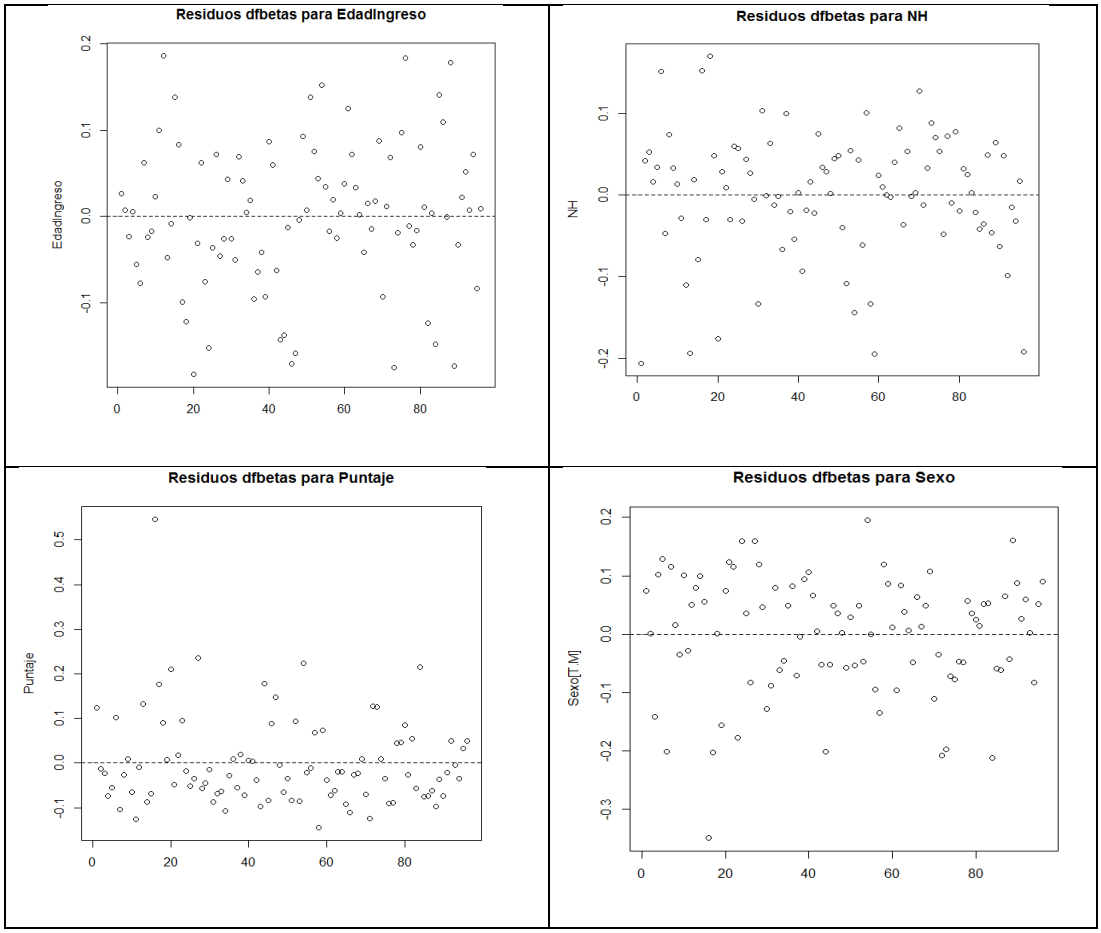
A partir de la información de la tabla anterior podemos decir que no existen evidencias significativas al 5% de que se viole el supuesto de riesgos proporcionales para NH, puntaje, género ni globalmente. Caso contrario sucede con la Edad de ingreso en donde se determina que el efecto de la edad de ingreso depende del tiempo. Así las cosas, se debe tener cuidado a la hora de interpretar el efecto de esta variable en el modelo.

Comprobación de la influencia sobre cada observación en el modelo

Otro uso de los residuos que se nos presenta es el de determinar la influencia de cada observación en el modelo ajustado. Hemos calculado, por medio de los residuos dfbeta, que están implementados en R, el cambio aproximado en el k-ésimo coeficiente (es decir, la k-ésima covariable) si la observación i-ésima se elimina del conjunto de datos y se vuelve a estimar el modelo sin esta

observación. Para cada covariable, se ha representado la observación (en orden de tiempo de fallo registrado) por el cambio de escala aproximada (dividiendo por el error estándar del coeficiente) del coeficiente después de la eliminación de la observación del modelo. Si la supresión de una observación hace que el coeficiente incremente, el residuo dfbeta es negativo y viceversa.

Gráfica 19: Gráficas de los dfbetas



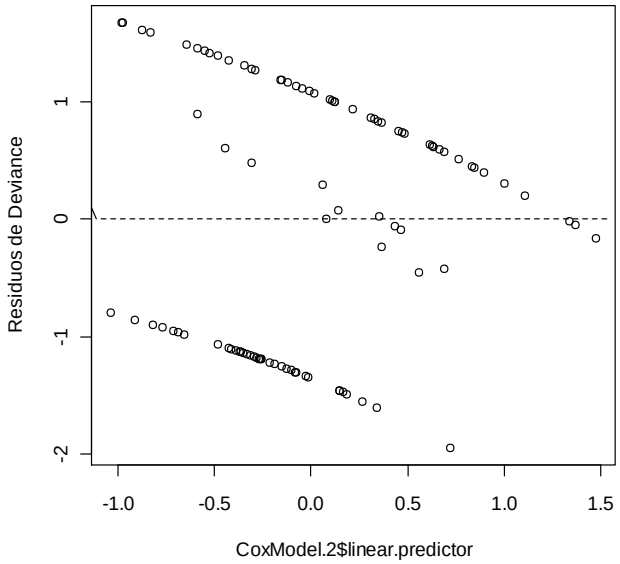
En la gráfica 19 se presentan los residuos *dfbeta* del modelo. Como vemos estos residuos están centrados con respecto al origen, y no presentan patrones definidos. En la variable puntaje del Icfes se presenta un dato alejado del origen, a excepción de esto no se aprecia ninguna irregularidad en las gráficas. Dado lo anterior se procede a efectuar la siguiente validación.

Comprobación de la existencia de outliers en el modelo

Se grafican los residuales de deviance versus el predictor lineal, obteniéndose lo que se presenta en la gráfica 20

No se aprecian residuales alejados del origen. Se puede concluir que no hay presencia de datos atípicos.

Gráfica 20: Residuos de deviance



4.2.4 Pronóstico a partir del modelo

Recordar que la estimación o ajuste del modelo de riesgos proporcionales está dado por la expresión $h_i(t, X, \beta) = h_0(t)\exp(X^T\beta)$. Las dos componentes del modelo $h_0(t)$ (función de riesgo inicial) y $\exp(X^T\beta)$ pueden ser estimadas por separado, se estiman primero los componentes del vector de parámetros y a partir de ellos se construye el estimador de la función de riesgo inicial. Nótese que la función de riesgo inicial depende del tiempo mientras que la segunda componente depende únicamente de las covariables. Sin embargo, en el modelo de riesgos proporcionales de Cox se debe tener en cuenta que los riesgos para dos conjuntos diferentes de valores de los covariables conservan la misma proporción a lo largo del tiempo, de ahí su denominación. (DOMÉNECH., M., Joseph. *Una aplicación del análisis de la* sobrevida *de la salud.* En: Anuarios de Psicología. 55[ta]Edición 1992. Universidad de Barcelona. Pág. 109-141

Con base en los resultados obtenidos hasta el momento, se tiene que el modelo estimado corresponde a:

$$h_i(t, X, \beta) = h_0(t)\exp\left(0.0704(Edadingreso) + 0.238(NH) - 0.0151(Puntaje) + 0.0476Sexo[\text{T.M}]\right)$$

Ecuación 1: Modelo estimado de Cox

El exponente del modelo de Cox, que particularmente para el estudio corresponde a la ecuación 15, se denomina *índice pronóstico (P_I)*, (BOJ, Eva del Val.

Evaluación de la hipotesis de riesgos Proporcionales. En: *El modelo de regresion de Cox..* 1 Ed., 2015. P. *34).*

De manera que la tasa de riesgo se puede expresar a través de este indicador en lugar de utilizar los valores del estudiante para cada variable. En la práctica es mejor usar el *índice de pronóstico centrado* (P_{IC}), que se obtiene centrando las variables predictoras:

$$P_{IC} = \beta_1(X_1 - \overline{X_1}) + \cdots + \beta_p(X_p - \overline{X_p})$$

que para el caso corresponde a:

$$P_{IC} = 0.0704(Edadingreso - 20.07) + 0.238(NH - 2.60) - 0.0151(Puntaje - 67.58) + 0.0476Sexo[T.M]$$

Un índice pronóstico centrado igual a cero corresponde a un estudiante "modal", mujer que ingresa al Programa a los 18 años aproximadamente, con 2 hermanos y con 67.58 de puntaje en el Icfes. La diferencia entre los índices pronósticos de dos estudiantes permite estimar su riesgo relativo (RR), veamos:

$$RR = \frac{h_i(t, X_i, \beta)}{h_j(t, X_j, \beta)} = \frac{h_0(t)\exp(P_{ICindividuoi})}{h_0(t)\exp(P_{ICindividuoj})} = \exp(P_{ICindividuoi} - P_{ICindividuoj})$$

Teniendo en cuenta lo anterior, a continuación, se presenta el RR para cinco estudiantes respecto a un estudiante "modal" del Programa, estudiantes seleccionados aleatoriamente, y que se encuentran cursando el primer semestre del programa, es decir, estudiantes que ingresaron en el I semestre de 2015. Es decir, se pretende resolver la pregunta de ¿cuánto más riesgo de desertar tiene un estudiante que recién ingresó al Programa respecto a un estudiante modal?, veamos. Los datos para los cinco estudiantes seleccionados aleatoriamente se presentan en la tabla 16.

Tabla 16: Pronósticos a partir del modelo

Estudiante	Edad de ingreso	Número de hermanos	Puntaje Icfes	Sexo	P_{IC}	RR respecto al estudiante modal
N° 1	20	1	80	M	-0.526	0.591
N° 2	29	2	74	M	0.437	1.547
N° 3	19	1	57	F	-0.249	0.780
N° 4	18	1	75	F	-0.591	0.554
N° 5	18	2	66	M	-0.217	0.805

A partir de la información de la tabla anterior se puede decir que la situación descrita en el estudiante N° 2, cuyo P_{IC}=0.437, es 1.547 veces más "riesgosa" que

la de un estudiante modal del Programa. Para los otros estudiantes se tiene que la situación de desertar es menos riesgosa que para un estudiante modal del Programa.

4.3 MODELO DE SOBREVIDA PARA LA GRADUACIÓN

Resulta también de interés analizar el tiempo hasta que un estudiante se gradúa del Programa. Al igual que para el modelo de deserción, en primer lugar, se presentará la aplicación de la metodología de Kaplan Meier para estimar la función de sobrevida, con la cual se puede analizar la evolución de la probabilidad de graduación con su respectivo intervalo de confianza. Posteriormente, se construirá el modelo de regresión de Cox para estimar el efecto de las variables de estudio sobre los tiempos de sobrevida a la graduación. En este caso las características del estudio quedan determinadas por:

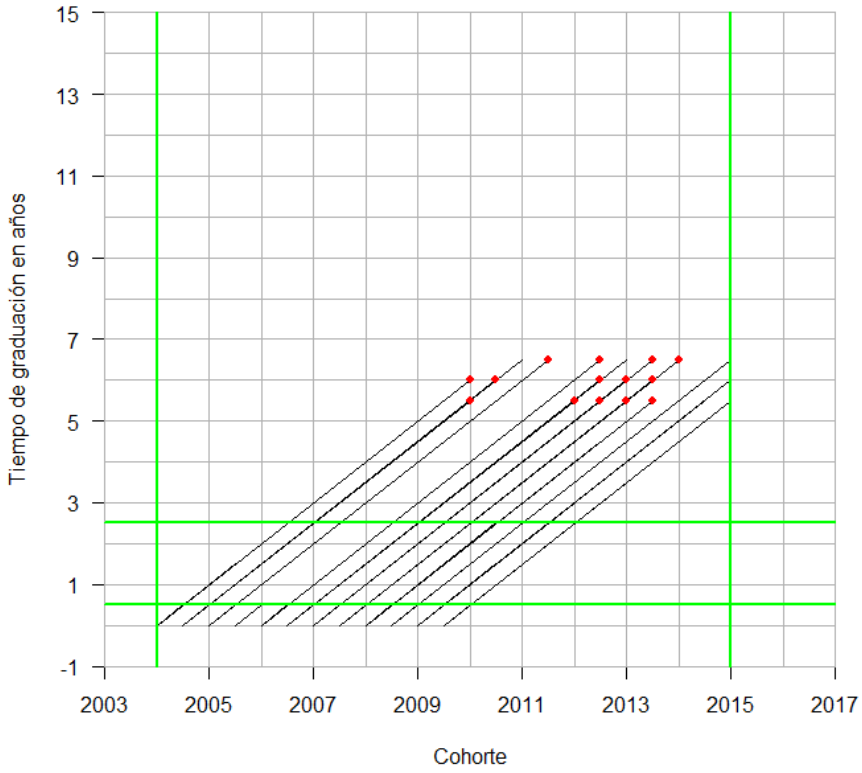
Tabla 17: Estructura del modelo de sobrevida para LME

Objeto de estudio o unidad sobre la cual se registra el evento: estudiante que haya ingresado a la Licenciatura en Matemáticas y Estadística (LME) entre el 2004 y 2009.
Evento de interés o falla*: obtención de grado en el programa por parte del estudiante (graduación).*
***La variable respuesta "Tiempo hasta la graduación"*:** Tiempo hasta que el estudiante se gradúa del programa. Cuya escala de medición es de razón de tipo discreto, ya que se mide en número de semestres cursados hasta que presenta el fenómeno de estudio.
Tiempo de origen del evento*:* primera matrícula del estudiante en el programa entre el primer semestre del 2004 y del 2009
Tiempo inicial del estudio*:* primer semestre del 2004
Tiempo final del estudio*:* primer semestre del 2015
Tipo de censura*:* Tipo I y a derecha. Los individuos entran al estudio en diferentes tiempos, es decir, diferentes cohortes, y el punto final del estudio es el mismo para todos. En este caso, el tiempo de censura para cada estudiante es conocido en el momento que ingreso al estudio, de manera que cada individuo tiene fijo y especificado su tiempo de censura. Se considera como censura al estudiante que desertó o que continúa estudiando.
Tiempo de censura*:* tiempo promedio de graduación en la LME, 13 semestres.

A continuación, se presenta la representación de los estudiantes del estudio mediante el diagrama de Lexis, cuya sintaxis R aparece en el Anexo B.

La gráfica 20 evidencia que los estudiantes bajo estudio no tienen el mismo tiempo de origen, se señala con los puntos rojos aquellos que presentaron el evento y el tiempo en que sucedió (expresado en años). Nótese que la mayoría de estudiantes de gradúan en el tiempo promedio del programa (13 semestres).

Gráfica 21: Diagrama de Lexis para graduación



4.3.1 Función de sobrevida

A continuación, se presenta el análisis del tiempo hasta la graduación, teniendo en cuenta toda la información disponible, es decir tanto los datos censurados como los no censurados. Las probabilidades de sobrevida en cada intervalo, así como la función de sobrevida se calculan con el estimador de Kaplan Meier, teniendo en cuenta que no se asumirá modelo probabilístico para el tiempo hasta la graduación y que se cuenta con datos censurados a derecha.

