UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"IDENTIFICACIÓN DE LAS VARIABLES DETERMINANTES EN EL CESE VOLUNTARIO DE UN COLABORADOR CON LA REGRESIÓN DE COX"

PRESENTADO POR

JAVIER NORBERTO PALOMINO GONZALES

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL

TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima – Perú

2020

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

"IDENTIFICACIÓN DE LAS VARIABLES DETERMINANTES EN EL CESE VOLUNTARIO DE UN COLABORADOR CON LA REGRESIÓN DE COX"

PRESENTADO POR

JAVIER NORBERTO PALOMINO GONZALES

TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL

TÍTULO DE INGENIERO ESTADÍSTICO E INFORMÁTICO

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

.....	.....
M.A. Fernando René Rosas Villena Presidente	Dr. Jorge Chue Gallardo Asesor
.....	.....
Mg. Iván Dennys Soto Rodríguez Miembro	Mg. Felipe De Mendiburu Delgado Miembro

Lima – Perú
2020

DEDICATORIA

A Dios por haberme brindarme su amor y fuerza a través de mi abuela Emma.

AGRADECIMIENTOS

Mis agradecimientos a todas las personas involucradas en la consecución del presente trabajo, profesores, profesoras y amigos del departo de Estadística e Informática.

Índice de Contenido

Índice de tablas

Índice de figuras

Índice de anexos

1. PRESENTACIÓN

El autor del presente trabajo de suficiencia profesional comenzó su experiencia laboral en una entidad bancaria de cobertura local, con muchos años de participación en el mercado peruano, haciendo uso intensivo de herramientas analíticas e informáticas. Las funciones desarrolladas en esta organización se centraron en implementar modelos analíticos para generar eficiencias tanto en uno de sus principales canales de atención, como en sus áreas de soporte. Por ejemplo, implementar un modelo de series de tiempo para el pronóstico de la demanda de uno de sus principales canales de atención, esto con el objetivo de proponer alternativas de planificación en la dotación del personal para el funcionamiento óptimo del canal. Otra de las principales funciones asumidas fue la creación y mantenimiento de reportes operativos cuyas fuentes se encontraron en diferentes tipos de bases datos y lenguajes de consulta, así como la respectiva visualización de los mismos.

Luego de 3 años de experiencia en esta entidad bancaria, el tesista pasó a trabajar a otra área analítica del sector bancario. En esta organización una de las principales funciones fue generar valor a partir de la optimización del tratamiento y análisis de los datos con el objetivo de impulsar el desarrollo óptimo del negocio. En esta empresa, se tuvo la oportunidad de haber aportado con implementar un proceso de extracción, transformación y carga de datos para varios procesos de construcción de indicadores, reduciendo considerablemente el tiempo de obtención de estos. También se elaboró herramientas de análisis de clasificación de sentimientos para ser aplicado en las preguntas abiertas de las diferentes encuestas de la organización. Adicionalmente se aplicó conocimientos de modelamiento de datos para identificar patrones y variables para la elaboración de estrategias en la retención de personal donde los valores de rotación eran demasiados altos. La experiencia del tesista en el sector bancario le ha permitido contemplar un mayor horizonte de las oportunidades acerca de la aplicación del modelamiento estadístico computacional, así como las principales debilidades a la hora de implementarlo. Con respecto al punto anteriormente mencionado, el tesista de la presente investigación considera importante seguir especializándose tanto en el conocimiento del negocio bancario, así como también el desarrollo de plataformas web que envuelva a la organización como parte de su desarrollo digital y dinámico; por todo ello, sus planes a futuro consisten en llevar una especialización en transformación digital y posteriormente una maestría en Estadística e Informática.

2. INTRODUCCIÓN

En presente trabajo de investigación se explica el desarrollo del modelamiento del evento renuncia voluntaria de los colaboradores de una organización financiera a través del modelo de regresión de Cox. Esta iniciativa fue de mucha utilidad para identificar cuáles son las variables que determinan este evento y cómo estas impactan en el evento renuncia. Así mismo es muy importante resaltar que para llegar a realizar este estudio, se tuvo que resolver varios desafíos respecto al conocimiento del negocio, así como la obtención los datos a través de una automatización de las consultas a las bases de datos de recursos humanos para obtener datos confiables y relevantes para el estudio. Dicha automatización fue de mucho aporte para la organización, ya que permitió ahorrar 15 días de proceso manual para su obtención, actualmente el proceso manual se ha reducido a actualizar un tablero de indicadores que solo demora 3 minutos en actualizar. Por ello consideramos que el presente estudio representa el trabajo de un año de comprender la dinámica interna de los procesos y a partir de estos proponer nuevas formas de abordar tanto los procesos, así como la formas de analizar los resultados del mismo.

Este trabajo se desarrolló a inicios del año 2019 y para su desarrollo fue necesario revisar los modelos conceptuales que enmarcan nuestro objeto de estudio: La rotación laboral y la regresión de Cox. La rotación laboral es un comportamiento normal dentro de las organizaciones, esta se define como "El retiro permanente de un trabajador de una organización, y puede ser voluntario o involuntario, una tasa de rotación elevada da como resultado costos más altos de reclutamiento, selección y capacitación" según (Robbins & Judge, 2013).

Para (Chiavenato, 2007) la rotación laboral es "La fluctuación de personal entre una organización y su ambiente; en otras palabras, el intercambio de personas entre la organización y el ambiente se determina por el volumen de personas que ingresan y salen de la organización. La rotación de personal se expresa mediante una relación porcentual entre los ingresos y las separaciones en relación con el número promedio de integrantes de la organización, en un periodo determinado. Casi siempre la rotación se concentra en índices mensuales o anuales, lo que permite comparaciones para elaborar diagnósticos, y prevenir o proporcionar alguna predicción". A partir de estas definiciones podemos identificar 2 principales elementos, una de ellas es; el evento decisión del colaborador en dejar la organización y la otra el tiempo de permanencia dentro de la organización, es decir desde el inicio del contrato hasta la fecha del

cese del colaborador. Entendido de esta manera la rotación laboral no sería un problema si es que esta ocurre dentro de los parámetros normales de la organización, según su sector de negocio y/o contexto en general. Por ejemplo, según la encuesta nacional de variación mensual del empleo que reporta mensualmente el Ministerio de Transporte y promoción del Empleo, el índice de rotación laboral para el sector industria manufacturera es por lo general mucho más elevado que la del sector transporte, almacenamiento y comunicaciones y esta a su vez pueden tener valores muchos más altos de acuerdo a su ubicación geográfica (rural, urbano) o incluso tomar valores de acuerdo al país donde se encuentren dichas organizaciones, ya que cada País agrega un factor cultural y legal laboral. En ese sentido, es importante notar que la rotación laboral se encuentra determinada por una complejidad de factores tanto endógenos como exógenos a la organización y su interacción entre ellas, es por ello que las organizaciones con un alto número de empleados, como es el caso del presente estudio, tienden a controlar las variables endógenas de la organización, promoviendo un adecuado clima laboral y un conjunto de estrategias de retención del personal, estableciendo todo un conjunto de esquema de beneficios que son utilizados en varios momentos del ciclo de vida del empleado. Es por esto que la conceptualización de la rotación y el esfuerzo por modelar su comportamiento tienen distintos enfoques y han venido desarrollándose cada vez más con el objetivo de entender cuáles son sus principales factores, como estas se relacionan para luego a partir de estas, identificar los valores de rotación permisibles o los cambios estructurales que deben realizarse para lograr cambios importantes en la reducción de dicho índice, si es que estos son demasiado altos. Ante esta complejidad por comprender las causas de la rotación y por cuanto el Perú es uno de los países de américa latina con las más altas tasas de rotación laboral, según (Maurizio, 2017), el autor ha visto conveniente identificar, a partir de un conjunto de variables disponibles de la organización, cuáles son las más significativas para explicar el evento renuncia del colaborador y cuantificar probabilísticamente su impacto en esta. En el presente trabajo de suficiencia profesional tiene como objetivo dar a conocer la metodología, como se indica en el objetivo general, así como en los objetivos específicos para luego pasar a explicar las funciones desempeñadas dentro de la organización que me permitieron aportar con el presente trabajo de suficiencia profesional. En a la sección 4.2 procedemos a detallar las características de la regresión de Cox, su supuesto de riesgos proporcionales, las características necesarias para su uso, así como la problemática a la cual respondió. Posteriormente revisaremos otros artículos

que abordan problemática similares pero haciendo uso de otras técnicas estadísticas, como por ejemplo, el uso de la regresión logística Multinivel de (Quispe Millones, 2014), para luego finalizar con el detalle paso a paso del uso de regresión de Cox en la rotación laboral en nuestro tema de aplicación.