Gráfica 22: Función de sobrevida estimada

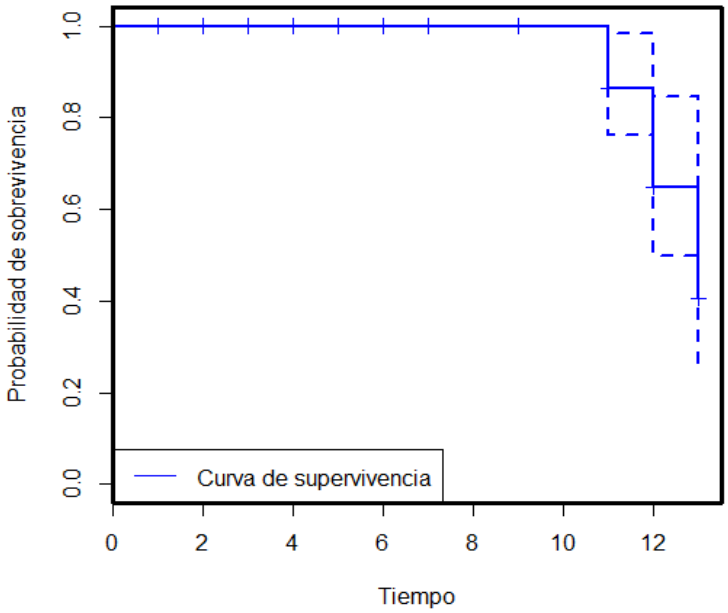


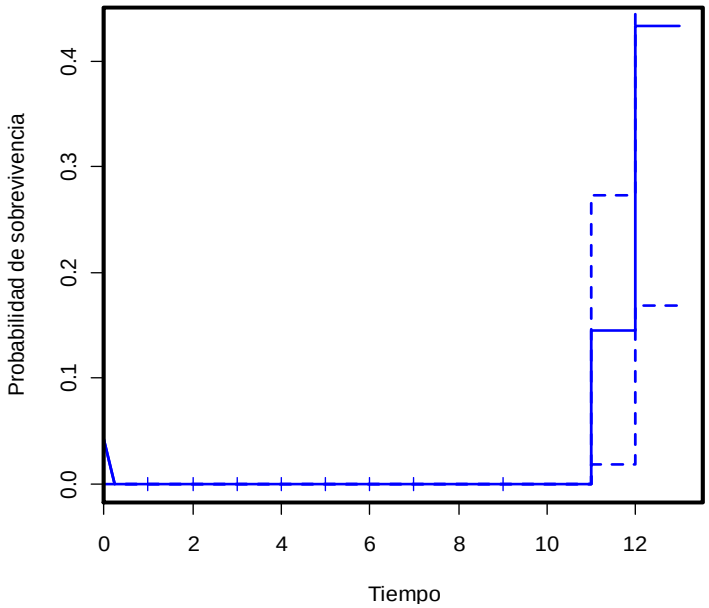
Tabla 18: Estimaciones función de sobrevida para LME

					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
11	37	5	0.865	0.0562	0.761	0.982
12	24	6	0.649	0.0873	0.498	0.844
13	16	6	0.000			

Se observa que la probabilidad de que un estudiante llegue a semestre 11 y se gradúe es del 86% y en semestre 12 es del 64%. Se calculó el cuantil 0.5 a partir de la función de sobrevida encontrándose que es en el semestre 13 donde se tiene un 50% de probabilidad de graduarse, es decir, que aproximadamente el 37% de los estudiante se gradúan entre los semestres 11 y 12.

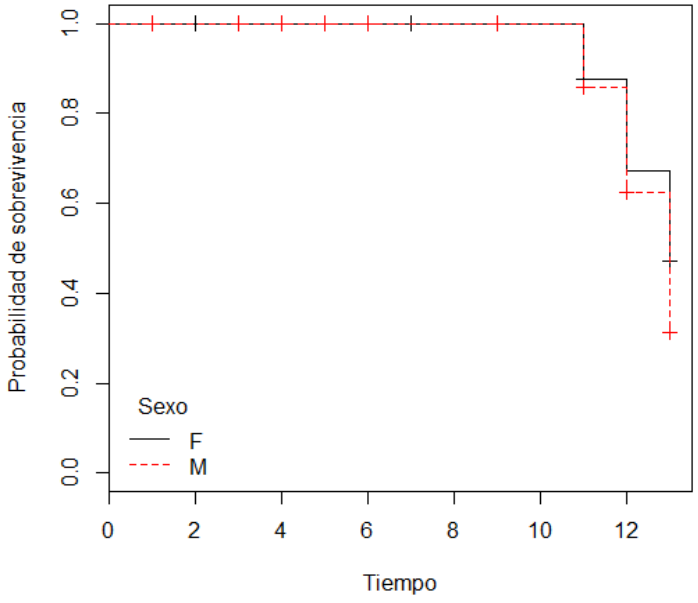
La función de riesgo, presentada en la gráfica 23, también conocida como la tasa instantánea de mortalidad, describe la forma en que cambia la tasa instantánea de la graduación al paso del tiempo. Con base en ésta se puede afirmar que la probabilidad que un estudiante se gradúe justo al terminar el semestre 11 es de 15%, es decir que hay una probabilidad baja de que el tiempo de permanencia extra de un estudiante sea de un semestre. La probabilidad de que el tiempo de permanencia extra en el programa sea de 2 semestres es del 43%.

Gráfica 23: Función de riesgo acumulada - graduación



En el fenómeno de la graduación también se hace necesario un análisis de la función de sobrevida separada por poblaciones de estudio, lo cual se presenta a continuación. En primer lugar, veremos la función de sobrevida por sexo

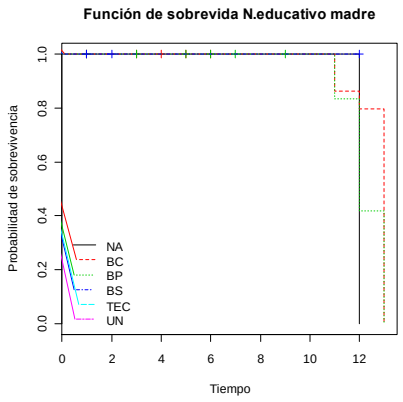
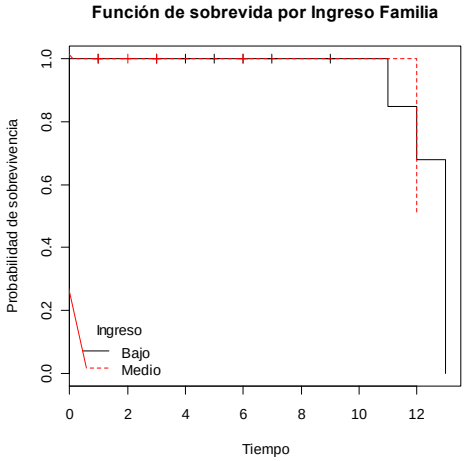
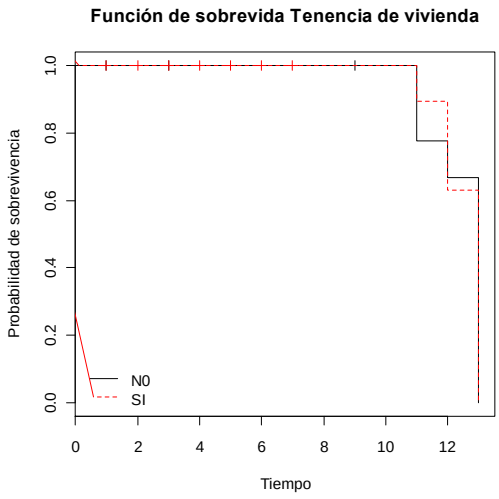
Gráfica 24: Función de sobrevida por sexo



Al comparar las curvas de sobrevida de los hombres y las mujeres se obtiene un valor de Chisq = 0.4 con 1 grado de libertad y un p valor de 0.516, estadística que nos permite afirmar que no podemos rechazar la hipótesis nula de igualdad de las

curvas de sobrevida. Por lo tanto, podemos afirmar que la sobrevida (a la graduación) de hombres y mujeres tiene el mismo comportamiento. También se presentan las gráficas y pruebas de hipótesis para nivel educativo de la madre, ingreso familiar y tenencia de vivienda.

Tabla 19: Comparación de curvas de sobrevida por categorías

Nivel Educativo de la madre	Ingreso Familiar
	
Estadístico de prueba	**Estadístico de prueba**
Chi-cuadrado= 5.9 con 3 grados de libertad, p= 0.114	Chi-cuadrado = 0.2 con 1 grados de libertad, p= 0.659
Tenencia vivienda	
	
Estadístico de prueba	
Chi-cuadrado = 0.2 con 1 grados de libertad, p= 0.635	

Se determina que para las diferentes categorías de las variables la sobrevida (a la graduación) tiene el mismo comportamiento.

4.3.2 Factores relacionados con el riesgo de graduación

Con el objetivo de medir los efectos de las variables consideradas en el estudio como explicativas del fenómeno de la graduación (edad de ingreso al Programa, nivel educativo de la madre, número de hermanos, tenencia o no de vivienda, ocupación a la hora de presentar el Icfes, ingreso familiar, puntaje Icfes estandarizado), a continuación, se presentan las estimaciones del modelo de riesgo proporcional utilizando el modelo semiparamétrico de Cox. Los resultados del modelo inicial se presentan en la tabla 20

Tabla 20: Análisis del modelo semiparamétrico de Cox de las para LI

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
EdadIngreso	-1.15E-01	8.91E-01	1.39E-01	-0.826	0.409
Ingresofami[T.Medio]	3.19E-01	1.38E+00	1.39E+00	0.23	0.818
NEM[T.BP]	-1.45E+00	2.35E-01	1.93E+00	-0.75	0.453
NEM[T.BS]	-5.32E-01	5.88E-01	1.94E+00	-0.274	0.784
NEM[T.TEC]	-1.69E+01	4.66E-08	2.51E+03	-0.007	0.995
NH	-1.38E-01	8.71E-01	6.86E-01	-0.201	0.841
Puntaje	2.08E-02	1.02E+00	1.55E-02	1.341	0.18
Sexo[T.M]	4.09E-01	1.51E+00	6.76E-01	0.604	0.546
tenenciavivi[T.SI]	-4.99E-02	9.51E-01	1.07E+00	-0.047	0.963
Trabajo[T.SI]	NA	NA	0.00E+00	NA	NA

Nótese que ninguna variable resulta significativa a la hora de explicar el tiempo de graduación. Seguidamente se procede a seleccionar el modelo más parsimonioso, utilizando el método hacia adelante y usando como criterio el AIC (Criterio de Información de Akaike), encontrándose que en el modelo óptimo no debe ir ninguna variable, es decir, se explica lo mismo con todas o algunas variables que sin ninguna de ellas. Se puede afirmar entonces que al momento de explicar el fenómeno de la graduación se deben considerar otras variables distintas a las consideradas en este estudio.

5 ANÁLISIS LICENCIATURA EN TECNOLOGÍA

A continuación, se realiza un análisis descriptivo de las variables consideradas, en donde podremos verlas resumidas en porcentajes y algunos estadísticos, como las medias, la mediana, mínimo y máximo, en el apartado gráfico se presentan algunos diagramas de barras para las variables.

La descripción se realiza a través de R-Comander. Todos los detalles de las sentencias en R figuran en el Anexo B al final del documento.

5.1 DESCRIPCIÓN DE LOS DATOS

La tabla 21 muestra el resumen numérico y descriptivo de las variables objeto de estudio para la Licenciatura en Tecnología.

Tabla 21: Resumen descriptivo de las variables de estudio de LT.

Variable	Unidades de medición /codificación	Resumen numérico
Género (n = 128)	F = femenino M = masculino	39 (30.46%) 89 (69.53%)
Nivel educativo de la madre	Básica Primaria = BP Básica Secundaria = BS Técnico = TEC Universitario = UN Sin Clasificar = SC	59 (46.46%) 35 (27.56%) 4 (3.15%) 9 (7.03%) 21 (16.40%)
Numero de Hermanos		Media = 2.6666, Desviación estándar = 1.611 Coeficiente de Variación = 0.698
Edad de ingreso al Programa	En años	Media = 20 Desviación estándar = 2.862 Coeficiente de Variación = 0.145
Edad de Presentación del Examen de Estado	En Años Sin Clasificar = SC	Media = 17.61 Min – Max = 12 - 29 Mediana = 17
Ingreso Familiar al Presentar el Examen de Estado	Bajo Medio Bajo: ingresos de 0,1,2 SMLV; Medio=3,4,5 SMLV, Alto: >5 SMLV Sin clasificar = SC	Bajo = 89 (69.53%) Medio = 19 (14.84%) SC = 20 (15.625%)
Trabajaba al presentar el examen de estado	Si = S No = N Sin clasificar = SC	S = 5 (3.09%) N = 105 (82.68%) SC = 18 (14.17%)
Vivienda Propia	Si = Si No = No Sin Clasificar = SC	S = 89 (69.53%) N =28 (22.05%) SC = 11 (8,59%)
Puntaje en el Examen de Estado	De 1 – 100 puntos porcentuales Sin Clasificar = SC	Media = 59.65 Min – Max = 4 - 100 Mediana = 62
Estado del estudiante	Matriculado = MT Desertor = DE Graduado = GR Terminación académica = TA Retirado con cupo reservado =RCR	MT = 11 (8.87%) DE = 78 (60.93%) GR = 30 (23.43%) TA = 6 (4.68%) RCR = 3 (2.242%)
Tiempo Hasta la Deserción	Semestres Académicos (16 semanas)	Media = 6.93 Min – Max = 0.68 – 19.9 Mediana = 9

Estado del estudiante Dicotomizado	Desertor=D, No Desertor=ND	D = 76 (60%) ND = 52 (40%)
En la consideración de estas medidas se debe tener en cuenta que debido a las censuras en los datos de esta variable, las estadísticas de resumen pueden no tener las propiedades estadísticas deseadas tales como insesgamiento. Por ejemplo, la media muestral ya no es un estimador insesgado de la media poblacional (del tiempo de sobrevida).		
Motivo de deserción académica	Perdió cupo Articulo 80 Literal B (Quien teniendo un promedio aritmético acumulado inferior a tres cero (3.0), obtenga un promedio aritmético semestral inferior a dos cero (2.0)) = LB Perdió cupo Articulo 80 Literal C (Quien pierda una asignatura que curse en calidad de repitente siendo su promedio acumulado inferior a tres cero (3.0). En el caso en que el promedio aritmético acumulado sea igual o superior a tres cero (3.0), la podrá cursar por tercera y última vez) = LC Perdió cupo Articulo 80 Literal D (Quien pierda en un mismo periodo académico dos asignaturas que se cursen en calidad de repitente) = LD Perdió cupo Articulo 80 Literal E (Quien pierda una asignatura por tercera vez) = LE	LB = 9 (7.26%) LC = 9 (7.26%) LD = 5 (4.03%) LE = 5 (4.03%)
Motivo de deserción no académica	No matriculado = NM Retiro definitivo = RD Sin Clasificar = SC	NM = 0% RD = 41 (33.06%) SC = 7 (5.7%)
Tipo de deserción	Precoz = P Temprana = Te Tardía = Ta	P = 45 (59.2%) Te = 11 (14.48%) Ta = 20 (23.30%)

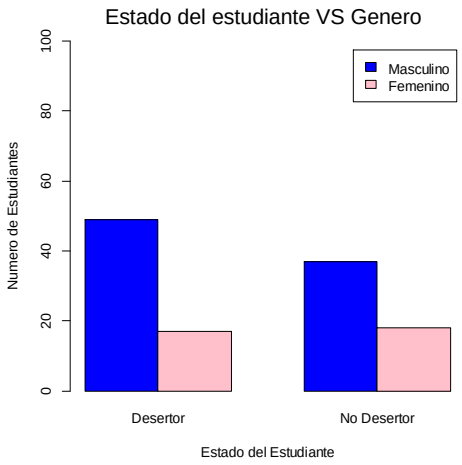
Con los resultados anteriores podemos ver que, de los estudiantes seleccionados en la muestra, la mayoría son hombres (69.29%). La edad media al momento de presentar el examen de Estado fue de 17 años. El 70.08% tenían vivienda propia y el 82.68% no trabajaban al momento de presentar el examen de Estado.

El puntaje medio en el examen de estado fue de 59.65 en una escala de 0 a 100. El 83.01% tenía como ingreso familiar al momento de presentar el examen de Estado, 1.2 Salarios Mínimos Mensuales Legales Vigentes (SMMLV). En estos mismos tiempos, se ha graduado el 28.3% de los estudiantes. Se puede ver que el 46.46% de los estudiantes tenían madres con nivel educativo de básica primaria.