3. OBJETIVOS

3.1. Objetivo general

Aplicación de la regresión de Cox para analizar e interpretar las variables determinantes en la rotación laboral de una posición comercial financiera.

3.2 Objetivos específicos

1. Usar el modelo de regresión de Cox para modelar el tiempo de permanencia de un colaborador de una posición masiva específica, considerando variables demográficas de la base de datos de recursos humanos de una organización bancaria.
2. Validar los supuestos de regresión de Cox y brindar una alternativa de procesamiento de ser el caso que existan variables que no cumplen con los supuestos.
3. Cuantificar e interpretar probabilísticamente el efecto de dichas variables demográficas e interpretar sus efectos en el evento cese del colaborador.

4. DESCRIPCIÓN DEL TRABAJO

Entre los principales aportes y experiencias profesionales obtenidos, fueron la introducción de herramientas analíticas, informáticas y computacionales a los diferentes procesos de la organización. Después de un breve tiempo en sectores no comerciales, la primera etapa laboral fue de carácter financiera, estas aplicaciones se desarrollaron muy de cerca a la parte comercial del banco a través de uno de sus canales estratégicos, por lo que se invirtió el tiempo suficiente para conocer la naturaleza de los datos, las reglas del negocio de cada uno de los productos involucrados en el canal, los repositorios de información, entre otras barreras para el adecuado uso de una solución integral para el negocio, como por ejemplo, una adecuada segmentación de

clientes de acuerdo al tipo de gestión realizado en dicho canal. Para ello previamente fue necesario la clasificación de todas las interacciones con el cliente para posteriormente identificar agrupaciones (clusters) y elaborar estrategias para cada una de ellas, con la mira de reducir la cantidad de interacciones para aquellos clusters en donde existen canales más económicos y que por muchos motivos, entre ellas, falta de conocimiento por parte del cliente, no lo usan.

Posteriormente pasé a otra empresa del mismo rubro, pero esta vez para prestar servicios a varias áreas de soporte de la organización, principalmente a sus áreas de gobierno. En esta empresa se logró romper con la caja negra que representaba, hasta en ese momento, algunos de los procesos para la elaboración de reportes. Se automatizó a través de consultas SQL (Structured Query Language), muchos de los reportes usados en el área, asimismo conectándolos a un tablero de visualización en Power BI (Business Inteligence).

En esta misma empresa también se implementó un modelo máquinas de aprendizaje para la clasificación de sentimientos, haciendo uso del algoritmo de clasificación de Bayes y del análisis de texto. Esta herramienta fue desarrollada con el software R y fue usada para la clasificación de respuestas en 5 tipos de emociones, así como en la polaridad de estas (sentimiento positivo y sentimiento negativo) aplicado a las encuestas de clima organizacional.

4.1. Funciones Desempeñadas

Las principales funciones desempeñadas en mi ejercicio profesional han tenido un alto componente del conocimiento del negocio para la correcta extracción, transformación y carga de datos. En una primera etapa fue la elaboración y validación de reportes de indicadores de resultados y en una segunda etapa para la elaboración de proyectos analíticos. Por ejemplo, para el proyecto de segmentación de clientes de acuerdo con el tipo de gestión realizado en uno de los canales de atención del cliente, los registros superaban los 450 mil registros por mes, por ello fue necesario trabajar la estructura de la data final a través de un motor de base de datos en Oracle, donde inicialmente se encentraban los datos. Para llegar a esta propuesta de segmentación, fue necesario antes conocer muy bien los principales reportes relacionados a estos, este conocimiento pudo ser obtenido gracias al conocimiento detrás de cada uno de los reportes y su mantenimiento día a día.

Las funciones en la actual organización se centran principalmente como líder de proyectos analíticos y el seguimiento de implementación de los principales proyectos de transformación digital de la organización, como parte de estas tareas también se hace el mantenimiento de los principales procesos para la obtención de indicadores.

4.2. Puesta en práctica de lo aprendido en la carrera

4.2.1. Descripción de las técnicas estadísticas y/o informáticas utilizadas en la solución situación problemática en el ejercicio de su actividad laboral.

En la presente sección pasaremos a describir la técnica de regresión de Cox. Si bien hemos realizado todo el análisis aplicando el paquete estadístico *"survival"* del software R, pasaremos a mencionar las bases conceptuales que permitieron Terry Therneau implementarlo en dicho software y que se encuentra disponible como recursos abierto en los servidores de R-cran. Antes de describir los puntos principales de dicha técnica, es importante mencionar que esta pertenece a un amplio espectro de técnicas para el análisis de supervivencia. Según (Borges, 2005), una técnica de análisis de supervivencia tiene como objeto de estudio el tiempo de seguimiento hasta la ocurrencia de un evento de interés y cobra vital importancia cuando existen observaciones censuradas. En el presente estudio usaremos el concepto de censura por la derecha, ya que también existen observaciones censuradas de otros tipos. Una observación censurada por la derecha es aquella unidad de estudio que no presentó el evento de interés hasta la finalización del estudio.

Este modelo, trabaja primordialmente con la función de riesgo, también llamado función de Hazard y es utilizado para detectar relaciones existentes entre el riesgo que se produce en un determinado individuo en el estudio y algunas variables independientes y/o explicativas; por lo que este modelo nos permite evaluar dentro de un conjunto de variables cuáles tienen relación, influencia sobre la función de riesgo y por ello también en la función de supervivencia, ya que ambas funciones están conectadas.

Si bien, la técnica de regresión de (Cox, 1972) es ampliamente utilizado en el área médica, también existen aplicaciones en otros ámbitos como lo son la industria biosanitaria, financiero, forestal, industria mecánica y como lo es en la presente investigación, a nivel empresarial, esto con el objetivo de modelar variables de gestión humana. Tal como lo menciona en una memoria

sobre las bases teórica de la regresión de (Cox, 1972) según (Velasco Álvarez, 2016). El modelo de Cox viene representado de la siguiente manera:

$$\lambda\left(t; Z_i\left(t\right)\right) = \lambda_0\left(t\right) e^{\beta^t Z_i(t)}$$

Donde $Z_i(t)$ es el vector de covariables para el i-ésimo individuo en el tiempo t. Se dice que es un modelo semiparamétrico debido a que incluye una parte paramétrica y otra no paramétrica; La parte paramétrica es $r_i\left(t\right) = e^{\beta' Z_i(t)}$, llamada puntaje de riesgo (*risk score*) y β es el vector de parámetros de la regresión. La parte no paramétrica es $\lambda_0\left(t\right)$, llamada función de riesgo base y es una función arbitraria no especificada.