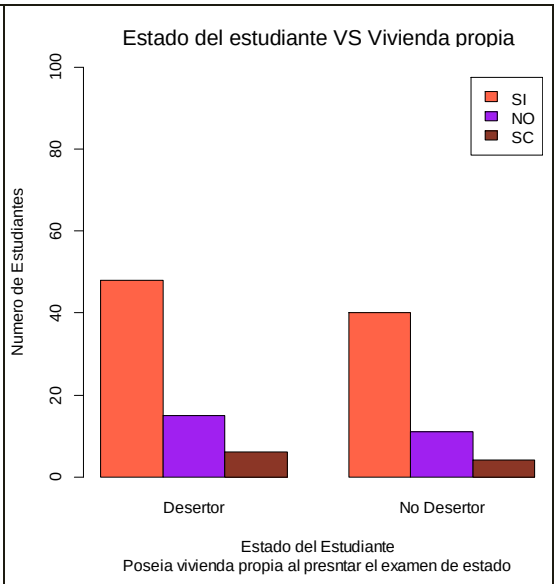
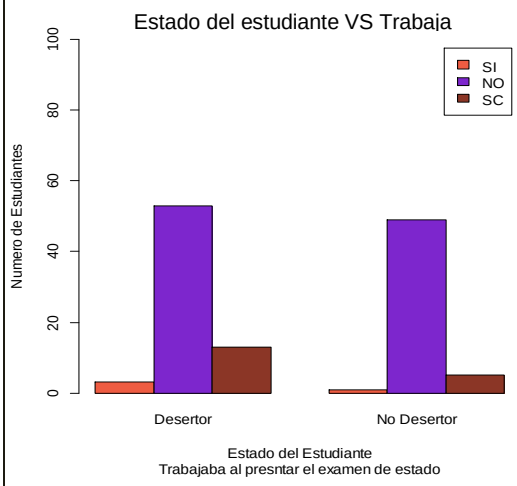
En promedio los estudiantes tenían dos hermanos al momento de presentar el examen de Estado. En resumen, se puede decir que un estudiante modal del programa de Licenciatura Tecnología es de género masculino, cuyas madres cuenta en su mayoría con básica primaria, con 3 hermanos en promedio, 20 años promedio de edad al momento de ingresar al programa, de ingresos familiares (al momento de presentar el Icfes) bajos (0,1 o 2 SMMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 59.65 sobre 100 puntos.

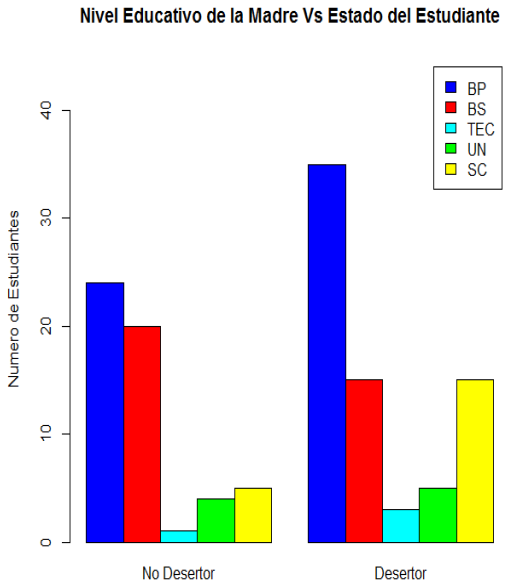
A continuación, se presenta una descripción bivariada de las variables objeto de estudio. En primer lugar, se pretende observar si las variables categóricas estudiadas se asocian con el estado del estudiante (desertor o no desertor), para lo cual se utilizarán pruebas de independencia y de asociación. Lo anterior se efectúa con el fin de advertir que variables están asociadas al hecho de desertar.

Tabla 22: Resumen bivariado de algunas variables de interés de LT

<table>
<tr><th colspan="2" rowspan="2">Variable</th><th colspan="2">Estado del estudiante</th><th colspan="2">Estadístico</th><th rowspan="2">Gráfica</th></tr>
<tr><th>Desertor</th><th>No desertor</th><th>Prueba Chi-cuadrado</th><th>Valor p</th></tr>
<tr><td rowspan="2">Género</td><td>M</td><td>49</td><td>37</td><td rowspan="2">0.7088, df = 1</td><td rowspan="2">0.3998</td><td rowspan="2"> 

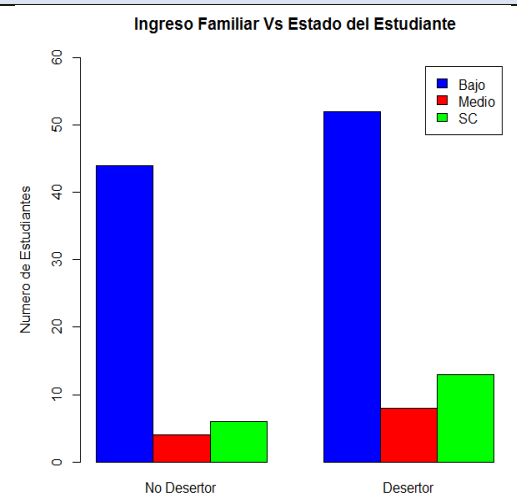
Gráfica 25: Diagrama para Estado del estudiante y Genero</td></tr>
<tr><td>F</td><td>17</td><td>18</td></tr>
<tr><td colspan="7">Se determina que el “Estado del Estudiante” y “Género” son independientes. Es decir, podemos concluir que el estado del estudiante no depende del género.</td></tr>
</table>

Vivienda Propia	Si	48	40	0.1641, df = 2	0.9212	 **Gráfica 26: Diagrama para Estado del estudiante y Vivienda Propia**
	No	15	11			
	SC	6	4			
Se determina que el “Estado del Estudiante” y “Vivienda propia” son independientes. Es decir, podemos concluir que el estado del estudiante no depende de la vivienda.						
Trabajaba al Presentar el Examen de Estado	Si	3	1	21.1218	0.1739	 **Gráfica 27: Diagrama para Estado del estudiante y Trabajaba”**
	No	53	49			
	SC	13	5			
	SC	13	5			
Se determina que el “Estado del Estudiante” y “Trabajaba al presentar el examen de estado” son independientes. Es decir, podemos concluir que el estado del estudiante no depende de si el estudiante trabajaba o no al momento de presentar el examen de estado						

Variable		Estado del estudiante		Estadístico		Gráfica
		No Desertor	Desertor	Prueba Chi-cuadrado	Valor p	
Nivel educativo madre	BP	24	35	6.1719, df = 4	0.1867	
	BS	20	15			
	TEC	1	3			
	UN	4	5			
	SC	5	15			

Gráfica 28: Diagrama para Estado del estudiante y Nivel educativo de la madre

Se determina que el “Estado del Estudiante” y el “Nivel educativo de la madre” son independientes

Variable		Estado del estudiante		Estadístico		Gráfica
		No Desertor	Desertor	Prueba Chi-cuadrado	Valor p	
Ingreso familiar	BAJO	44	52	1.7762, df = 2	0.4114	
	MEDIO	4	8			
	SC	6	13			

Gráfica 29: Diagrama para Estado del estudiante e Ingreso familiar

Se determina que el “Estado del Estudiante” y el “ingreso familiar” son independientes

A partir de la información anterior se puede afirmar que el género del estudiante, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el hecho de o no desertar del Programa.

5.2 MODELO DE SOBREVIDA PARA LA DESERCIÓN

Como el propósito del estudio es identificar los factores que están relacionados al tiempo de la deserción de un estudiante de la Licenciatura en Tecnología En primer lugar, se presentará la aplicación de la metodología de Kaplan Meier para estimar la función de sobrevida, con la cual se puede analizar la evolución de la probabilidad de deserción con su respectivo intervalo de confianza. Posteriormente, se construirá el modelo de regresión de Cox para estimar el efecto de las variables de estudio sobre los tiempos de sobrevida a la deserción. Recordemos las características del estudio:

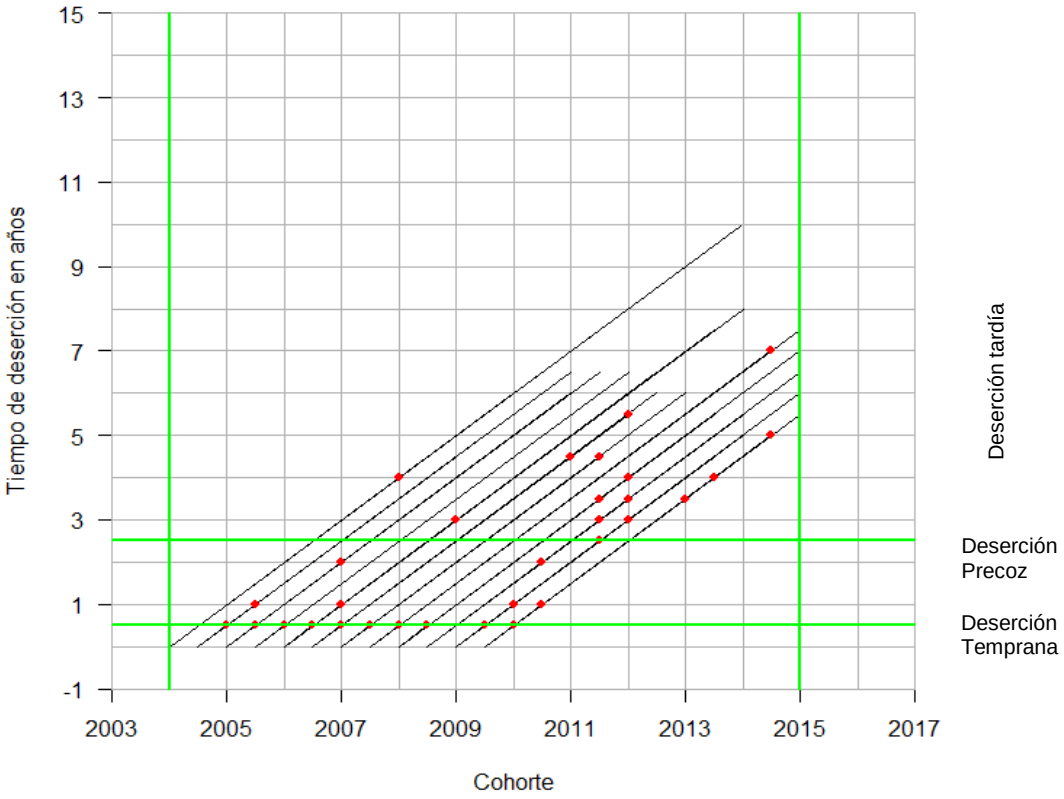
Tabla 23: Estructura del modelo de sobrevida para LT

Objeto de estudio o unidad sobre la cual se registra el evento: estudiante que haya ingresado a la Licenciatura en Tecnología (LT) entre el 2004 y 2009.
Evento de interés o falla*: Abandono del programa por parte del estudiante (deserción).*
***La variable respuesta "Tiempo hasta la deserción"*:** Tiempo hasta que el estudiante se va del programa ya sea porque lo abandona por causas académicas o no académicas. Cuya escala de medición es de razón de tipo discreto, ya que se mide en número de semestres cursados hasta que presenta el fenómeno de estudio.
Tiempo de origen del evento*:* primera matrícula del estudiante en el programa entre el primer semestre del 2004 y del 2009
Tiempo inicial del estudio*:* primer semestre del 2004
Tiempo final del estudio*:* primer semestre del 2015
Tipo de censura*:* Tipo I y a derecha. Los individuos entran al estudio en diferentes tiempos, es decir, diferentes cohortes, y el punto final del estudio es el mismo para todos. En este caso, el tiempo de censura para cada estudiante es conocido en el momento que ingreso al estudio, de manera que cada individuo tiene fijo y especificado su tiempo de censura. Se considera como censura al estudiante graduado o que continúa estudiando.
Tiempo de censura*:* tiempo promedio de graduación en la LT, 12 semestres.

A continuación, se presenta la representación de los estudiantes del estudio mediante el diagrama de Lexis, cuya sintaxis R aparece en el anexo B. Diagrama

que refleja el tiempo calendario en el eje horizontal y la longitud del tiempo de vida, representada por una línea a 45°. El tiempo que un individuo pasa en el estudio es representado por la altura del rayo en el eje vertical.

Gráfica 30: Diagrama de Lexis - Deserción para LT



La gráfica 24 evidencia que los estudiantes bajo estudio no tienen el mismo tiempo de origen, se señala con los puntos rojos aquellos que presentaron el evento y el tiempo en que sucedió (expresado en años). Nótese que la mayoría de estudiantes desertan tardíamente, seguido de la deserción precoz, y unos pocos se registran con deserción temprana.

A continuación, se presenta el análisis del tiempo hasta la deserción, teniendo en cuenta toda la información disponible, es decir tanto los datos censurados como los no censurados. Las probabilidades de sobrevida en cada intervalo, así como la función de sobrevida se calculan con el estimador de Kaplan Meier, teniendo en cuenta que no se asumirá modelo probabilístico para el tiempo hasta la deserción y que se cuenta con datos censurados a derecha.

5.2.1 Función de sobrevida

El estimador de Kaplan Meier para la función de sobrevida es obtenido en el paquete estadístico R mediante la función survfit, en el anexo B se muestra la sintaxis R. En el gráfico 25 y la tabla 20 se pueden observar las probabilidades de sobrevida en cada intervalo. Time representa el tiempo para el que se presenta la información, n risk indica la cardinalidad del conjunto en riesgo o del número de individuos que continua en el estudio al tiempo correspondiente, n event corresponde al número de fallas que se presentan entre cada tiempo, survival indica el valor que toma la función de sobrevida estimada por el método de Kaplan Meier en el tiempo correspondiente, std err corresponde al error estándar estimado para la función de sobrevida en el tiempo respectivo y lower y upper 95% CI denotan el límite inferior y superior respectivamente del intervalo de confianza al 95% para la función de sobrevida.

Gráfica 31: Función de sobrevida estimada

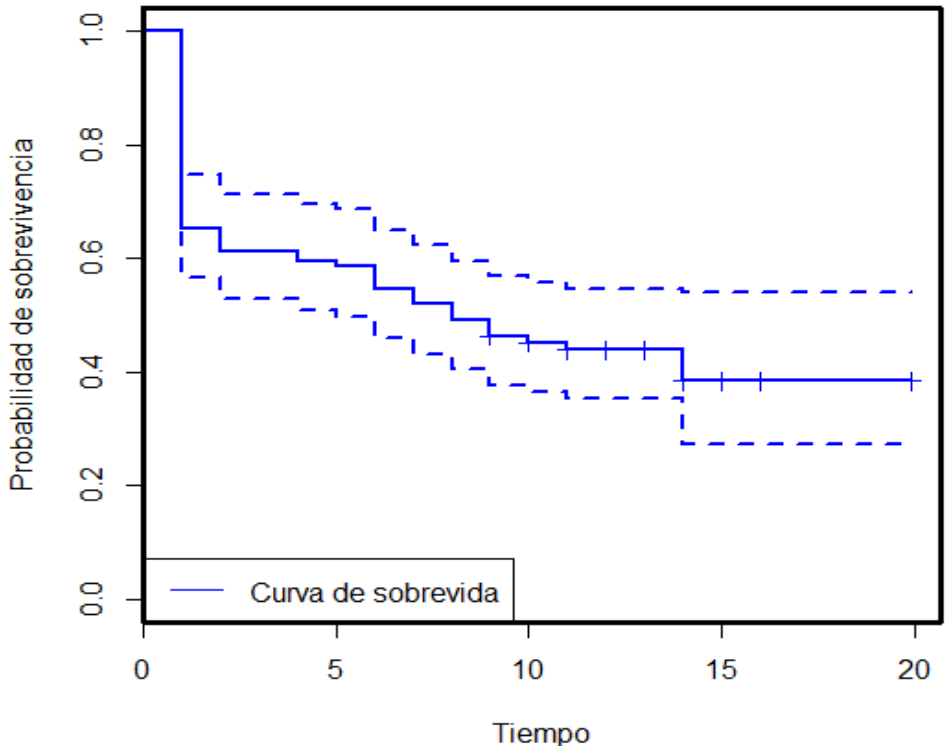


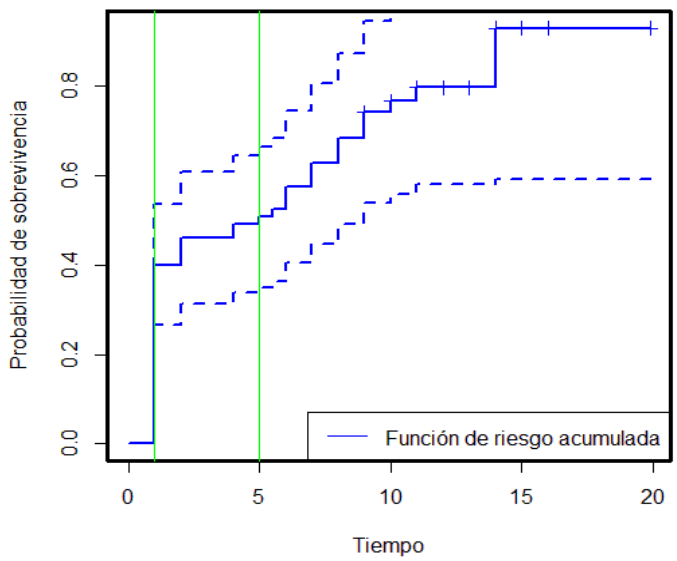
Tabla 24: Estimaciones función de sobrevida para LT

					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	103	34	0.670	0.0463	0.585	0.767
2	69	4	0.631	0.0475	0.544	0.731
4	65	2	0.612	0.0480	0.524	0.713
5	63	1	0.602	0.0482	0.514	0.704
6	61	3	0.563	0.0489	0.0475	0.668
7	58	3	0.534	0.0492	0.446	0.640
8	55	3	0.505	0.0493	0.417	0.611
9	52	3	0.476	0.0492	0.388	0.583
10	42	1	0.464	0.0493	0.377	0.572
11	34	1	0.451	0.0497	0.363	0.560
14	8	1	0.394	0.0683	0.281	0.554

Se observa que a medida que aumenta el número de semestres la probabilidad de sobrevida disminuye. Nótese que es más rápido el decrecimiento en los primeros cinco semestres y tiende a estabilizarse a partir del 6 semestre donde está alrededor del 56.3%. Se deja de presentar información para los tiempos mayores que 9 ya que es el tiempo mayor de deserción para el estudio, pero con la función de sobrevida estimada se puede calcular la misma para cualquier tiempo mayor que cero. Se calculó el cuantil 0.5 a partir de la función de sobrevida, ver en el anexo B la sintaxis R, encontrándose que hay una probabilidad del 50% de “estar vivo” o “no presentar deserción” hasta el noveno semestre y hay una probabilidad del 75% de “estar vivo” o “no presentar deserción” hasta el sexto semestre.