El modelo de regresión de Cox es también llamado modelo de riesgos proporcionales debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante sobre el tiempo, es decir:

$$\frac{\lambda\left(t; Z_i\left(t\right)\right)}{\lambda\left(t; Z_j\left(t\right)\right)} = \frac{\lambda_0\left(t\right) e^{\beta' Z_i(t)}}{\lambda_0\left(t\right) e^{\beta' Z_j(t)}} = \frac{e^{\beta' Z_i(t)}}{e^{\beta' Z_j(t)}}$$

Respecto al contraste de hipótesis, a partir de la función de verosimilitud parcial podemos obtener una estimación de los coeficientes $\hat{\beta}$ cuya distribución es aproximadamente normal de media β y matriz de varianzas y covarianzas $\sum = \phi^{-1}\left(\beta\right)$, que puede ser estimada por $\hat{\Sigma} = \phi^{-1}\left(\underline{\hat{\beta}}\right)$. Para contrastar la hipótesis $H_0 : \beta_j = 0$ *v*s. $H_1 : \beta_j \neq 0$, es decir, la significación de la *j*-ésima covariante en el modelo, se puede utilizar el estadístico de Wald, contraste de razón de verosimilitudes y el contraste score, también conocido como log Rank. Para el presente estudio utilizaremos el contraste de Wald ya que tiene una interpretación más directa que los otros dos mencionados. El contraste de Wald se basa en $\hat{\beta}$ que sigue asintóticamente una distribución aproximadamente normal. Se considera el estadístico:

$$X_w = \left(\hat{\beta} - \beta_0\right)' \phi\left(\hat{\beta}\right)\left(\hat{\beta} - \beta_0\right)$$

que, bajo hipótesis nula, sigue una distribución chi-cuadrado con p grados de libertad.

Luego de mencionar como se calcula el modelo de regresión de Cox, se indica cómo se selecciona el mejor modelo, ya que no todas las variables presentes en el estudio tienen un

impacto significativo en el evento estudiado, que para el trabajo profesional es la renuncia del colaborador. El método para identificar el mejor modelo en la presente aplicación, fue propuesta por (Collet, 1994) y consta de los siguientes pasos:

1.- Ajustar todos los modelos con una sola covariable. Luego, incluir todas las covariables cuya contribución resultó significativa a un nivel del 5% o el nivel de significación seleccionado.

2.- Las covariables que contribuyen en forma significativa en el paso 1 se incluyen en el modelo y se ajusta conjuntamente. La presencia de ciertas covariables puede hacer que otras no contribuyan significativamente. Entonces se ajustan modelos reducidos, excluyendo una única covariable en cada ajuste que no es significativa. Solamente aquellas que contribuyan significativamente permanecerán en el modelo.

3.- Se ajusta un nuevo modelo con las covariables retenidas en el paso 2. En esta etapa las variables excluidas en el paso 2 retornan al modelo para confirmar que sus contribuciones no son estadísticamente significativas.

4.- Las eventuales covariables significativas en el paso 3 son incluidas al modelo conjuntamente con aquellas retenidas en el paso 2. En este paso se retornan las variables excluidas en el paso 1, para confirmar si ellas contribuyen o no significativamente al modelo.

5.- Ajustar un modelo incluyendo las covariables que contribuye significativamente en el paso 4. En este paso se prueba si algunas de ellas pueden ser retiradas del modelo.

6.- Utilizando las covariables que fueron retenidas en el paso 5 se ajusta el modelo final para los efectos principales. Para completar el modelo se debe verificar la posibilidad de incluir términos de interacción. Se prueba cada interacción de dos posibles covariables entre aquellas incluidas en el modelo. El modelo final queda determinado por los efectos principales identificado en el paso 5 y los términos de interacción que contribuye en forma significativa que fueron identificados en este paso.

Por último, según (Therneau, Grambsch, & Felimng, 1990) se pueden usar varios tipos de residuales para validar 4 características principales del modelo final, como son; descubrir la correcta forma funcional de los predictores, identificar a los sujetos peores predichos del modelo, identificar los puntos influyentes y finalmente evaluar las suposiciones de la proporcionalidad tanto global como por cada covariable a través de una prueba estadística donde

se busca no rechazar la hipótesis planteada, ya que si se rechaza ésta se concluiría que se incumple el supuesto de riesgos proporcionales. La interpretación gráfica de los residuales de Schoenfeld la realizaremos según (Harrel & Lee, 1986), que nos menciona que si los residuos mantienen un patrón aleatorio, es decir no sistemático, proporcionan una evidencia de que el efecto de la covariable no cambia respecto del tiempo, algo que presupone el modelo de Cox, por lo contrario si hay algún tipo de patrón sistemático, sugiere que el efecto de la covariable cambia a lo largo de tiempo. Por ende, los residuales no deberían mostrar tendencias temporales, es decir, la pendiente el plot de los residuos frente al tiempo debe ser nula. Por eso los gráficos pueden acompañar a la prueba de hipótesis global del supuesto de riesgos proporcionales. De ser el caso de existan variables que no soporten la prueba de hipótesis de riesgos proporcionales tenemos la opción de usar el modelo de Cox estratificado, que como bien resume (Roque, 2009) admite que la forma de la función de riesgo varíe según los estratos o niveles de las covariables Lo cual significa que: Supongamos que tenemos un predictor X, por su naturaleza propia, éste X puede ser categorizado en varios niveles, es decir, en subpredictores secundarios, esto conlleva a que el modelo sea ajustado para cada subpredictor Z. De modo que el modelo de Cox para cada estrato será definido de la manera siguiente:

$$\lambda_{ij}(t) = \lambda\left(t / X_j, \mathrm{Z} = j\right) = \lambda_{0j}(t) e^{Z_{ij}(t)^t \beta}$$

El modelo de Cox estratificado asume que las covariables actúan de modo similar en la función de riesgo de base de cada estrato, es decir, se asume que β es común para todos los estratos lo cual, ésta suposición debe ser probada realizando una vez más la prueba de validación de supuestos de riesgos proporcionales, como se describe en el artículo de (Roque, 2009).

4.2.2 Revisión de literatura

En (Borges, 2005)**,** se aplicó el análisis de supervivencia clásico de Kaplan Meier, y modelos de regresión basados en técnicas más recientes como la regresión de Cox. La suposición principal fue de riesgos proporcionales. La verificación gráfica de este supuesto de riesgos proporcionales puede observarse en las partes (a), (b) y (c) del anexo 3. Borges menciona que en estos gráficos no se observa incumplimiento del supuesto en cada una de las covariables. Sin embargo, para el caso de la edad, se observa quizás un patrón cíclico muy atenuado, lo que da indicios de la necesidad de ajustar un modelo de Cox con covariables dependientes del tiempo,

pudiendo en este caso utilizarse los modelos frágiles (frailty models), debido a la variación temporal de los residuos de Schoenfeld. Luego Borges procede a la verificación del supuesto de riesgos proporcionales a través de un contraste de hipótesis, donde la hipótesis nula está asociada al cumplimiento del supuesto de riesgos proporcionales. Los resultados de este contraste indican que no se incumple el supuesto de riesgos proporcionales para ninguna de las tres covariables. Los p-valores asociados a este contraste para diabetes, edad e índice de Quetellet son 0.776, 0.305 y 0.633, respectivamente, observándose que todos son mayores que 0.10, con lo que no se estaría rechazando la hipótesis de riesgos proporcionales para ninguna de las covariables. Este contraste permite verificar el incumplimiento global del supuesto de riesgos proporcionales de todas las covariables. En este caso se obtiene un p-valor de 0.71, y por ser éste mayor que 0.10, no se rechazaría la hipótesis nula de cumplimiento conjunto del riesgo proporcional de las tres covariables.

Tabla 1: Modelo de Cox para DPA y muerte como evento de interés.

Covariables	Coeficiente	*p*- valor
Edad	0.0315	0.0011
Índice de Quetelet	-0.0969	0.013
Diabetes	0.5492	0.087

Fuente: (Borges, 2005)

Los datos del estudio de Borges corresponden a 246 pacientes en diálisis peritoneal (DPA) que acudían al Servicio del Hospital Clínico Universitario de Caracas entre 1980 y 1997. Se hizo un seguimiento a los pacientes desde el comienzo de sus sesiones de diálisis hasta alcanzar la muerte como evento de interés, o hasta la terminación del estudio, por lo que algunas observaciones resultan censuradas. En el análisis inicial se incluyeron 100 covariables dicotómicas y 16 continuas. Las etapas del proceso de aplicación de la técnica estadística comenzaron con el ajuste de varios modelos de Cox para obtener las covariables significativas, eliminando las variables no significativas mediante el procedimiento paso a paso hacia atrás, concluyendo que para el caso de los pacientes que acudían al servicio de diálisis peritoneal del Hospital Clínico Universitario de Caracas, Venezuela entre los años 1980 y 1997, las covariables significativas en el modelo de Cox fueron la diabetes la edad y el índice de Quetellet. Estas covariables son las que estarían modificando el riesgo de muerte en los pacientes en

diálisis peritoneal. Se concluye además que el modelo de riesgos proporcionales presentado es adecuado ya que todos los supuestos se verifican.

En la investigación realizada por (Quispe Millones, 2014), el objetivo de la investigación fue identificar las características de los trabajadores que se retiran de la empresa antes del término del periodo de prueba, de modo que se pueda mejorar el proceso de selección, implementar programas de retención y comprobar la diferencia de perfiles de trabajadores que cesan durante el periodo de prueba, de acuerdo al área donde trabajaron. Los supuestos del modelo de regresión logística fueron: la relación entre la variable respuesta y los predictores es lineal, las varianzas de los errores no son constantes (heterocedasticidad) y no necesariamente se distribuyen normalmente. El modelo presentado considera como variable respuesta una variable continua con distribución normal y cuya medición se da en el primer nivel.

La población está formada por los trabajadores de la empresa en estudio, que se desvincularon, cesaron o renunciaron, durante el tiempo de actividad de la empresa. No fue necesario realizar una encuesta para la recolección de datos por la accesibilidad al sistema de información existente. Se consideraron las desvinculaciones de trabajadores presentadas durante el periodo enero 2010 - agosto 2012. Los datos fueron obtenidos de una fuente secundaria, el sistema de información EXACTUS de donde se obtuvieron campos como fecha de nacimiento, edad, sexo, dirección, estado civil, número de hijos, motivo de cese, fecha de ingreso, fecha de cese, puesto y área asociados al trabajador que se desvinculó de la empresa. Se clasificaron todos los casos de desvinculaciones en dos grupos, ceses y renuncias. En el primer grupo se consideraron los casos donde la empresa decidió la desvinculación, anulación de ingreso o cambio de modalidad de contrato. En el grupo renuncias, se consideraron las desvinculaciones a solicitud del trabajador por motivos asociados a estudios, otro trabajo, viaje, salud u otros. La empresa cuenta con más de 20 áreas organizacionales, sin embargo, se decidió considerar las 10 áreas relacionadas a los negocios más importantes, donde existe mayor rentabilidad y riesgo en la operación. Para el desarrollo y análisis de resultados, se consideraron 2249 renuncias en el periodo enero 2010 - agosto 2012, distribuidas en 10 áreas.

Tabla 2: Distribución de renuncias por área en el periodo enero 2010 – agosto 2012

Área	Total
Canales	391

Área	Total
Distribución	368
Blindados	267
Sucursales	267
Multiser	251
Administrativos	176
Seguridad	170
Procesamiento	167
Bpo	114
Seguridad Externa	78
Total	2249

Fuente: (Quispe Millones, 2014)

Tabla 3: Definición de variables

Campo Inicial	Variable	Observaciones
Fecha De Nacimiento	Edad	Edad del trabajador
Sexo	Sexo	Sexo del trabajador
Dirección	Distancia	Distancia del domicilio del trabajador a LA EMPRESA. La clasificación se realizó tomando como criterio la distancia entre el distrito limeño de residencia y el distrito donde se ubica el centro de trabajo. Solo en el caso de provincias se consideró como sucursales.
Estado Civil	Estado Civil	Casado: se consideran los casados y convivientes. Soltero: se considera solo a los trabajadores que hayan declarado ser solteros a su ingreso.
Número de Hijos	Número de Hijos	Número de hijos registrados en el sistema de la empresa
Fecha de Ingreso Vs Fecha de Cese	Renunció antes del Periodo de Prueba	Indica si el trabajador renunció pasado el periodo de prueba o no. El periodo de prueba es de 6 meses desde su ingreso.
Puesto	Área	Área de trabajo, sector de negocio.
Área	Grupo Ocupacional	Clasificación del puesto que desempeña el trabajador de acuerdo a la misión del puesto.
Horario De Trabajo	Escala Remunerativa	Intervalo remunerativo asignado para fijar el total de ingresos percibidos por el colaborador de acuerdo al área y puesto que desempeña.
Evaluación de Desempeño	Evaluación Del Jefe	Puntaje obtenido en la última evaluación de desempeño realizada al jefe del área.
Número de Trabajadores	Número de Trabajadores	Número de trabajadores por área.

Fuente (Quispe Millones, 2014)

El análisis que aplicó (Quispe Millones, 2014) tuvo dos fases, en la primera se realizó el estudio descriptivo de las variables involucradas. Con la finalidad de validar la correcta selección de variables y analizar la asociación de las variables regresoras propuestas con el hecho de renunciar antes de culminar el periodo de prueba, se realizó el análisis de regresión logística. Por último, se planteó un modelo de regresión logística multinivel, de intercepto aleatorio, y se evaluó la existencia de diferencias en las características de los renunciantes entre áreas. Se compararon los resultados del modelo de regresión logística con los obtenidos con el modelo de regresión logística multinivel. Para la definición del modelo se consideraron 2249 (N=2249) retiros presentados en el periodo establecido agrupados en 10 áreas (J=10). Las variables en el primer nivel fueron 5 (P=5): X1: Edad del trabajador, X2: Sexo del trabajador, X3: Distancia, X4: Estado civil, X5: Número de hijos. Las variables dummy creadas a partir de la variable Distancia son: X31: Distancia lejos, X32: Distancia medio, X33: Distancia sucursales. Las variables en el segundo nivel fueron 5 (L=5): W 1: Grupo ocupacional, W2: Escala remunerativa, W3: Evaluación del jefe, W4: Número de trabajadores, W5: Cuenta con beneficios adicionales. El modelo finalmente ajustado fue:

$$\begin{aligned}\text{Logit}(\pi(X)) = {} & 1.386 + 1.180 * \text{Grupo ocupacional operativo}_i + 0.020 \\ & * \text{Escala remunerativa}_i + 0.0638 * \text{Evaluación del jefe}_i - 0.295 \\ & * \text{Número de trabajadores mayor a } 300_i - 0.013 \\ & * \text{Beneficios adicionales no recibidos}_i + 0.583 \\ & * \text{Edad del trabajador de 18 a } 30_i + 0.169 \\ & * \text{Sexo del trabajador masculino}_i + 0.099 * \text{Distancia lejos}_i - 0.224 \\ & * \text{Distancia medio}_i - 0.315 * \text{Distancia sucursal}_i - 0.376 \\ & * \text{Estado civil}_i - 0.068 * \text{Número de hijos}_i + \varepsilon_i\end{aligned}$$