La función de riesgo, también conocida como la tasa instantánea de mortalidad, describe la forma en que cambia la tasa instantánea de la deserción al paso del tiempo. La función de riesgo acumulado, permite tener información del comportamiento del riesgo a lo largo del tiempo. A continuación, se presenta la correspondiente a la Licenciatura en Tecnología, para efectos de interpretación se indica en la gráfica el semestre 1 (deserción precoz) y semestre 5 (deserción temprana).

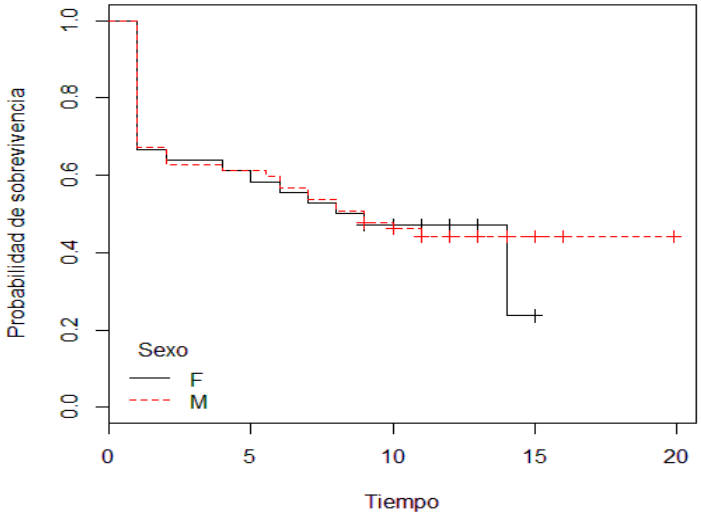
Gráfica 32: Función de riesgo acumulada - deserción



Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 40%, es decir que hay una probabilidad alta que deserte precozmente del Programa.

Teniendo en cuenta que el fenómeno de la deserción no se está estudiando de manera aislada, se hace necesario un análisis de la función de sobrevida separado por poblaciones de estudio, lo cual se presenta a continuación. En primer lugar, veremos la función de sobrevida por sex

Gráfica 33. Función de sobrevida por sexo



Al comparar las curvas de sobrevida de los hombres y las mujeres se obtiene un valor de Chisq = 0 con 1 grado de libertad y un p valor de 0.917, estadística que nos permite afirmar que no podemos rechazar la hipótesis nula de igualdad de las curvas de sobrevida. Por lo tanto, podemos afirmar que la sobrevida (a la deserción) de hombres y mujeres tiene el mismo comportamiento.

Tal como se presenta en el Anexo B se efectuaron las pruebas de hipótesis para nivel educativo de la madre, ingreso familiar y tenencia de vivienda, encontrándose que para las diferentes categorías de las variables la sobrevida (a la deserción) tiene el mismo comportamiento. En la tabla 25 se observa el análisis para la función de sobrevida por sexo.

Tabla 25: Análisis de la función de sobrevida por sexo

Sexo=F					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	36	12	0.667	0.0786	0.5292	0.840
2	24	1	0.639	0.0801	0.4998	0.817
4	23	2	0.611	0.0812	0.4709	0.793
5	22	1	0.583	0.0822	0.4426	0.769
6	21	1	0.556	0.0828	0.4148	0.744
7	20	1	0.528	0.0832	0.3875	0.719
8	19	1	0.500	0.0833	0.3607	0.693
9	18	1	0.472	0.0832	0.3343	0.667
14	2	1	0.236	0.1721	0.0566	0.985
Sexo=M					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	67	22	0.672	0.0574	0.568	0.794
2	45	3	0.627	0.0591	0.521	0.754
4	42	1	0.612	0.0595	0.506	0.740
5.5	41	1	0.597	0.0599	0.490	0.727
6	40	2	0.567	0.0605	0.460	0.699
7	38	2	0.537	0.0609	0.430	0.671
8	36	2	0.507	0.0611	0.401	0.642
9	34	2	0.478	0.0610	0.372	0.614
10	29	1	0.461	0.0611	0.0356	0.598
11	24	1	0.442	0.0615	0.336	0.581

De igual manera se efectuó la prueba para la hipótesis nula de igualdad de las curvas de sobrevida respecto a si trabajaban o no a la hora de presentar el Icfes, encontrándose que, con Chisq= 0.6 con 1 grados de libertad y un p valor de 0.434, estadística que nos permite afirmar que no podemos rechazar la hipótesis nula de igualdad de las curvas de sobrevida. A continuación, se presenta la información que sustenta la afirmación.

Gráfica 34: Función de sobrevida por ocupación al presentar el Icfes

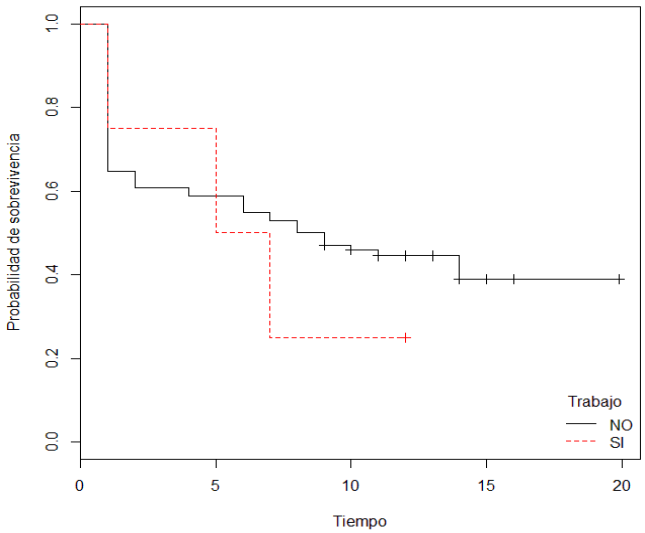


Tabla 26: Análisis de la función de sobrevida para "Trabajo"

Trabajo=NO					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	99	33	0.667	0.0474	0.580	0.766
2	66	4	0.626	0.0486	0.538	0.729
4	62	2	0.606	0.0491	0.517	0.710
5.5	60	1	0.596	0.0493	0.507	0.701
6	59	3	0.566	0.0498	0.476	0.672
7	56	2	0.545	0.0500	0.456	0.653
8	54	3	0.515	0.0502	0.426	0.624
9	51	3	0.485	0.0502	0.396	0.594
10	41	1	0.473	0.0504	0.384	0.583
11	3	1	0.459	0.0508	0.369	0.570
14	8	1	0.401	0.0697	0.286	0.564
Trabajo=SI					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
1	4	1	0.75	0.217	0.4259	1
5	3	1	0.50	0.250	0.1877	1
7	2	1	0.25	0.217	0.0458	1

Se puede inferir que la probabilidad de graduarse del Programa es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable y su consideración en el modelo de Cox (se presenta más adelante) se deben tomar con precaución ya que, en primer lugar, no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad. En segundo lugar, de los 104 estudiantes del estudio tan sólo 4 presentaron la condición de estar trabajando a la hora de presentar el Icfes.

5.2.2 Factores relacionados con el riesgo de deserción

Con el objetivo de medir los efectos de las variables consideradas en el estudio como explicativas del fenómeno de la deserción (edad de ingreso al Programa, nivel educativo de la madre, número de hermanos, tenencia o no de vivienda, ocupación a la hora de presentar el Icfes, ingreso familiar, puntaje Icfes estandarizado), a continuación, se presentan las estimaciones del modelo de riesgo proporcional utilizando el modelo semiparamétrico de Cox. Los resultados del modelo inicial se presentan en la tabla 27

Tabla 27: Análisis del modelo semiparamétrico de Cox para LT

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
EdadIngreso	-0.035267	0.965348	0.062335	-0.566	0.5716
Ingresofami[T.Medio]	0.490339	1.632869	0.653786	0.750	0.4533
NEM[T.BS]	-0.713145	0.490100	0.417635	-1.708	0.0877
NEM[T.TEC]	-0.736344	0.478862	0.837339	-0.879	0.3792
NEM[T.UN]	-0.853316	0.426000	0.865212	-0.986	0.3240
NH	-0.129652	0.878401	0.124322	-1.043	0.2970
Puntaje	-0.006629	0.993393	0.006003	-1.104	0.2694
Sexo[T.M]	0.039172	1.039950	0.294454	0.133	0.8942
tenenciavivi[T.SI]	-0.291663	0.747020	0.382027	-0.763	0.4452
Trabajo[T.SI]	0.437844	1.549363	0732774	0.598	0.5502

Nótese que ninguna variable resulta significativa a la hora de explicar el fenómeno de la deserción. Seguidamente se procede a seleccionar el modelo más parsimonioso, utilizando el método hacia adelante y usando como criterio el AIC (Criterio de Información de Akaike), encontrándose que en el modelo óptimo no debe ir ninguna variable, es decir, se explica lo mismo con todas, algunas variables que sin ninguna de ellas. Se puede afirmar entonces que al momento de explicar el fenómeno de la deserción se deben considerar otras variables distintas a las consideradas en este estudio

5.3 MODELO DE SOBREVIDA PARA LA GRADUACIÓN

Como otro de nuestros objetivos es identificar los factores que están relacionados con la graduación de un estudiante de Licenciatura en Matemáticas y Estadística y Licenciatura en Tecnología. En primer lugar, se presentará la aplicación de la metodología de Kaplan Meier para estimar la función de sobrevida, con la cual se puede analizar la evolución de la probabilidad de graduación con su respectivo intervalo de confianza. Posteriormente, se construirá el modelo de regresión de Cox para estimar el efecto de las variables de estudio sobre los tiempos de sobrevida a la graduación. Recordemos las características del estudio:

Tabla 28: Estructura del modelo de sobrevida para LT

Objeto de estudio o unidad sobre la cual se registra el evento: estudiante a Licenciatura en Tecnología (LT) entre el 2004 y 2009.
Evento de interés o falla*: Abandono del programa por parte del estudiante (Graduación).*
***La variable respuesta "Tiempo hasta la graduación"*:** Tiempo hasta que el estudiante se va del programa porque se ha graduado. Cuya escala de medición es de razón de tipo discreto, ya que se mide en número de semestres cursados hasta que presenta el fenómeno de estudio.
Tiempo de origen del evento*:* primera matrícula del estudiante en el programa
Tiempo inicial del estudio*:* primer semestre del 2004
Tiempo final del estudio*:* primer semestre del 2015
Tipo de censura*:* Tipo I y a derecha. Los individuos entran al estudio en diferentes tiempos, es decir, diferentes cohortes, y el punto final del estudio es el mismo para todos. En este caso, el tiempo de censura para cada estudiante es conocido en el momento que ingreso al estudio, de manera que cada individuo tiene fijo y especificado su tiempo de censura. Se considera como censura al estudiante que abandona sus estudios por causas académicas o no académicas.
Tiempo de censura*:* tiempo promedio de deserción en LT 2 semestres.

A continuación, se muestra la representación de los estudiantes del estudio mediante el diagrama de Lexis, cuya sintaxis R aparece en el anexo B. Diagrama que refleja el tiempo calendario en el eje horizontal y la longitud del tiempo de vida, representada por una línea a 45°. El tiempo que un individuo pasa en el estudio es representado por la altura del rayo en el eje vertical.

Gráfica 35: Diagrama de Lexis - Graduación

La gráfica 35 evidencia que los estudiantes bajo estudio no tienen el mismo tiempo de origen, se señala con los puntos rojos aquellos que presentaron el evento y el tiempo en que sucedió (expresado en años). Nótese que la mayoría de estudiantes de gradúan en el tiempo promedio del programa (12 semestres), seguido de los estudiantes que están por encima de este promedio, unos pocos se gradúan en un tiempo muy superior al promedio.

A continuación, se presenta el análisis del tiempo hasta la deserción, teniendo en cuenta toda la información disponible, es decir tanto los datos censurados como los no censurados. Las probabilidades de sobrevida en cada intervalo, así como la función de sobrevida se calculan con el estimador de Kaplan Meier, teniendo en cuenta que no se asumirá modelo probabilístico para el tiempo hasta la graduación y que se cuenta con datos censurados a derecha.

5.3.1 Función de sobrevida

El estimador de Kaplan Meier para la función de sobrevida es obtenido en el paquete estadístico R mediante la función survfit, en el anexo B se muestra la sintaxis R. En el gráfico 36 y la tabla 29 se pueden observar las probabilidades de sobrevida en cada intervalo. Time representa el tiempo para el que se presenta la información, n risk indica la cardinalidad del conjunto en riesgo o del número de individuos que continua en el estudio al tiempo correspondiente, n event corresponde al número de fallas que se presentan entre cada tiempo, survival

indica el valor que toma la función de sobrevida estimada por el método de Kaplan Meier en el tiempo correspondiente, std err corresponde al error estándar estimado para la función de sobrevida en el tiempo respectivo y lower y upper 95% CI denotan el límite inferior y superior respectivamente del intervalo de confianza al 95% para la función de sobrevida.