La investigación desarrollada por (Flores Flores, 2011) está relacionada al uso de la regresión de Cox con métodos flexible en pacientes con linfoma no Hodgkin. El objetivo del trabajo fue analizar el efecto de las covariables en la supervivencia de un grupo de pacientes con linfoma no Hodgkin, utilizando métodos flexibles como los P-splines y polinomio fraccional para aproximar el efecto de las covariables en el contexto del modelo de Cox, cuando el supuesto de los riesgos proporcionales no se verifica y el efecto de las covariables no presenta una estructura

de relación lineal. Los datos recopilados de las variables relacionadas a las características del paciente y del tumor (características clínicas) fueron;

i. Edad: en años.
ii. Género: Femenino o masculino.
iii. Zubrod: Estado funcional del paciente, según la escala ECOG.
iv. Primario: Localización ganglionar o extra ganglionar.
v. Tumor: Diámetro mayor del tumor.
vi. Número de ganglios afectados.
vii. Número de sitios extra ganglionares.
viii. Estadio clínico: Extensión de la enfermedad, según la clasificación on Ann Arbor.
ix. Sitios de metástasis: Extensión del tumor a otras regiones o órganos.
x. Síntomas: Fiebre, sudoración nocturna o baja de peso sin causa alguna.
xi. Tipo de LNH: Subtipo histológico, según la clasificación disponible.
xii. VIH/SIDA: Infección por VIH o SIDA.
xiii. Hemoglobina: en g/dl.
xiv. Leucocitos: Número de leucocitos $/mm^3$
xv. Linfocitos: Linfocitos en porcentaje.
xvi. Deshidrogenasa láctica: en UI/L.
xvii. β 2-micro globulina: en mg/L

(Flores Flores, 2011) menciona en su investigación una etapa importante previa de exclusión de algunas variables como tamaño del tumor, número de ganglios y el sitio de metástasis, debido a que estas variables ya están reflejadas en el estadio clínico. Así mismo, no se incluye el subtipo histológico y el genotipo (células T o B) debido a que los pacientes fueron diagnosticados con tres criterios de clasificación histopatológica diferentes (Rappaport y Kiel, formulación de trabajo (WF) y la clasificación REAL). La variable β2M tampoco fue incluida debido a que este dato había sido solicitado en muy pocos pacientes. En los resultados se observa que todas las covariables, a excepción del foco primario ($p = 0.580$), presentan un efecto significativo ($p<0.05$) en la supervivencia de los pacientes con LNH. La razón de riesgo de estas variables implica que, los pacientes de sexo masculino presentan un riesgo de mortalidad de HR=1.2 (IC5 %:1.1-1.4) veces más que los pacientes de sexo femenino. Los pacientes con zubrod 2-4

presentan un riesgo de mortalidad de HR=2.0 (IC95 %: 1.7-2.4) veces más que los pacientes con zubrod 0-1. Los pacientes con enfermedad avanzada (EC III-IV) presentan un riesgo de mortalidad de HR=1.6 (IC95 %: 1.3-1.8) veces más que los pacientes con enfermedad temprana (EC I-II). Para el logaritmo de los leucocitos, el riesgo de mortalidad se incrementa en HR=1.24 por cada unidad que incrementa los leucocitos, así mismo, para el logaritmo del DHL el riesgo de mortalidad se incrementa en 1.3 por cada unidad que incrementa la DHL. También se verificó el supuesto de riesgo proporcional basado en los residuos de Schoenfeld escalado y test de no proporcionalidad de Therneau y Grambsch y la forma funcional del efecto de las covariables en la función de riesgo basados en los residuos martingalas. En los resultados se observa que el efecto de las covariables no es constante ($p < 0.05$). Por tanto, los resultados de las estimaciones bajo el modelo de Cox son discutibles, debido a que el modelo no cumple el supuesto de riesgos proporcionales (Test de no proporcionalidad global: $p < 0.001$).

4.2.3 Propuesta de alternativa de solución a la situación problemática siguiendo las seis fases de la Metodología para el Desarrollo de Proyectos:

Para comprender el contexto de la presente investigación es necesario mencionar la importancia de un proyecto previo que consistió en automatizar las consultas de las bases de datos del personal, esta experiencia permitió conocer que datos y de qué forma se almacenan, de tal manera que no hubo una dependencia de otros reportes que podrían tener menos confiabilidad. Ante la disponibilidad de datos de rotación del personal de toda la empresa se corroboró que la rotación de las posiciones masivas comerciales fueron significativamente altas comparados a otras familias de puestos, ante esta problemática y las razones mencionadas en la introducción se procedió a focalizar e identificar las variables para este tipo de posiciones que se presentan en la tabla 4.

Tabla 4: Definición de variables del estudio

Variable	Descripción
Dias_Perm	Tiempo (en Días) de permanencia en la posición comercial.
Censura	Indica 1 si el colaborador(a) renunció en algún momento del estudio, 0: Si el colaborador permanece.
Age	Edad del colaborador(a) la hora de la renuncia o al finalizar el estudio.
Gender	Sexo del colaborador(a).
L_P	Si el colaborador(a) pertenece y labora en Lima.

Variable	Descripción
Region_N	Si el colaborador(a) pertenece y labora en la región norte del País.
Region_C	Si el colaborador(a) pertenece y labora en la región central del País.
Region_S	Si el colaborador(a) pertenece y labora en la región sur del País.
Union	Si el colaborador(a) es sindicalizado o no.
EDD_Category	Si el colaborador(a) tiene o no una evaluación de desempeño.
EDD	La nota de la evaluación de desempeño.
Q_Children_Number	Número de hijos del colaborador(a).
Age_Leader	Edad del líder del colaborador(a).
Leader_Gender	Sexo del colaborador(a).
Q_Children_Leader	Número de hijos del colaborador(a).
Cat_Leader_Edd	Si el leader del colaborador(a) tiene o no una evaluación de desempeño.
Leader_EDD	La nota de la evaluación de desempeño del líder del colaborador.
Leader_Union	Si el líder del colaborador es sindicalizado o no.
FTE	Si el colaborador(a) tiene jornada a tiempo completo.
Status_couple	Si el colaborador(a) pertenece al estado Casado o conviviente.
Status_couple_Leader	Si el líder del colaborador(a) pertenece al estado Casado o conviviente.
Relation_location	Si la localidad del colaborador y su respectivo lugar de trabajo pertenecen al mismo distrito.
Relation_Generation_1	Si la generación del líder pertenece a una mayor que la del colaborador(a).
Relation_Generation_2	Si el líder y colaborador(a) pertenecen a la misma generación.
Relation_Generation_3	Si la generación del colaborador(a) es mayor que la de su respectivo líder.

Fuente de elaboración propia

Para efectos de esta investigación se tomaron solo los ceses voluntarios, es decir se retiraron los registros de ceses no voluntarios, así como los ceses tempranos menores a 3 meses quienes fueron retirados por no pasar la etapa de prueba, tampoco se consideraron a los colaboradores activos menores a 3 meses por no tener la certeza de que hayan pasado el periodo de prueba. Se tomaron los datos históricos de 10 años de contrataciones, es decir las contrataciones realizadas entre el 2008 y finales del 2018. En el análisis se excluyeron a los empleados sindicalizados, que son los empleados que tienen un sesgo en su tiempo de permanencia. Bajo estas consideraciones se construyó una base de 1023 colaboradores.

Respecto a las variables, estas fueron seleccionadas de acuerdo a la opinión de los especialistas, tomando en cuenta su relación con el cese voluntario y que contenían información en la base de datos de la organización. Algunas de estas variables también fueron validadas por el estudio de (Quispe Millones, 2014), cuya investigación se menciona en la revisión de la literatura.

La preparación de los datos se desarrolló en el software R (R Core Team, 2018) con la librería *dplyr* que forma parte de una serie de paquetes orientados a la ciencia de datos. Con este paquete se verificó que no existían valores nulos, luego se renombraron las variables categóricas como Gender, que se encontraban nombradas por letras, pero que se decidió renombrar, por ejemplo, para el caso de esta variable por 0 al género femenino y 1 al género masculino. Luego se utilizaron las funciones *filter* y *mutate* para construir las variables Relation_Generation_1 y Relation_location.

Una vez obtenido los datos, se obtuvo el modelo usando la función *coxph* de la librería survival. Cómo se menciona en la parte teórica para seleccionar el mejor modelo se utilizó el método de (Collet, 1994), resultando al final las nueve variables significativas que aparecen en la tabla 5 y que se mencionan a continuación: edad del colaborador, género del colaborador, el género del líder del colaborador, si el colaborador trabaja a la región sur del Perú, la cantidad de hijos del colaborador, la edad del líder, si el colaborador es de jornada completa o parcial, EDD_Category y la variable Cat_Leader_Edd.

Tabla 5: Resultados del modelo final

Variables	coef	exp(coef)	se(coef)	z	Pr(>|z|)	ns
Age	-0.15665	0.855000	0.024354	-6432	1.26e-10	***
Gender1	0.224555	1251766	0.127510	1761	0.078226	.
Region_S1	0.528829	1696944	0.185770	2847	0.004418	**
EDD_Category1	-1005105	0.366006	0.138272	-7269	3.62e-13	***
Q_Children_Number	-0.31579	0.729210	0.123409	-2559	0.010500	*
Age_Leader	-0.03242	0.968098	0.008343	-3886	0.000102	***
Cat_Leader_Edd1	-0.82166	0.439701	0.150410	-5463	4.69e-08	***
FTE1	-0.50983	0.600594	0.270963	-1882	0.059895	.
Leader_Gender1	0.411788	1509514	0.129996	3168	0.001536	**

n= 1023, number of events= 274

Fuente de elaboración propia

Como se puede apreciar todas las variables son significativas a valores inferiores de nivel de significancia 0.06, luego se procedió a verificar gráficamente el comportamiento de los residuales para cada una de las variables y los supuestos de riesgos proporcionales del modelo, usando la librería *survminer*. Usando la función *ggcoxzph* se obtiene el gráfico 1, donde se observa el comportamiento de los residuales para las variables edad, género y Región_S1 respectivamente. Cada gráfico de los residuales viene acompañado por su respectivo valor de test individual que se muestra en la figura 1.

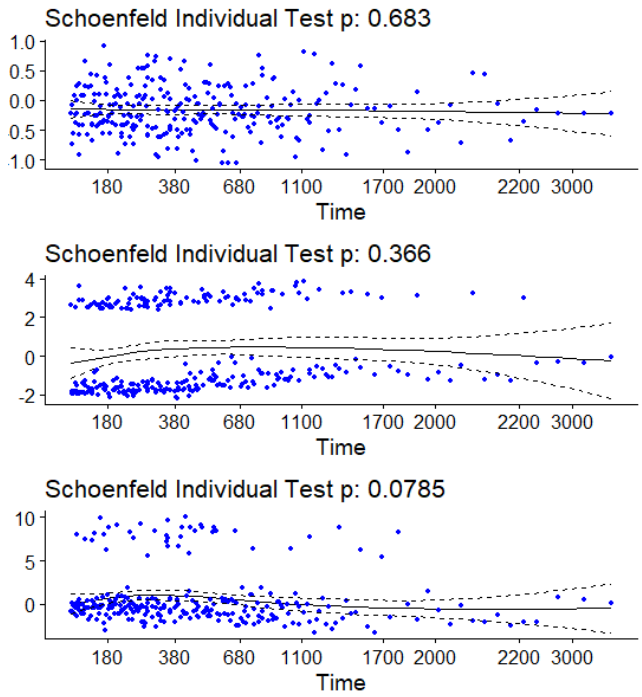


Figura 1: Residuales de Schoenfeld para las variables Edad, Sexo y Región Sur
Fuente: Elaboración propia

En el gráfico 1, se puede observar el comportamiento de los residuales de Schoenfeld para las variables edad, género y la variable indicadora de la región sur. Según este gráfico no se evidencia un comportamiento sistemático de los residuales para la variable edad, pero si una leve pendiente para la variable género y la variable indicadora de la región sur, pero esto parece ser leve ya que sus respectivos p-valores son significativos para valores mayores de 0.07. Ahora bien, si se pone atención a todos los p-valores de la tabla 6, se notará que existen 3 variables que tienen un p-valor menor a 0.02 con lo que se rechazaría la hipótesis del cumplimiento del

supuesto de riesgos proporcionales individuales para esas variables y además de la prueba global de riesgos no proporcionales con un p-valor de 4.62e-07.

Tabla 6: Prueba para el Supuesto de riesgos proporcionales

Variables	rho	chisq	p-valor
Age	-0.02499	1.67e-01	6.83e-01
Gender1	0.05404	8.17e-01	3.66e-01
Region_S1	-0.10290	3.10e+00	7.85e-02
EDD_Category1	0.15330	4.91e+00	2.68e-02
Q_Children_Number	-0.00152	6.47e-04	9.80e-01
Age_Leader	0.05268	8.19e-01	3.65e-01
Cat_Leader_Edd1	0.31114	2.51e+01	5.36e-07
FTE1	-0.02121	1.21e-01	7.28e-01
Leader_Gender1	-0.13791	5.19e+00	2.27e-02
GLOBAL	NA	4.66e+01	4.62e-07

Fuente: Elaboración propia

Ahora, obsérvese la siguiente figura 2 en donde se presentan a las variables Cat_Leader_Edd1, EDD_Category1 y Leader_Gender1 que son las que no cumplen con el supuesto de riesgos proporcionales y para los cuales se muestran su respectivo gráfico de residuales en la figura 2, mientras que para los gráficos de las variables que si cumplieron con el supuesto de riesgos proporcionales se encuentran en los anexos 1 y 2.

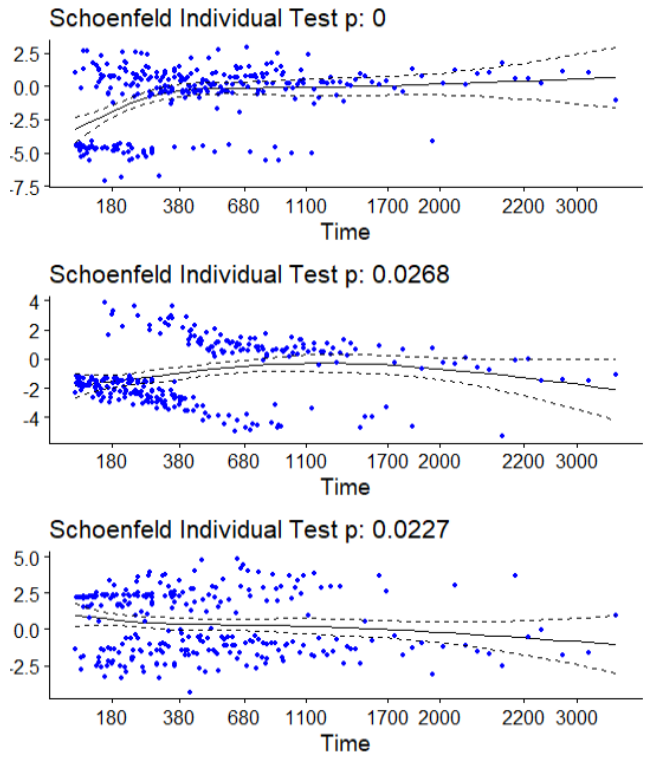


Figura 2. Residuales de Schoenfeld para las variables Cat_Leader_Edd1, EDD_Category1 y Leader_Gender1.

Si se observan los residuales de Schoenfeld de las variables que tienen un p-valor menor a 0.02 que se muestran en la figura 2, se nota que para Cat_Leader_Edd1 existe una pendiente positiva, aunque parece ser leve, la línea ajustada de los residuales tiende a crecer. De la misma manera la variable EDD_Category1 y Leader_Gender1 pero con pendiente negativa. Por último, la prueba global de riesgos proporcionales para estas variables, mostrada en la tabla 6, vemos que el p-valor es muy pequeño, con lo cual corroboraría que el modelo global no cumple con el supuesto de riesgos proporcionales. Ante esta situación se procedió a estratificar las variables categóricas Cat_Leader_Edd1, Leader_Gender1 y por último EDD_Category1 ya que fueron estas las variables que no soportaron la prueba de riesgos proporcionales y se ejecutó nuevamente un modelo con dichas estratificaciones resultando un modelo con todas las variables significativas para niveles menores a 0.07 tal como se muestra en la tabla 7.