Gráfica 36: Función de sobrevida estimada

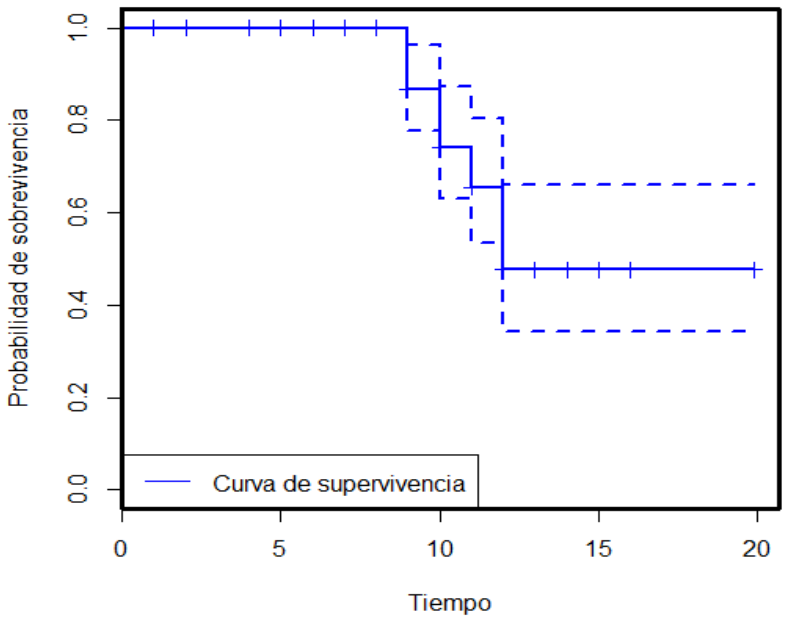


Tabla 29: Estimaciones función de sobrevida para LT

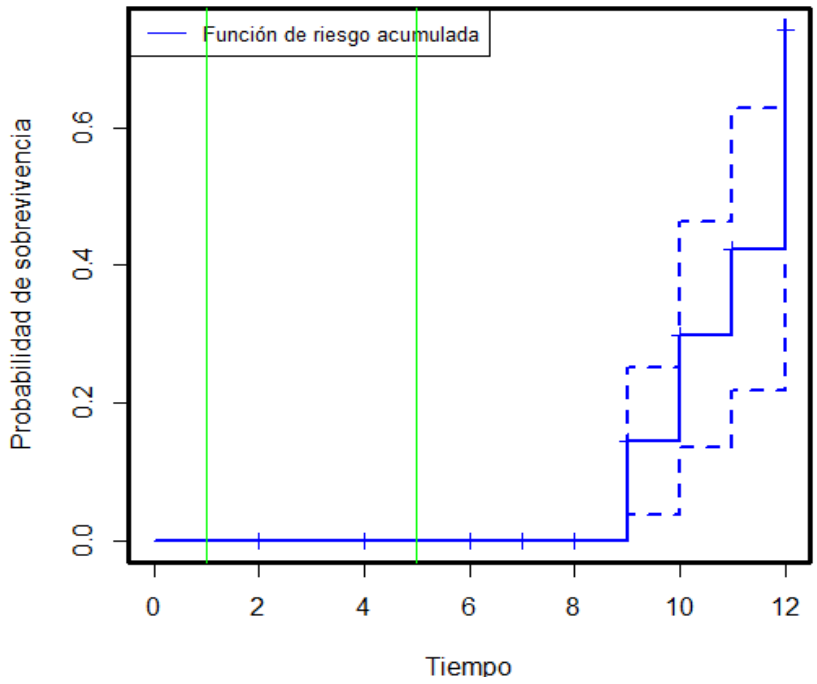
					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
9	52	7	0.865	0.0473	0.777	0.963
10	42	6	0.742	0.0619	0.630	0.874
11	34	4	0.654	0.0683	0.533	0.803
12	22	6	0.476	0.0795	0.343	0.660

Se observa que la probabilidad de que un estudiante llegue a semestre 9 y se gradúe es del 86.5% y en semestre 12 es del 47.6%. Se calculó el cuantil 0.5 a partir de la función de sobrevida encontrándose que es en el semestre 12 donde se tiene un 50% de probabilidad de graduarse.

La función de riesgo, también conocida como la tasa instantánea de mortalidad, describe la forma en que cambia la tasa instantánea de la graduación al paso del

tiempo. La función de riesgo acumulado, permite tener información del comportamiento del riesgo a lo largo del tiempo. A continuación, se presenta la correspondiente a la Licenciatura en Tecnología.

Gráfica 37: Función de riesgo acumulada – graduación



Con base en la gráfica 37, se puede afirmar que la probabilidad que un estudiante se gradué justo al terminar el semestre 9 es de 14.5%, es decir que hay una probabilidad baja de que el tiempo de permanencia extra de un estudiante sea de un semestre. La probabilidad de que el tiempo de permanencia extra en el programa sea de 2 semestres es del 39%.

Como se ha realizado en las funciones de sobrevida en la deserción, en la graduación también estudian otras variables asociadas al estudio, se hace necesario un análisis de la función de sobrevida separado por poblaciones de estudio, lo cual se presenta a continuación. En primer lugar, veremos la función de sobrevida por sexo:

Gráfica 38: Función de sobrevida por sexo

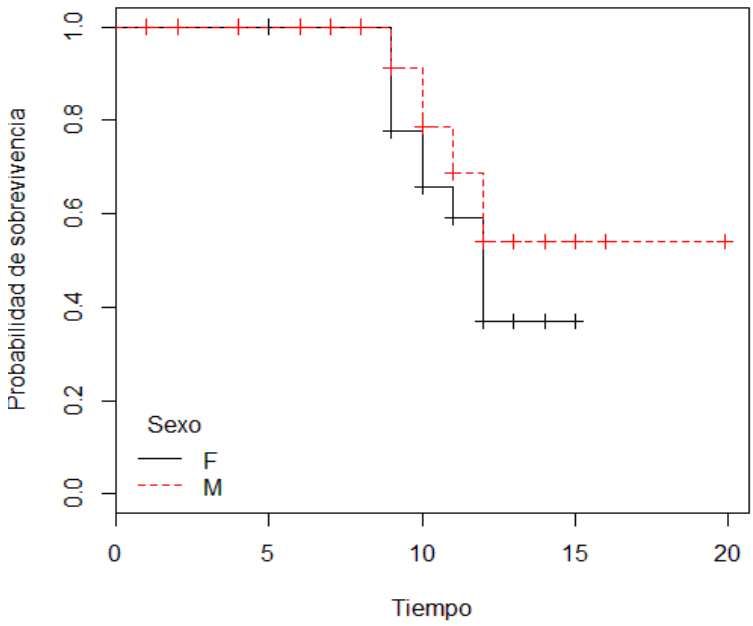


Tabla 30: Análisis de la función de sobrevida por sexo

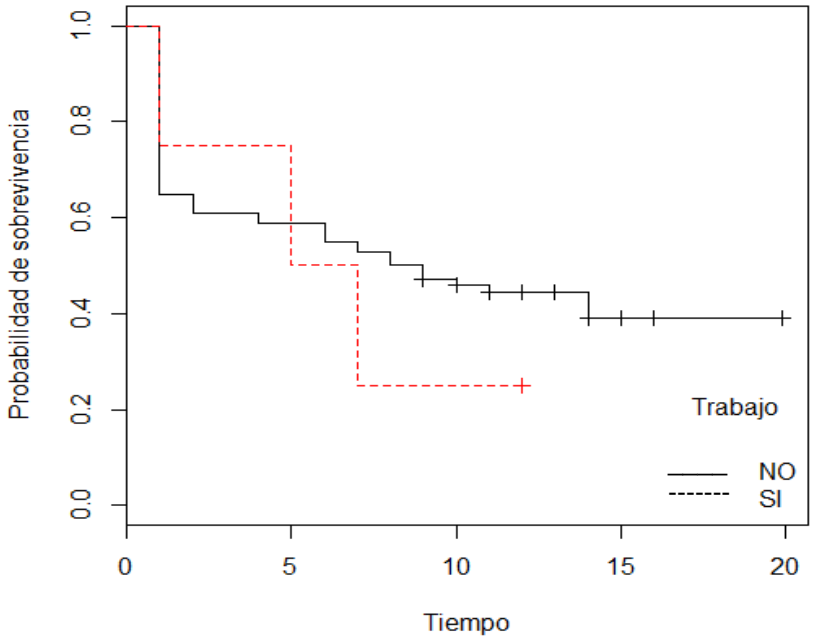
Sexo=F					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
9	18	4	0.778	0.098	0.608	0.996
10	13	2	0.658	0.114	0.469	0.923
11	10	1	0.592	0.120	0.398	0.881
12	8	3	0.370	0.126	0.190	0.722
Sexo=M					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
9	34	3	0.912	0.0486	0.821	1.000
10	29	4	0.786	0.0719	0.657	0.940
11	24	3	0.688	0.0823	0.544	0.870
12	14	3	0.540	0.0993	0.377	0.775

Al comparar las curvas de sobrevida de los hombres y las mujeres se obtiene un valor de Chisq = 1.3 con 1 grado de libertad y un p valor de 0.257, estadística que nos permite afirmar que no podemos rechazar la hipótesis nula de igualdad de las

curvas de sobrevida. Por lo tanto, podemos afirmar que la sobrevida (a la graduación) de hombres y mujeres tiene el mismo comportamiento.

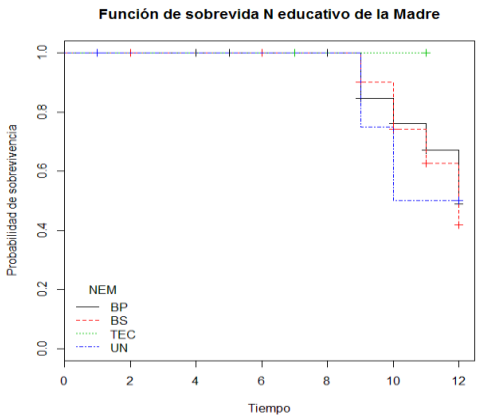
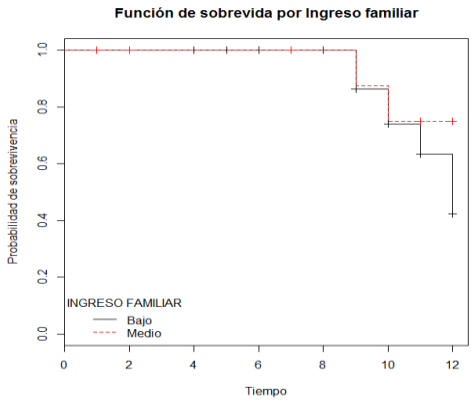
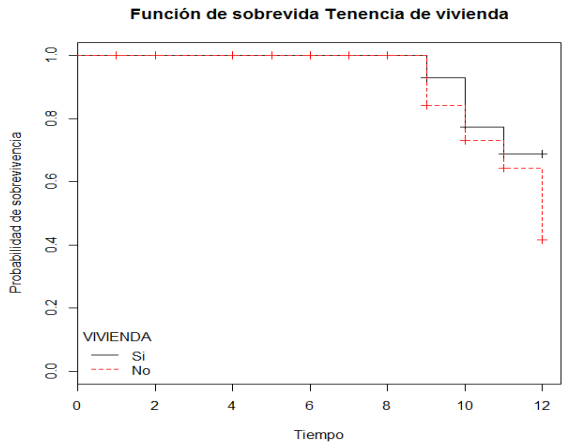
De igual manera se efectuó la prueba para la hipótesis nula de igualdad de las curvas de sobrevida respecto a si trabajaban o no a la hora de presentar el Icfes, encontrándose que, con Chisq= 0.8 y un p valor de 0.364, podemos afirmar que las curvas de sobrevida son distintas para la condición laboral. A continuación, se presenta la gráfica que sustenta la afirmación.

Gráfica 39: Función de sobrevida por ocupación al presentar el Icfes



Mediante las gráficas de las curvas de sobrevida presentadas en la tabla número 32 se determina que para las diferentes categorías de las variables la sobrevida (a la graduación) tiene el mismo comportamiento.

Tabla 31: Comparación de curvas de sobrevida por categorías.

Nivel Educativo de la madre	Ingreso Familiar
	
Estadístico de prueba	**Estadístico de prueba**
Chi-cuadrado= 1.1 con 3 grados de libertad, valor p= 0.786	Chi-cuadrado = 1.3 con 1 grados de libertad, valor p= 0.258
Tenencia de vivienda	
	
Estadístico de prueba	
Chi-cuadrado = 1.1 con 1 grados de libertad, valor p= 0.3	

5.3.2 Factores relacionados con el riesgo de graduación

Con el objetivo de medir los efectos de las variables consideradas en el estudio como explicativas de la graduación (edad de ingreso al Programa, nivel educativo de la madre, número de hermanos, tenencia o no de vivienda, ocupación a la hora de presentar el Icfes, ingreso familiar, puntaje Icfes estandarizado), a continuación, se presentan las estimaciones del modelo de riesgo proporcional utilizando el

modelo semiparamétrico de Cox. Los resultados del modelo inicial se presentan en la tabla 32

Tabla 32: Análisis del modelo semiparamétrico de Cox para LT

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
EdadIngreso	-1.994E-01	8.192E-01	1.511E-01	-1.320	0.187	
Ingresofami[T.Medio]	-1.855E+01	8.803-E09	1.060E+04	-0.002	0.999	
NEM[T.BS]	6.170E-01	1.853+00	5.666E-01	1.089	0.276	
NEM[T.TEC]	6.139E-01	1.848E+00	1.446E+04	0.000	1.000	
NEM[T.UN]	1.898EE+01	1.747E+08	1.060E+04	0.002	0.999	
NH	-1.816E-02	9.820E-01	1.699E-01	-0.017	0.915	
Puntaje	5.409E-03	1.005E+00	1.043E-02	0.519	0.604	**
Sexo[T.M]	-2.221E-01	8.009E-01	4.709E-01	1.805	0.637	
tenenciavivi[T.SI]	1.136E+00	3.113E+00	6.291E-01	1.805	0.071	
Trabajo[T.SI]	9.856E-01	2.679E+00	1.499E+04	0.000	1.000	

Seguidamente se procede a seleccionar el modelo más parsimonioso, utilizando el método hacia adelante y usando como criterio el AIC (Criterio de Información de Akaike), encontrándose que el modelo óptimo queda determinado por la edad de ingreso al programa y tener vivienda o no al presentar el Icfes. La tabla 33 resume el modelo óptimo:

Tabla 33: Resumen del modelo óptimo para LT

	Coef	exp(coef)	se(coef)	Z	Pr(>|z|)
EdadIngreso	-0.2489	0.7797	0.1286	-1.935	0.0529
Tenenciavivi	0.7409	2.0978	0.5577	1.329	0.1840
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	

	exp(coef)	exp(-coef)	Límite inferior 0.95	Límite superior 0.95
EdadIngreso	0.7797	1.2826	0.6060	1.003
Tenenciavivi	2.0978	0.4767	0.7032	6.258

En cuanto a las características observadas en los estudiantes se tiene que ninguna variable es significativa a la hora de explicar el evento de graduación.

Tabla 34: Análisis paramétrico del modelo óptimo

R - cuadrado= 0.048 (máximo posible= 0.783)			
Prueba de verosimilitud=	5.18	con 2 grados de libertad	Valor p=0.07486
Prueba de Wald =	4.81	con 2 grados de libertad	Valor p=0.09026
Prueba de puntaje de Log Rank =	4.5	con 2 grados de libertad	Valor p=0.1053

Se observa que el modelo es aceptable para cualquiera de los tres criterios y que las variables seleccionadas para el estudio tan sólo explican el 4.8% de la variabilidad en el tiempo de graduación

Tabla 35: Supuestos de riesgos proporcionales del modelo Cox

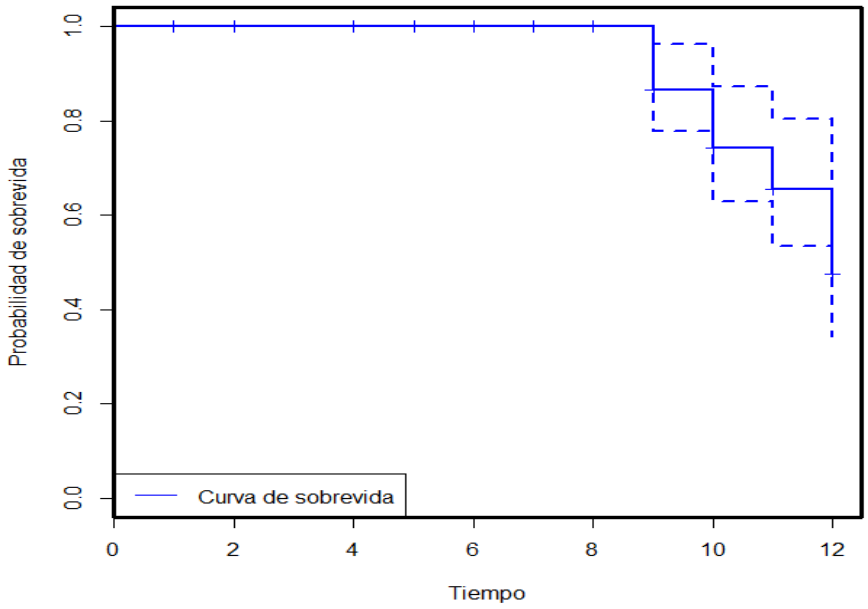
	Rho	Chi-cuadrado	Valor p
EdadIngreso	0.3441	2.1490	0.143
tenenciavivi[T.SI]	0.0649	0.0919	0.762
GLOBAL	NA	2.4460	0.294

Una vez estimado el modelo, se desea saber cuándo es más probable que ocurrirá la graduación. A continuación, se presenta la estimación “no paramétrica” de la función de sobrevida y la función de riesgo.

Tabla 36: Estimación no paramétrica de la función de sobrevida y de riesgo

					95% IC	
Tiempo	cardinalidad	Número de eventos	Valor de la función de sobrevida	Error estándar	Límite inferior	Límite superior
9	52	7	0.865	0.0473	0.777	0.963
10	42	6	0.742	0.0619	0.630	0.874
11	34	4	0.654	0.0683	0.533	0.803
12	22	6	0.476	0.0795	0.343	0.660

Gráfica 40: Función de sobrevida del modelo óptimo para LT



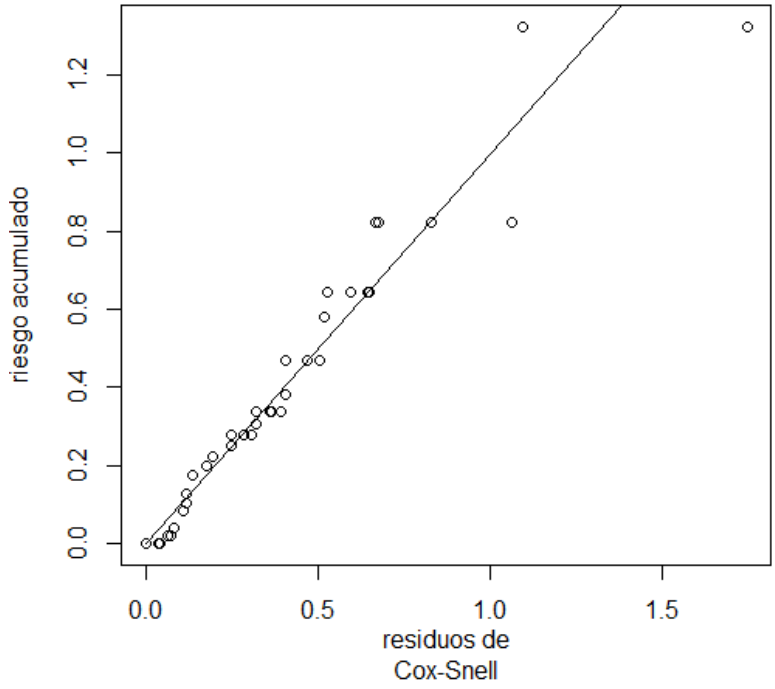
Se observa que la curva de sobrevida que incluye las variables consideradas en el modelo no presenta mayores variaciones con la estimada de Kaplan Meier presentada en la gráfica 37. Esto se debe a que la edad de ingreso al programa y tener o no vivienda al momento de presentar el Icfes (las variables que resultaron significativas) influyen en el tiempo hasta la graduación. Como se puede ver en la gráfica 40 los estudiantes se gradúan en el 9 semestre independiente de la edad de ingreso y los estudiantes que no tenían casa al momento de presentar el Icfes pueden durar mucho tiempo en el Programa.

5.3.3 Evaluación del modelo

Una vez se ha seleccionado el modelo más parsimonioso procedemos a evaluarlo. Las pruebas y los diagnósticos gráficos para evaluar el modelo de riesgos proporcionales se pueden basar en los residuales de Cox-Snell, los de martíngala, los de deviance, los de score y los de Schoenfeld. Residuales que se pueden utilizar para evaluar el ajuste global del modelo; la forma funcional apropiada de los predictores continuos; identificar los sujetos que están pobremente predichos por el modelo; identificar los puntos de influencia y verificar el supuesto de riesgo proporcional, (MONTOYA, L., Antonio. *Comparación de dos modelos de regresión en fiabilidad*. Granada, España. 2011, p. 31 Trabajo de gado de Master en Estadística Aplicada. De Granada de España. Facultad de Ciencias. Departamento de Estadística e Investigación Operativa)

Después de ajustar el modelo, tenemos que calcular los residuos de Cox-Snell con el fin de evaluar el ajuste del modelo de riesgos proporcionales. Si el modelo es correcto y la estimación de los $\beta's$ son cercanas a los valores reales, entonces los residuales deberían corresponder con una muestra censurada de observaciones de una distribución exponencial. Al aplicar el estimador de Nelson-Aalen de la tasa de riesgo acumulado de los residuos de Cox-Snell y gráficarlos versus los residuales de Cox-Snell, si una distribución exponencial ajusta a los datos, entonces, este estimador debería aproximadamente describir una línea de pendiente igual a 1. A partir de la gráfica 41 se puede afirmar que el modelo propuesto ajusta bien a los datos

Gráfica 41: Residuos de Cox Snell para LT



Comprobación de la hipótesis de riesgos proporcionales por cada variable explicativa

Ahora estamos interesados en evaluar la hipótesis de riesgos proporcionales para cada variable explicativa del modelo, es decir se observará si para cada variable explicativa, el riesgo de desertar puede variar con el tiempo, en el sentido de que el correspondiente coeficiente β puede no ser constante, es decir que $\beta(t)$ no depende del tiempo. La función cox.zph de R calcula la prueba de riesgos proporcionales para cada variable explicativa a partir de la correlación entre los residuales estandarizados de Schoenfeld (MONTOYA, L., Antonio. *Comparación de dos modelos de regresión en fiabilidad*. Granada, España. 2011, p. 31 Trabajo de gado de Master en Estadística Aplicada. De Granada de España. Facultad de Ciencias. Departamento de Estadística e Investigación Operativa)

Tabla 37: Prueba de riesgos proporcionales

	rho	Chi-cuadrado	Valor p
EdadIngreso	0.3441	2.1490	0.143
tenenciavivi[T.SI]	0.0649	0.0919	0.762
GLOBAL	NA	2.4460	0.294

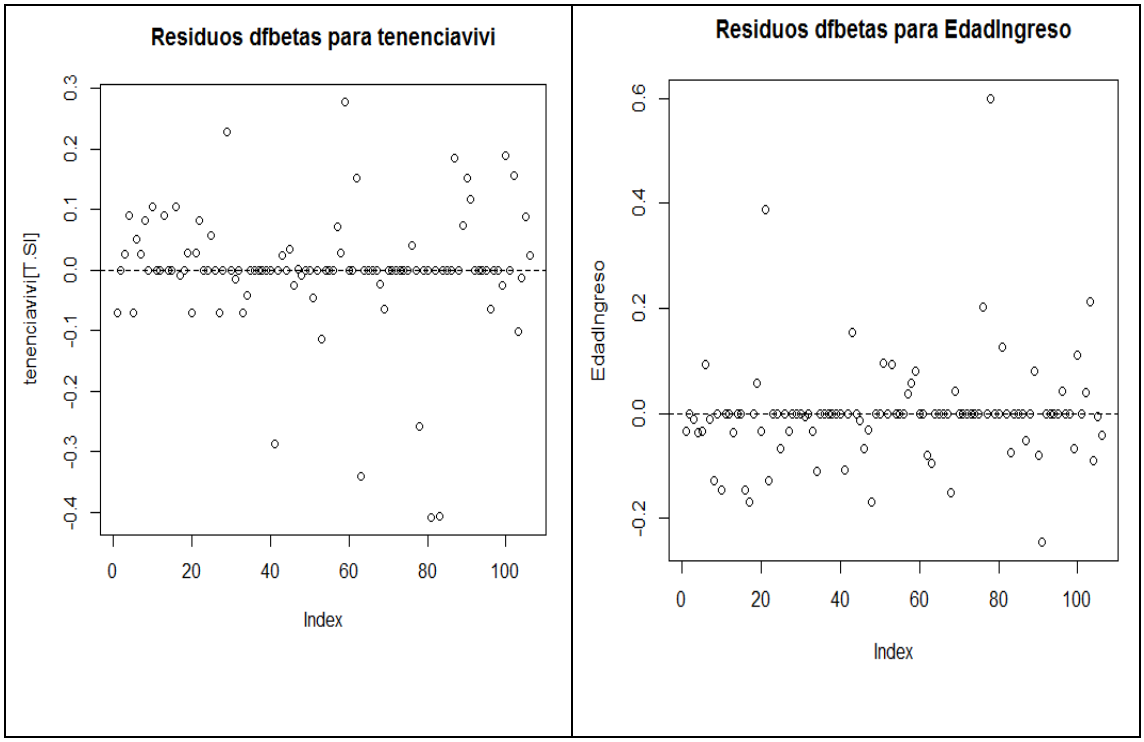
A partir de la información de la tabla anterior podemos decir que no existen evidencias significativas al 5% de que se viole el supuesto de riesgos proporcionales para tenenciavivi ni globalmente. Caso contrario sucede con la Edad de ingreso en donde se determina que el efecto de la edad de ingreso

depende del tiempo. Así las cosas, se debe tener cuidado a la hora de interpretar el efecto de esta variable en el modelo

Comprobación de la influencia sobre cada observación en el modelo

Otro uso de los residuos que se nos presenta es el de determinar la influencia de cada observación en el modelo ajustado. Hemos calculado, por medio de los residuos dfbeta, que están implementados en R, el cambio aproximado en el k-ésimo coeficiente (es decir, la k-ésima covariable) si la observación i-ésima se elimina del conjunto de datos y se vuelve a estimar el modelo sin esta observación. Para cada covariable, se ha representado la observación (en orden de tiempo de fallo registrado) por el cambio de escala aproximada (dividiendo por el error estándar del coeficiente) del coeficiente después de la eliminación de la observación del modelo. Si la supresión de una observación hace que el coeficiente incremente, el residuo dfbeta es negativo y viceversa.

Gráfica 42: Residuos dfbetas

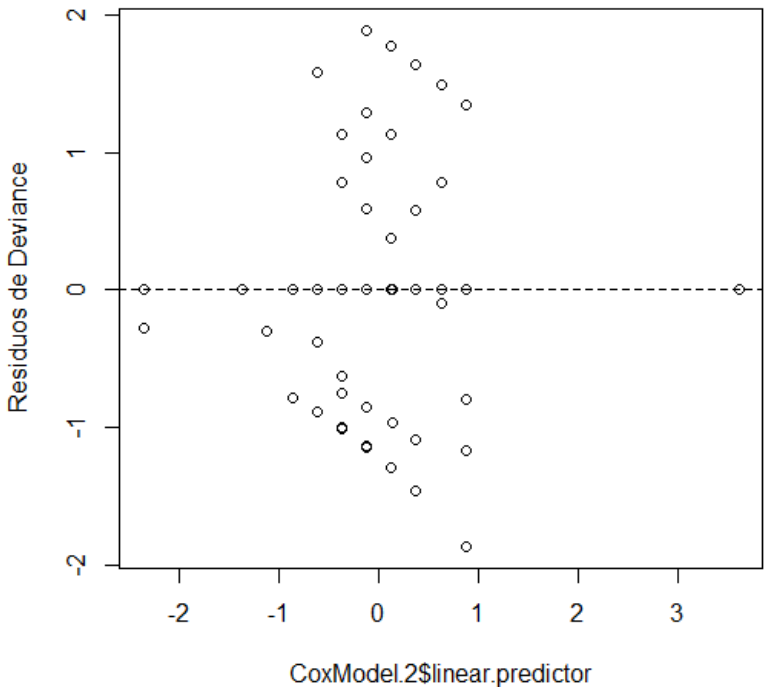


En la gráfica 42 se presentan los residuos *dfbeta* del modelo. Como vemos estos residuos están centrados con respecto al origen, y no presentan patrones definidos. En la variable Edad ingreso se presenta un dato alejado del origen, a excepción de esto no se aprecia ninguna irregularidad en las gráficas. Dado lo anterior se procede a efectuar la siguiente validación.

Comprobación de la existencia de outliers en el modelo

Se gráfican los residuales de deviance versus el predictor lineal, obteniéndose lo que se presenta en la gráfica 43. No se aprecian residuales alejados significativamente de cero. Se puede concluir que no hay presencia de datos atípicos.

Gráfica 43: Residuos de deviance



6 CONCLUSIONES Y RECOMENDACIONES

Como se muestra en este trabajo se puede apreciar que los factores que afectan la deserción (puntaje Icfes, Número de hermanos, edad de ingreso al programa y sexo) son exógenos, difícilmente o simplemente no pueden pueden ser cambiados por la Institución. Es decir, los riesgos se asocian más con cuestiones de tipo estructural de nuestra sociedad y que se refieren a la población estudiantil que tiene la Facultad Seccional Duitama. Así las cosas, es importante que la UPTC encamine sus esfuerzos a mecanismos alternativos de estudio que le permita a estudiantes con ciertas características reducir los riesgos de deserción.

Un estudiante modal del programa de Licenciatura en Matemáticas y Estadística es de género masculino, cuyas madres cuentan en su mayoría con básica primaria, con 3 hermanos en promedio, 22 años promedio de edad, de ingresos

familiares (al presentar el Icfes) bajos (0,1 o 2 SMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 67.58 sobre 100 puntos. Se encontró que el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el hecho de o no desertar del Programa. Las variables correlacionas con el tiempo hasta desertar son el número de hermanos y el puntaje en el Icfes, el valor de la correlación indicó que a medida que aumenta el número de hermanos el tiempo hasta desertar disminuye. De igual manera se tiene que a medida que aumenta el puntaje del Icfes también aumenta el tiempo hasta la deserción.

Para caso del programa de Licenciatura en Matemáticas y Estadística, se encontró que a medida que aumenta el número de semestres la probabilidad de sobrevivir a la deserción disminuye. Es más rápido el decrecimiento en los primeros cuatro semestres y tiende a estabilizarse a partir del 5 semestre donde está alrededor del 46%. El cuantil 0.5 a partir de la función de sobrevida, determino que hay una probabilidad del 50% de no presentar deserción hasta el tercer semestre y hay una probabilidad del 75% de no presentar deserción hasta el segundo semestre. Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 58%, es decir, que hay una probabilidad alta que deserte precozmente del Programa.

Al comparar las curvas de sobrevida por género, nivel educativo de la madre, ingreso familiar y tenencia de vivienda, se encontró que, para las diferentes categorías de las variables, la sobrevida a la deserción tiene el mismo comportamiento. La prueba para la hipótesis nula de igualdad de las curvas de sobrevida respecto a si trabajaban o no a la hora de presentar el Icfes, determinó que las curvas son distintas para la condición laboral. Se puede inferir que la probabilidad de permanecer en la Licenciatura en Matemáticas y Estadística es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable se deben tomar con precaución ya que no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad.

El modelo óptimo que explica el tiempo hasta que un estudiante deserta de la Licenciatura en Matemáticas y Estadística queda determinado por la edad de ingreso al Programa, el número de hermanos (NH), el puntaje estandarizado que obtuvo en el Icfes y el sexo del estudiante. El número de hermanos incide en el riesgo a desertar, el cual aumenta a medida que el estudiante incrementa su número de hermanos. Si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un hermano más, se multiplica por 1.269 la probabilidad de desertar. Los resultados también indican que el cambio proporcional en la función de riesgo que resulta de un aumento en un punto del puntaje estandarizado del Icfes con el que un estudiante ingresa al Programa, es negativo. Lo anterior significa que el puntaje del Icfes incide en el

riesgo a desertar, el cual disminuye a medida que el estudiante incrementa su puntuación. Si comparamos dos estudiantes, manteniendo las demás variables constantes, para aquel estudiante que tenga un punto menos en la prueba, se multiplica por 1.0152 la probabilidad de desertar.

Un estudiante modal del programa de la Licenciatura en Tecnología es de género masculino, cuyas madres cuentan en su mayoría con básica primaria, con 3 hermanos en promedio, 20 años promedio de edad, de ingresos familiares (al presentar el Icfes) bajos (0,1 o 2 SMLV), que al momento de presentar el Icfes no trabajaban, contaban con vivienda propia y obtuvieron un puntaje Icfes promedio de 59.65 sobre 100 puntos. Se encontró que el género del estudiante, nivel educativo de la madre, ingreso familiar, contar con vivienda propia o haber trabajado al momento de presentar el Icfes no tienen relación con el hecho de o no desertar del Programa.

Para caso de la Licenciatura en Tecnología, se encontró que a medida que aumenta el número de semestres la probabilidad de sobrevivir a la deserción disminuye. Es más rápido el decrecimiento en los primeros seis semestres y tiende a estabilizarse a partir del 7 semestre donde está alrededor del 53%. El cuantil 0.5 a partir de la función de sobrevida, determino que hay una probabilidad del 50% de no presentar deserción hasta el noveno semestre y hay una probabilidad del 75% de no presentar deserción hasta el sexto semestre. Con base en la función de riesgo acumulada se puede afirmar que la probabilidad que un estudiante deserte justo al terminar el primer semestre es de 40%, es decir que hay una probabilidad alta que deserte precozmente del programa.

Al comparar las curvas de sobrevida para el sexo, nivel educativo de la madre, ingreso familiar y tenencia de vivienda, se encontró que, para las diferentes categorías de las variables, la sobrevida a la graduación tiene el mismo comportamiento. La prueba para la hipótesis nula de igualdad de las curvas de sobrevida respecto a si trabajaban o no a la hora de graduarse el Icfes, determinó que las curvas son distintas para la condición laboral. Se puede inferir que la probabilidad de graduarse de la Licenciatura en Tecnología es superior en quienes no trabajan que en los que si lo hacen. Sin embargo, los resultados de esta variable se deben tomar con precaución ya que no implica necesariamente el estado actual del estudiante ya que su recolección corresponde a un periodo anterior al ingreso a la universidad.