Tabla 7: Resultados para modelo estratificado de Cox.

Variables	coef	exp(coef)	se(coef)	z	Pr(>|z|)	ns
Age	-0.15326	0.857902	0.024378	-6287	3.23e-10	***
Gender1	0.237397	1.267945	0.129185	1838	0.066114	.
Region_S1	0.642508	1.901243	0.187136	3433	0.000596	***
Q_Children_Number	-0.29210	0.746690	0.124331	-2349	0.018803	*
Age_Leader	-0.03259	0.967931	0.008827	-3692	0.000222	***
FTE1	-0.66756	0.512959	0.277087	-2409	0.015987	*

Fuente de elaboración propia

Como puede apreciarse las variables que fueron estratificadas no se muestran en el modelo ya que como se mencionó en la parte teórica los individuos fueron agrupados según los estratos de las variables categóricas Cat_Leader_Edd1, Leader_Gender1 y EDD_Category1. De estos estratos para corroborar que el vector β es común a todos los estratos, es corroborado usando la función cox.zph, cuyo resultado se muestra a continuación en la tabla 8.

Tabla 8: Prueba para el Supuesto de riesgos proporcionales.

Variables	Rho	Chisq	p
Age	-0.05413	0.772726	0.379
Gender1	0.047947	0.637559	0.425
Region_S1	-0.08748	2063357	0.151
Q_Children_Number	0.000765	0.000163	0.990
Age_Leader	0.032129	0.339315	0.560
FTE1	0.006360	0.010640	0.918
GLOBAL	NA	4875045	0.560

Fuente de elaboración propia

De acuerdo a los resultados de la tabla 8, se observa que las pruebas estadísticas no son significativas, tanto en las pruebas individuales como para la prueba global, esto respalda el supuesto de riesgos proporcionales; es decir, no se rechaza la hipótesis planteada que indica que el modelo cumple con el supuesto de riesgos proporcionales y por consiguiente que el vector de coeficientes pueda considerarse los mismos para cada estrato. También se corrobora gráficamente que se cumple con el supuesto de riesgos proporcionales a través del análisis de residuos del modelo de Cox estratificado que se presentan en las figuras 3 y 4. Nótese que para cada variable no existe ningún patrón sistemático en el comportamiento de los residuales con el tiempo.

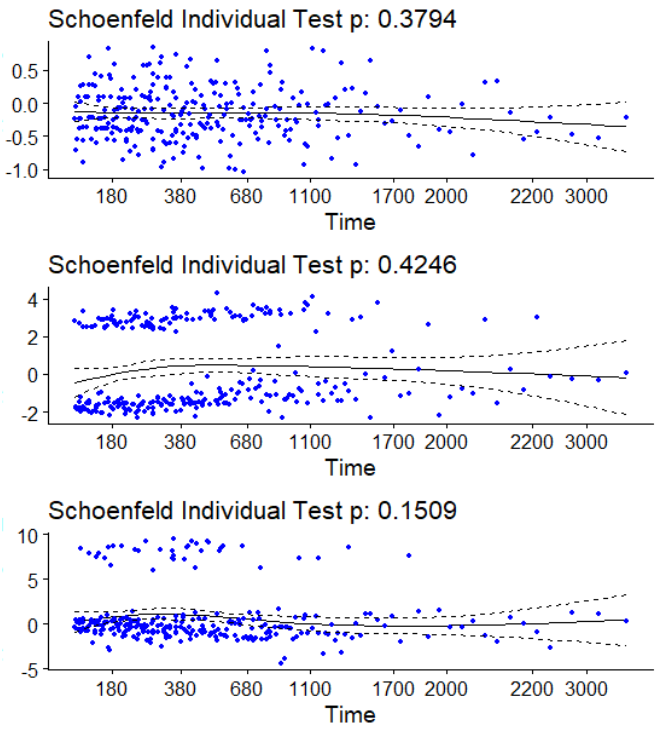


Figura 3. Residuales de Schoenfeld para Edad, Sexo y Región_S1

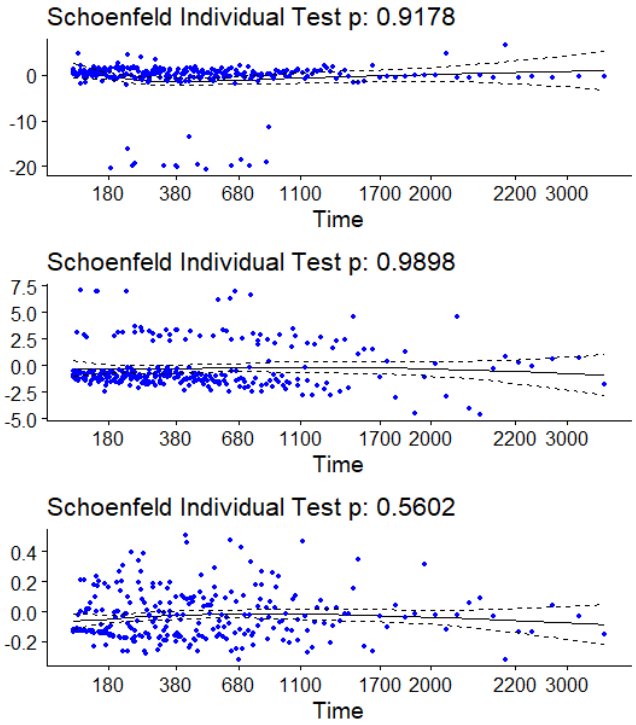


Figura 4. Residuales de Schoenfeld para Age_Leader, Q_children_Number y FTE_1

A continuación, se procedió a detectar si existen puntos influyentes en el modelo a través del análisis de residuos dfbeta, los resultados se presentan en la figura 5. Para identificar si hay algunos puntos influyentes se calcula $2/\sqrt{n}$ donde n es la cantidad de colaboradores que es igual a 1023, entonces $2/\sqrt{n}$ =0.06. Los puntos influyentes serán los |dfbeta|>0.06 que se encuentran en cada una de las 6 variables.