El modelo óptimo que explica el tiempo hasta que un estudiante se gradúa de la Licenciatura en Tecnología queda determinado por la edad de ingreso al Programa y la Tenencia de vivienda al momento de presentar el Icfes.

Construir a futuro nuevos modelos de sobrevida a partir de la recolección de información a través del instrumento, el cual está disponible en los Anexos C y D,

y que tiene como propósito recoger información sobre aspectos importantes que no fueron abordados en este proyecto.

En este estudio se usaron modelos no paramétricos, sería conveniente en los trabajos posteriores asumir una distribución de probabilidad para las variables tiempo hasta la deserción y graduación

7 BIBLIOGRAFÍA

CERON, Deisy. Modelamiento estadístico en el rendimiento académico de los estudiantes los programas de administración de la Universidad Pedagógica y Tecnológica de Colombia – Duitama, segundo semestre de 2012., Licenciada en Matemáticas y Estadística. Duitama: Universidad Pedagógica y Tecnológica de Colombia. Facultad Seccional Duitama, Departamento de Educación. 2012. P.35

BOJ, Eva del Val. Evaluación de la hipotesis de riesgos Proporcionales. En: *El modelo de regresion de Cox..* 1 Ed., 2015. P. *34*

COLOSIMO, Enricon A. y GIOLO Ruíz, Suely. Conceptos básicos y ejemplos, técnicas no paremétricas y modelos probabilísticos. En: *Análisis de sobrevida aplicada.* 1 Ed.1997. P. *6-123.*

DÍAZ, Guillermo., (agosto 2013). *Análisis estadístico de datos "tiempo para un evento" univariados y multivariados*, II Encuentro Internacional de Matemáticas, Estadística y Educación Matemática, Tunja, Colombia.

DOMÉNECH., M., Joseph. *Una aplicación del análisis de la supervivencia de la salud.* En: Anuarios de Psicología. 55taEdición 1992. Universidad de Barcelona. Pág. 109-141

GODOY, Ángel. *Introducción al análisis de supervivencia con R*. Distrito Federal de México. 2009, p. 10 -28. Trabajo de gado de Actuario. Universidad Nacional Autónoma de México. Facultad de Ciencias. Departamento de Matemáticas y Estadística

GUZMÁN, Carolina., DURÁN Diana., FRANCO, Jorge., CASTAÑO, Elkin., GALLÓN, Santiago., GÓMEZ, Karoll. y VÁSQUEZ, Johanna. Una introducción a los modelos de duración para el análisis de la deserción estudiantil. En: *Deserción estudiantil en la educación superior colombiana*. 1 ed. Bogotá: Ministerio de educación nacional, 2009. p. 17, 40-43.

RICO HIGUITA, Alberto D. *Caracterización de la deserción estudiantil en la Universidad Nacional de Colombia sede Medellín*. Medellín: Colombia. Universidad Nacional de Colombia sede Medellín. 2006. p. 8-13.

Ministerio de Educacion Nacional de Colombia. (30 de Octubre de 2007). *Ministerio de Educacion Nacional, Republica de Colombia.* Obtenido de http://www.mineducacion.gov.co/1621/article-133057.html)

MONTOYA, L., Antonio. *Comparación de dos modelos de regresión en fiabilidad*. Granada, España. 2011, p. 30-39 Trabajo de gado de Master en Estadística

Aplicada. De Granada de España. Facultad de Ciencias. Departamento de Estadística e Investigación Operativa

OSORIO, Ana María. BOLANCE, Catalina., y CAICEDO CASTILLO Maribel. Deserción y graduación estudiantil Universitaria: Una aplicación de los modelos de supervivencia. En: Revista Iberoamericana de Educación Superior. Julio, 2012. Vol. 3 No 6, p. 41-43.

REBELLON BARRERA, Mauricio. *Análisis de supervivencia aplicado al problema de la deserción estudiantil en la Universidad Tecnológica de Pereira*, Magister en Investigación de Operativa y Estadística. Pereira: Universidad Tecnológica de Pereira. Facultad Ingeniera Industrial. 2008. p. 3-25

R Development Core Team (2009). R: *A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org*.

TORRES C., Yeferson J., SIAUCHÓ L., Miguel A. (2013). *Estudio de permanencia y deserción estudiantil en el programa de Licenciatura en Matemáticas y Estadística*. Trabajo de Grado. Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

8 ANEXOS

ANEXO A: MUESTRA DE CADA PROGRAMA

Códigos Licenciatura en Matemáticas y Estadística			
44220	59293	200720915	200910621
44235	64043	200721006	200910007
44238	64045	200720059	200910423
44231	64048	200720958	200910887
44233	64049	200720620	200910922
44237	64052	200720951	200920239
44219	64053	200720089	200921265
44215	64055	200720046	200921468
44222	64057	200720090	200921490
44226	64058	200810594	200921283
49042	64059	200810567	200921295
49043	64064	200810585	200921407
49045	69028	200810049	200920990
49046	69029	200810644	200921460
49050	69032	200810615	200920780
49053	69033	200810601	200921301
49246	69034	200810577	200921480
49057	69035	200820331	200921326
49058	69255	200820605	200922221
49044	69037	200820005	200921458
49245	69038	200820406	200921355
54043	69041	200820379	200921550
54039	200712081	200820358	200910981
54299	200711233	200820045	200910696
54040	200711486	200820007	59271
54048	200711595	200910786	
54046	200711375	200910785	
54266	200711576	200910634	
54258	200711641	200910816	
54298	200711434	200910287	
54049	200711343	200910690	
54045	200711325	200910409	
59265	200711623	200911039	
59263	200720948	200720529	

Códigos Licenciatura en Tecnología			
44151	64254	200720626	200910978
44162	64257	200720947	200910610
44155	64258	200720885	200910314
44158	64259	200720745	200910776
44163	64261	200720117	200921271
44166	64262	200720900	200921408
44170	64263	200720077	200921394
44173	64264	200810574	200921753
49111	64278	200810520	200920176
49114	64265	200810624	200921307
49118	64266	200810659	200921652
49119	64268	200810570	200921554
49121	64269	200810411	200921358
49106	64271	200810584	200921529
49108	69042	200810633	200921466
49110	69046	200810031	200920149
49113	69048	200811366	200921354
49116	69051	200810038	200921578
49115	69052	200810576	200921346
49104	69053	200820148	200921422
49107	69055	200820660	200921316
54056	69056	200820184	200921031
54057	69060	200820512	59246
54058	200711702	200820177	200720517
54301	200711499	200820403	200910975
54061	200711583	200820637	
54303	200711289	200910911	
54063	200711137	200911651	
54064	200720216	200911750	
54306	200720886	200910643	
59236	200720067	200910789	
59239	200720094	200910485	
59242	200720322	200910271	
59245	200720223	200910535	

ANEXO B: SENTENCIAS DEL R

********** Sintaxis en "R"**********

MODELAMIENTO#

Sentencias del R para el modelamiento del programa de Licenciatura en Matemáticas y Estadística para la deserción.

#DIAGRAM DE LEXIS#

>LL <- Lexis.diagram(age=c(-1,15), date=c(2003,2017), entry.age= EdadIng, exit.age= EdadSal, birth.date=cohorte, fail=(censura), lwd.life=1, cex.fail=0.8, col.fail=c("blue","red"), alab="Tiempo de deserción en años", dlab="Cohorte", int=c(1,1),data=sobremateDes) abline(v=c(2004,2015), h=c(0.5,2.5), col="green",lwd=2) # Identify the persons' entry and exits text(LL$exit.date, LL$exit.age, paste(1:nrow(LL)), col="Black", font=10, adj=c(0,1))

#CONSTRUCCIÓN DE CURVA DE SOBREVIDA#

>Survfit <- survfit(Surv(Tdeserción,censura) ~ 1, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateDes) summary(.Survfit) plot(.Survfit, mark.time=TRUE, main="Función de sobrevida", sub="Licenciatura en Matemáticas y Estadística", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue") box(lwd=3, col = "black") legend("bottomleft",c("Curva de sobrevida"),lty=1, col="blue")

#ESTIMACIÓN DE CUANTILES PARA LA FUNCIÓN DE SOBREVIDA#

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

#COMPARACIÓN POR SEXO DE LA FUNCIÓN DE SOBREVIDA#

>Survfit <- survfit(Surv(Tdeserción,censura) ~ Sexo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateDes) summary(.Survfit) plot(.Survfit, col=1:2, lty=1:2, mark.time=TRUE,main="Función de sobrevida por sexo", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomleft", legend=c("F","M"), title="Sexo", col=1:2, lty=1:2, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ Sexo, rho=0, data=sobremateDes)

#COMPARACIÓN POR NEM DE LA FUNCIÓN DE SOBREVIDA#

>Survfit <- survfit(Surv(Tdeserción,censura) ~ NEM, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateDes)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ NEM, rho=0, data=sobremateDes)

#COMPARACIÓN POR OCUPACIÓN AL PRESENTAR EL ICFES DE LA FUNCIÓN DE SOBREVIDA#

>Survfit <- survfit(Surv(Tdeserción,censura) ~ Trabajo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateDes)

>summary(.Survfit)

>plot(.Survfit, col=1:3, lty=1:3, conf.int=FALSE, mark.time=TRUE, main="Función de sobrevida ocupación al presentar Icfes", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomright", legend=c("","NO","SI"), title="Trabajo", col=1:3, lty=1:3, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ Trabajo, rho=0, data=sobremateDes)

#CONSTRUCCIÓN DE LA FUNCIÓN DE RIESGO ACUMULADA#

>Survfit <- survfit(Surv(Tdeserción,censura) ~ 1, conf.type="log", conf.int=F, type="kaplan-meier", error="greenwood", data=sobremateDes,fun="cumhaz")

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de riesgo acumulada", sub="Licenciatura en Matemáticas y Estadística", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue", fun="cumhaz") box(lwd=3, col = "black") legend("bottomright",c("Función de riesgo acumulada"),lty=1, col="blue") abline(v=c(1,5),col="green")

#CONSTRUCCIÓN MODELO COMPLETO#

>CoxModel.1 <- coxph(Surv(Tdeserción,censura) ~ EdadIngreso + Ingresofami + NEM + NH + Puntaje + Sexo + tenenciavivi + Trabajo, method="efron", data=sobremateDes)

>summary(CoxModel.1)

#Para efectuar los cálculos se omitieron de la base los datos faltantes#

#SELECCIÓN MODELO ÓPTIMO#

>stepwise(CoxModel.1, direction='backward/forward', criterion='AIC')

#CONSTRUCCIÓN MODELO ÓPTIMO

>CoxModel.2 <- coxph(Surv(Tdeserción,censura) ~ EdadIngreso + NH + Puntaje + Sexo, method="efron", data=sobremateDes)

>summary(CoxModel.2)

>cox.zph(CoxModel.2)

Residuos de Cox-Snell

>estado<-sobremateDes$censura

>mresi<-residuals(CoxModel.2, type="martingale")

>csresi<-estado-mresi

>hazard.csresi<-survfit(Surv(csresi,estado)~1,type="fleming-harrington")

>plot(hazard.csresi$time,-log(hazard.csresi$surv), xlab='residuos de Cox-Snell', ylab='riesgo acumulado',lty = 1:4, main="Representación de los residuos de Cox-Snell") lines(c(0,5),c(0,5))

Residuos dfbeta

>dfbeta <- residuals(CoxModel.2, type="dfbetas") par(mfrow=c(1,1)) for (j in 1:1){ plot(dfbeta[,j], main="Residuos dfbetas para EdadIngreso",ylab= names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))}

>dfbeta <- residuals(CoxModel.2, type="dfbetas"), par(mfrow=c(1,1)) for (j in 1:2){ plot(dfbeta[,j], main="Residuos dfbetas para NH",

ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))

>dfbeta <- residuals(CoxModel.2, type="dfbetas") par(mfrow=c(1,1)) for (j in 1:3){ plot(dfbeta[,j], main="Residuos dfbetas para Puntaje",ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))}

>dfbeta <- residuals(CoxModel.2, type="dfbetas") par(mfrow=c(1,1)) for (j in 1:4){ plot(dfbeta[,j], main="Residuos dfbetas para Sexo",ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))}

Residuos de deviance

>devresi <- resid(CoxModel.2, type="deviance")

>plot(CoxModel.2$linear.predictor, devresi, ylab="Residuos de Deviance", main='Residuos de deviance') abline(h=0,lty=2, col='black') dfbeta <- residuals(CoxModel.2, type="dfbetas")

>par(mfrow=c(1,1)) for (j in 1:4) {plot(dfbeta[,j], main="Residuos dfbetas para ",ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))}

Residuos Martingale

>NullModel <- update(CoxModel.2, ~ 1) .residuals <- residuals(.NullModel, type="martingale") .X <- padNA(model.matrix(CoxModel.2), residuals(CoxModel.2)) .mfrow <- par(mfrow = mfrow(1)) scatter.smooth(.X["EdadIngreso"], .residuals, xlab="EdadIngreso", ylab="Martingale residuals from null model", family="gaussian") abline(lm(.residuals ~ .X["EdadIngreso"]), lty=2)

>plot(.X["NH"], .residuals, xlab="NH", ylab="Martingale residuals from null model") abline(lm(.residuals ~ .X["NH"]), lty=2) scatter.smooth(.X["Puntaje"], .residuals, xlab="Puntaje", ylab="Martingale residuals from null model", family="gaussian") abline(lm(.residuals ~ .X["Puntaje"]), lty=2)

>plot(.X["Sexo[T.M]"], .residuals, xlab="Sexo[T.M]", ylab="Martingale residuals from null model") abline(lm(.residuals ~ .X["Sexo[T.M]"]), lty=2)

>par(mfrow=.mfrow)

>remove(.NullModel, .residuals, .X, .mfrow)

Sentencias del R para el modelamiento del programa de Licenciatura en Matemáticas y Estadística para la graduación.

#DIAGRAMA DE LEXIS#

>LL <- Lexis.diagram(age=c(-1,15), date=c(2003,2017), entry.age= EdadIng, exit.age= EdadSal, birth.date=cohorte, fail=(censura), lwd.life=1, cex.fail=0.8, col.fail=c("blue","red"), alab="Tiempo de graduación en años", dlab="Cohorte", int=c(1,1),data=sobremateGr) abline(v=c(2004,2015), h=c(0.5,2.5), col="green",lwd=2) # Identify the persons' entry and exits text(LL$exit.date, LL$exit.age, paste(1:nrow(LL)), col="Black", font=10, adj=c(0,1))

#CONSTRUCCIÓN DE CURVA DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ 1, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateGr)

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de sobrevida", sub="Licenciatura en Matemáticas y Estadística", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue") box(lwd=3, col = "black") legend("bottomleft",c("Curva de sobrevida"),lty=1, col="blue")

#ESTIMACIÓN DE CUANTILES PARA LA FUNCIÓN DE SOBREVIDA#

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

#COMPARACIÓN POR SEXO DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ Sexo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateGr)

>summary(.Survfit)

>plot(.Survfit, col=1:2, lty=1:2, mark.time=TRUE,main="Función de sobrevida por sexo", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomleft", legend=c("F","M"), title="Sexo", col=1:2, lty=1:2, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ Sexo, rho=0, data=sobremateGr)

#COMPARACIÓN POR NEM DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ NEM, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateGr)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

<survdiff(Surv(Tgraduación,censura) ~ NEM, rho=0, data=sobremateGr)

#COMPARACIÓN POR OCUPACIÓN AL PRESENTAR EL ICFES DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ Trabajo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobremateGr)

>summary(.Survfit)

>plot(.Survfit, col=1:3, lty=1:3, conf.int=FALSE, mark.time=TRUE, main="Función de sobrevida ocupación al presentar Icfes", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomright", legend=c("","NO","SI"), title="Trabajo", col=1:3, lty=1:3, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ Trabajo, rho=0, data=sobremateGr)

#CONSTRUCCIÓN DE LA FUNCIÓN DE RIESGO ACUMULADA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ 1, conf.type="log" conf.int=F, type="kaplan-meier", error="greenwood", data=sobremateGr,fun="cumhaz")

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de riesgo acumulada", sub="Licenciatura en Matemáticas y Estadística", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue", fun="cumhaz") box(lwd=3, col = "black") legend("bottomright",c("Función de riesgo acumulada"),lty=1, col="blue") abline(v=c(1,5),col="green")

#CONSTRUCCIÓN MODELO COMPLETO#

>CoxModel.1 <- coxph(Surv(Tgraduación,censura) ~ EdadIngreso + Ingresofami + NEM + NH + Puntaje. + Sexo + tenenciavivi + Trabajo, method="efron", data=sobremateGr)

>summary(CoxModel.1)

#Para efectuar los cálculos se omitieron de la base los datos faltantes#

#SELECCIÓN MODELO ÓPTIMO#

>stepwise(CoxModel.1, direction='backward/forward', criterion='AIC')

Sentencias del R para el modelamiento del programa de Licenciatura en Tecnología para la deserción.

#DIAGRAMA DE LEXIS#

>LL <- Lexis.diagram(age=c(-1,15), date=c(2003,2017),entry.age= EdadIng, exit.age= Edadsal, birth.date=cohorte, fail=(censura), lwd.life=1, cex.fail=0.8, col.fail=c("blue","red"), alab="Tiempo de deserción en años", dlab="Cohorte", int=c(1,1),data=sobreindusDes) abline(v=c(2004,2015), h=c(0.5,2.5), col="green",lwd=2) # Identify the persons' entry and exits text(LL$exit.date, LL$exit.age, paste(1:nrow(LL)), col="Black", font=10, adj=c(0,1))

#CONSTRUCCIÓN DE CURVA DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tdeserción,censura) ~ 1, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusDes)

summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de sobrevida", sub=" Licenciatura en Tecnología ", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue") box(lwd=3, col = "black") legend("bottomleft",c("Curva de sobrevida"),lty=1, col="blue")

#ESTIMACIÓN DE CUANTILES PARA LA FUNCIÓN DE SOBREVIDA#

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

#COMPARACIÓN POR SEXO DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tdeserción,censura) ~ Sexo, conf.type="log" conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusDes)

>summary(.Survfit)

>plot(.Survfit, col=1:2, lty=1:2, mark.time=TRUE,main="Función de sobrevida por sexo", xlab="Tiempo", ylab="Probabilidad de sobrevida"), legend("bottomleft", legend=c("F","M"), title="Sexo", col=1:2, lty=1:2, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ Sexo, rho=0, data=sobreindusDes)

#COMPARACIÓN POR NEM DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tdeserción,censura) ~ NEM, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusDes)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ NEM, rho=0, data=sobreindusDes)

#COMPARACIÓN POR OCUPACIÓN AL PRESENTAR EL ICFES DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tdeserción,censura) ~ Trabajo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusDes)

>summary(.Survfit)

>plot(.Survfit, col=1:3, lty=1:3, conf.int=FALSE, mark.time=TRUE, main="Función de sobrevida ocupación al presentar Icfes", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomright", legend=c("","NO","SI"), title="Trabajo", col=1:3, lty=1:3, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tdeserción,censura) ~ Trabajo, rho=0, data=sobreindusDes)

#CONSTRUCCIÓN DE LA FUNCIÓN DE RIESGO ACUMULADA#

>.Survfit <- survfit(Surv(Tdeserción,censura) ~ 1, conf.type="log", conf.int=F, type="kaplan-meier", error="greenwood", data=sobreindusDes,fun="cumhaz")

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de riesgo acumulada", sub=" Licenciatura en Tecnología ", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue", fun="cumhaz") box(lwd=3, col = "black") legend("bottomright",c("Función de riesgo acumulada"),lty=1, col="blue") abline(v=c(1,5),col="green")

#CONSTRUCCIÓN MODELO COMPLETO#

>CoxModel.1 <- coxph(Surv(Tdeserción,censura) ~ EdadIngreso + Ingresofami + NEM + NH + Puntaje. + Sexo + tenenciavivi + Trabajo, method="efron", data=sobreindusDes)

>summary(CoxModel.1)

#Para efectuar los cálculos se omitieron de la base los datos faltantes#

#SELECCIÓN MODELO ÓPTIMO#

>stepwise (CoxModel.1, direction='backward/forward', criterion='AIC')

Sentencias del R para el modelamiento del programa de Licenciatura en Tecnología para la graduación.

#DIAGRAMA DE LEXIS#

>LL <- Lexis.diagram(age=c(-1,15), date=c(2003,2017), entry.age= EdadIng, exit.age= Edadsal, birth.date=cohorte, fail=(censura), lwd.life=1, ex.fail=0.8, col.fail=c("blue","red"), alab="Tiempo de graduación en años", dlab="Cohorte", int=c(1,1),data=sobreindusGr) abline(v=c(2004,2015), h=c(0.5,2.5), col="green",lwd=2) # Identify the persons' entry and exits text(LL$exit.date, LL$exit.age, paste(1:nrow(LL)), col="Black", font=10, adj=c(0,1))

#CONSTRUCCIÓN DE CURVA DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ 1, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusGr)

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de sobrevida", sub=" Licenciatura en Tecnología ", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue") box(lwd=3, col = "black") legend("bottomleft",c("Curva de sobrevida"),lty=1, col="blue")

#ESTIMACIÓN DE CUANTILES PARA LA FUNCIÓN DE SOBREVIDA#

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

#COMPARACIÓN POR SEXO DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ Sexo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusGr)

>summary(.Survfit)

>plot(.Survfit, col=1:2, lty=1:2, mark.time=TRUE,main="Función de sobrevida por sexo", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomleft", legend=c("F","M"), title="Sexo", col=1:2, lty=1:2, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ Sexo, rho=0, data=sobreindusGr)

#COMPARACIÓN POR NEM DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ NEM, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusGr)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>plot(.Survfit, col=1:4, lty=1:4, mark.time=TRUE ,main="Función de sobrevida N educativo de la Madre", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomleft", legend=c("BP","BS","TEC","UN"), title="NEM", col=1:4, lty=1:4, bty="n")

>quantile(.Survfit, probs=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ NEM, rho=0, data=sobreindusGr)

#COMPARACIÓN POR INGRESO FAMILIAR DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~Ingresofami , conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusGr)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>plot(.Survfit, col=1:4, lty=1:4, mark.time=TRUE ,main="Función de sobrevida por Ingreso familiar", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomleft", legend=c("Bajo","Medio"), title="INGRESO FAMILIAR", col=1:4, lty=1:4, bty="n")

>quantile(.Survfit, probs=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ Ingresofami, rho=0, data=sobreindusGr)

#COMPARACIÓN POR TENECIA DE VIVIENDA DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~tenenciavivi , conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusGr)

>summary(.Survfit)

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>plot(.Survfit, col=1:4, lty=1:4, mark.time=TRUE ,main="Función de sobrevida Tenencia de vivienda", xlab="Tiempo", ylab="Probabilidad de sobrevida"), legend("bottomleft", legend=c("Si","No"), title="VIVIENDA", col=1:4, lty=1:4, bty="n")

>quantile(.Survfit, probs=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ tenenciavivi, rho=0, data=sobreindusGr)

#COMPARACIÓN POR OCUPACIÓN AL PRESENTAR EL ICFES DE LA FUNCIÓN DE SOBREVIDA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ Trabajo, conf.type="log", conf.int=0.95, type="kaplan-meier", error="greenwood", data=sobreindusDes)

>summary(.Survfit)

>plot(.Survfit, col=1:3, lty=1:3, conf.int=FALSE, mark.time=TRUE, main="Función de sobrevida ocupación al presentar Icfes", xlab="Tiempo", ylab="Probabilidad de sobrevida") legend("bottomright", legend=c("","NO","SI"), title="Trabajo", col=1:3, lty=1:3, bty="n")

>quantile(.Survfit, quantiles=c(.25,.5,.75))

>remove(.Survfit)

>survdiff(Surv(Tgraduación,censura) ~ Trabajo, rho=0, data=sobreindusGr)

#CONSTRUCCIÓN DE LA FUNCIÓN DE RIESGO ACUMULADA#

>.Survfit <- survfit(Surv(Tgraduación,censura) ~ 1, conf.type="log", conf.int=F, type="kaplan-meier", error="greenwood", data=sobreindusDes,fun="cumhaz")

>summary(.Survfit)

>plot(.Survfit, mark.time=TRUE, main="Función de riesgo acumulada", sub="Licenciatura en Tecnología ", xlab="Tiempo", ylab="Probabilidad de sobrevida", lwd=2, col="blue", fun="cumhaz") box(lwd=3, col = "black") legend("bottomright",c("Función de riesgo acumulada"),lty=1, col="blue") abline(v=c(1,5),col="green")

#CONSTRUCCIÓN MODELO COMPLETO#

>CoxModel.1 <- coxph(Surv(Tgraduación,censura) ~ EdadIngreso + Ingresofami + NEM + NH + Puntaje. + Sexo + tenenciavivi + Trabajo, method="efron", data=sobreindusGr)

>summary(CoxModel.1)

#Para efectuar los cálculos se omitieron de la base los datos faltantes#

#SELECCIÓN MODELO ÓPTIMO#

>stepwise(CoxModel.1, direction='backward/forward', criterion='AIC')

#CONSTRUCCIÓN MODELO ÓPTIMO#

>CoxModel.2 <- coxph(Surv(Tgraduación,censura) ~ EdadIngreso+ tenenciavivi, method="efron", data=sobreindusGr)

>summary(CoxModel.2)

>cox.zph(CoxModel.2)

Residuos de Cox-Snell

>estado<-sobreindusGr$censura

>mresi<-residuals(CoxModel.2, type="martingale")

>csresi<-estado-mresi

>hazard.csresi<-survfit(Surv(csresi,estado)~1,type="fleming-harrington")

>plot(hazard.csresi$time,-log(hazard.csresi$surv), xlab='residuos de Cox-Snell', ylab='riesgo acumulado',lty = 1:4, main="Representación de los residuos de Cox-Snell") lines(c(0,5),c(0,5))

Residuos dfbeta

>dfbeta <- residuals(CoxModel.2, type="dfbetas") par(mfrow=c(1,1)) for (j in 1:1){ plot(dfbeta[,j], main="Residuos dfbetas para EdadIngreso",ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black' lines(c(0,0),c(0,0)) }

>dfbeta <- residuals(CoxModel.2, type="dfbetas")

>par(mfrow=c(1,1)) for (j in 1:2){ plot(dfbeta[,j], main="Residuos dfbetas para tenenciavivi",ylab=names(coef(CoxModel.2))[j]) abline(h=0, lty=2, col='black') lines(c(0,0),c(0,0))}

Residuos de deviance

>devresi <- resid(CoxModel.2, type="deviance")

>plot(CoxModel.2$linear.predictor, devresi, ylab="Residuos de Deviance", main='Residuos de deviance') abline(h=0,lty=2, col='black')

Residuos Martingale

>mres<-residuals(CoxModel.2, type=c("martingale")) plot(sobreindusGr[,1], mres, xlab=c("tenecniavivi")[1], ylab="Residuos martingale", main="Residuos de Martingala") abline(h=0, lty=2) lines(lowess(sobreindusGr[,1], mres, iter=0))

ANEXO C: FORMATO DE REGISTRO DE DATOS

Teniendo en cuenta que el análisis efectuado en este proyecto indicó un pobre ajuste, es importante indicar algunas otras variables que podrían estar afectando el riesgo de deserción y la graduación. Lo anterior con el fin de que cada programa recoja la información pertinente a sus estudiantes y a futuro la utilice para construir modelos de sobrevida con una mayor bondad de ajuste.

Las variables que se proponen se dividen en 4 grupos, así:

I. Información Personal

- Semestre de ingreso: año del semestre en el que un estudiante se matriculó al programa, ejemplo: 2016-I, significa primer semestre del 2016.
- Código: Valor numérico asignado por registro en el momento de la matrícula.
- *Género,* los valores que toma son: Masculino (M), Femenino (F).
- *Edad:* valor numérico de la edad en la que se matriculó en el programa.
- Estado Civil: Soltero(a), Casado(a), Divorciado(a), Viudo(a)
- Si tiene hijos, indicar el número de estos.
- Si tiene hermanos, señalar el número de hermanos
- Posición entre hermanos: No tiene (0), primer lugar (1), segundo (2), así sucesivamente.
- Estado de la salud Física: Bueno, Regular, Malo.
- *Población Vulnerable*: Desplazado (DESPL), Afrodescendiente (AFRO), LGTBI, Ninguno.
- *Embarazos no planeados*: si ha experimentado tales, Si o No.
- *Calamidad y/o problema doméstico* los valores que toma: Si presentó algunos de los siguientes eventos, Muerte de los padres (MP), Muerte de un Hijo (MH), Ninguno.
- *Expectativas no satisfechas*: Percepciones de estudiante antes de ingresar al programa, posibles valores que puede tomar: no le gusta el programa(NGPR), no era como se lo esperaba(NESP), es muy duro el cambio del colegio a la Universidad(CAMB_U)
- *Discapacidad:* si el estudiante manifiesta tenerla, posibles valores a tomar: Física, Sensorial: dentro de esta se encuentran (auditiva, visual), psíquica, intelectual o mental.

II. Información Socioeconómica

- Situación Laboral del estudiante al momento de ingresar a la universidad: valores que toma la variable: Si, No.
- ¿Cuánto es el nivel de ingresos?: Bajo (0, 1 SMMLV], Medio (1,3 SMMLV], Alto (3, en adelante], SMMLV: Salario Mínimo Mensual Legal Vigente.
- ¿Con quién vive? Madre y padre (1), Madre o padre (2), Otro familiar (3), ningún familiar (4)
- Si se tienen personas a cargo, indicar el número de estas
- Nivel del Sisben, indica el nivel de este, valores que toma (1,2,3)
- Clasificación del Estrato socioeconómico los valores que toma son (1,2,3,4,5)
- Nivel de Ingresos de los Padres, indica cual es el ingreso familiar de los padres, y los valores que puede tener: Bajo (0, 1 SMMLV], Medio (1,3 SMMLV], Alto (3, en adelante]
- Máximo Nivel de estudios alcanzado por el Padre, estos valores son: Ninguno, Primaria, Básica, Media, Técnico, Universitario-Superior.
- Máximo Nivel de estudios alcanzado por el Madre, estos valores son: Ninguno, Primaria, Básica, Media, Técnico, Universitario-Superior.
- Ocupación del padre: Jubilado, hogar, estudiante, busca empleo (Baja), trabajador independiente, empleado (Media), empresario, administrador, gerente, profesional independiente (Alta).
- Ocupación de la madre: Jubilada, hogar, estudiante, busca empleo (Baja), trabajador independiente, empleada (Media), empresaria, administradora, gerente, profesional independiente (Alta).
- Tipo de vivienda, Si es propia o No.
- Número de Hermanos, indica la cantidad de hermanos que tiene.
- Tipo de relación con los padres, la opinión del estudiante y su relación con sus progenitores, puede tomar los siguientes valores: Buena, regular, mala.

III. Variables Académicas

- *Tiempo de ingreso a la Universidad:* indica el número de semestres que el estudiante demoró en ingresar a la Universidad.
- Puntaje estandarizado de la prueba saber 11.
- Tipo de Colegio: índica el tipo de colegio de egreso del estudiante, tiene dos categorías Público o Privado.
- *Rendimiento académico durante el Colegio:* Cómo considera que fue este, pude ser: Excelente, Sobresaliente, Aceptable, Insuficiente.
- *¿Ha recibido algún tipo de orientación vocacional*? Indique Si o No.
- Énfasis del colegio de egreso de secundaria del estudiante: técnico agropecuario, comercial=1, académico=2, turístico=3, técnico=4.

- Orientación Vocacional previa recibida antes de ingresar a la universidad. Tiene dos categorías posibles: Sin O.V.=1 si no recibió ninguna orientación vocacional, O.V.=1 si recibió orientación.
- Empezar Universidad: Esta variable cuenta con cuatro categorías. E1: es la primera vez que el alumno comienza un estudio superior, COC: si además de la carrera iniciada, continua otra carrera, TOC: en caso en que el estudiante, al iniciar esta carrera, ya haya terminado previamente algún otro estudio superior. Y por último, AC: en caso que abandono una carrera anteriormente.

IV. **Información Institucional**

- ¿Ha recibido beca o financiamiento de la Universidad? Si, No.
- ¿El programa académico que cursa, cuenta con registro calificado? Si, No.
- ¿Considera que usted tiene una buena relación con los docentes? Si, No.
- ¿Considera que usted tiene una buena relación con sus compañeros? Si, No.
- ¿Ha iniciado otro programa académico en esta Institución u otra?, Indicar Si o no, además. Mencione ¿cuál? _____

ANEXO D: DISEÑO HOJA EXCEL PARA REGISTRO DE INFORMACIÓN DE LOS ESTUDIANTES (ANEXO TIPO DIGITAL)