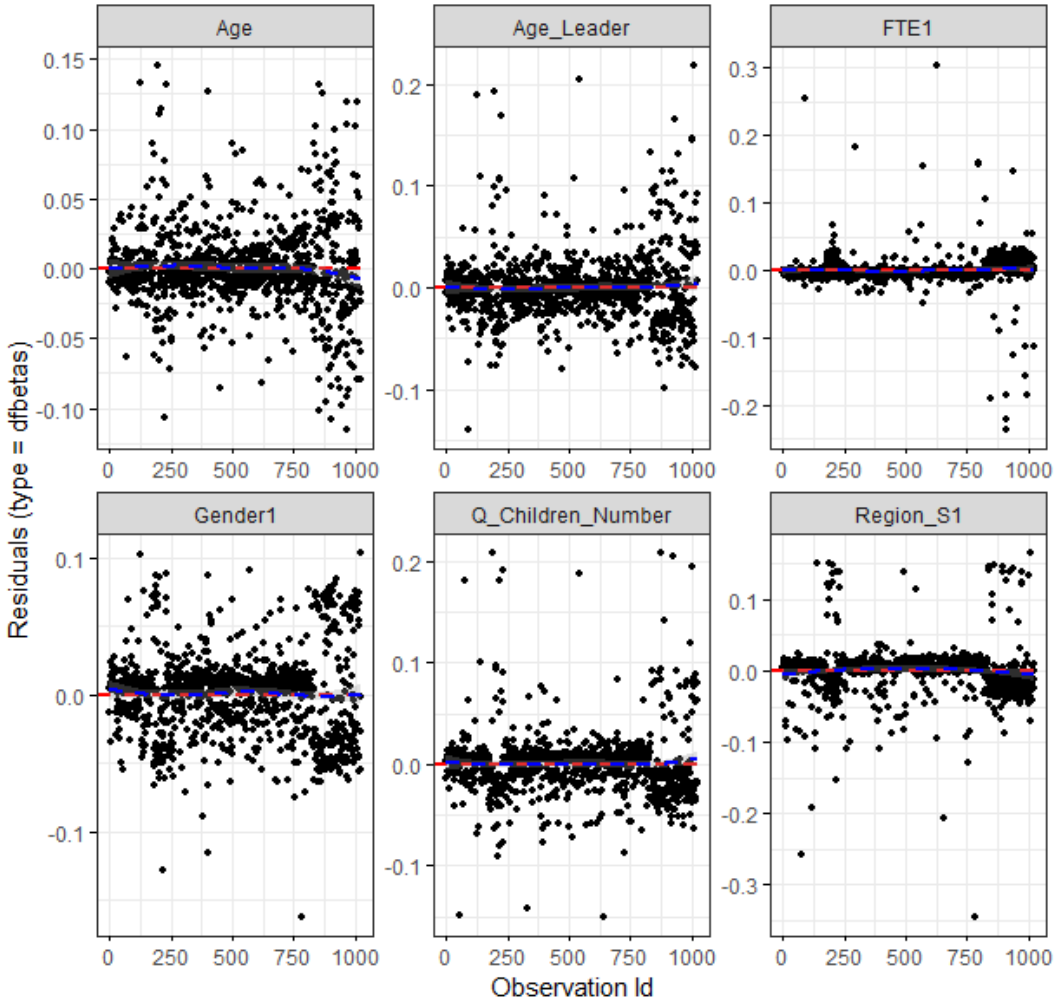


Figura 5. Residuales tipo dfbetas para las variables del modelo final.

Por último, se identifican los peores valores predecidos del modelo final estratificado a través de la figura 5 que son los residuales de tipo deviance versus las observaciones ordenadas.

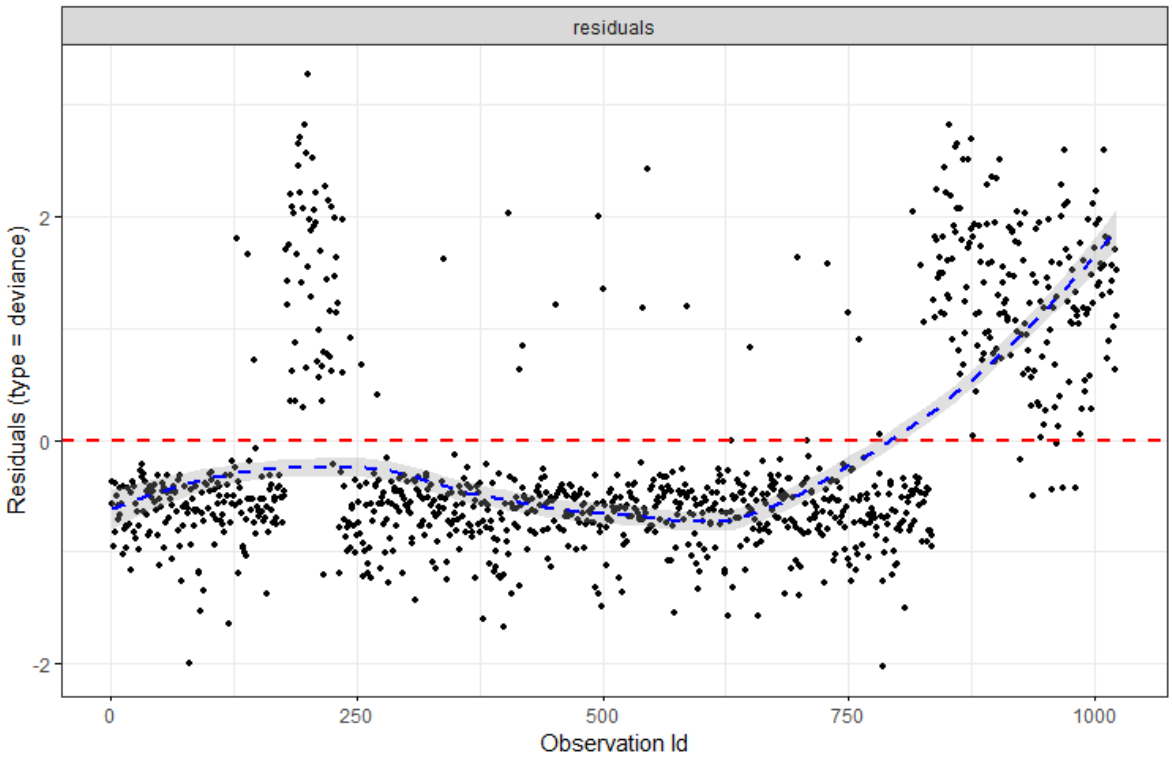


Figura 6. Residuales tipo deviance para las variables del modelo estratificado de Cox
Fuente de elaboración propia

Los residuales deviance sirven para identificar los perores valores predecidos. Los valores predecidos son aquellos que el modelo aproxima para cada uno de los individuos. Los residuales deviance que son mayores a 2 o menores a -2 son los peores valores predecidos. Se observa que solo hay 2 valores menores a -2 y más de 40 valores mayores a 2 que serían los peores valores predecidos del modelo.

A continuación, se interpreta cada uno de los coeficientes del modelo final estratificado de Cox:

i. Para la variable Age (Edad): en este caso $e^{\beta_1} = 0.857$ se interpreta de la siguiente manera: Cuando la edad se incrementa en un año, se estima que el riesgo de renuncia de un colaborador disminuye 14.3%, manteniendo constante las otras variables.

ii. Para la variable Gender (Sexo): $e^{\beta_2} = 1.267$ se estima que el riesgo de renuncia de un colaborador es 1.267 veces el riesgo de renuncia que la de una colaboradora, manteniendo constante las otras variables.

iii. Para la variable indicadora pertenencia Región Sur: $e^{\beta_3} = 1.901$ se estima que el riesgo de renuncia de un colaborador de la región sur del Perú es 1.901 veces el riesgo de renuncia de un colaborador que no pertenece a la región sur.

iv. Para la variable Q_Children_Number (Número de hijos del colaborador): $e^{\beta_4} = 0.746$ cuando el número de hijos de un colaborador se incrementa una unidad, se estima que el riesgo de renuncia de un colaborador disminuye 25.4%.

v. Para la variable Age Leader (Edad del líder): $e^{\beta_6} = 0.967$ cuando la edad del líder se incrementa una unidad, se estima que el riesgo de renuncia de un colaborador disminuye 3.3%.

vi. Para la variable FTE (Tipo de jornada del colaborador): $e^{\beta_6} = 0.513$ Se estima que el ri esgo de renuncia de un colaborador a tiempo completo es 0.513 veces el riesgo de renu ncia de un colaborador que no tiene contrato a tiempo completo.

4.3. Contribución en la solución de situaciones problemáticas

A partir de la investigación desarrollada se implementó un nuevo esquema de pesos de valoración del perfil en la etapa de selección del personal para estas posiciones, generando eficiencias en el proceso de reclutamiento y selección en tiempo y costes e impactando en la disminución de la rotación del personal de las posiciones masivas comerciales en un 20% en promedio. A partir de este estudio se ha propuesto implementar el análisis para otras posiciones masivas comerciales, pero que tienen otro tipo de estructura y otra dinámica respecto a la organización y al perfil de las personas en este rubro.

4.4. Análisis de la contribución en términos de competencias y habilidades

La presente investigación ha permitido conocer con mayor profundidad el dinamismo del negocio, respecto a la gestión del personal de una de las posiciones comerciales más importantes de la organización, el cual concentra un importante interés debido al costo operativo y al ingreso que implica su adecuada gestión. También ha permitido un mayor conocimiento de la técnica de regresión de Cox, la validación de sus supuestos, su interpretación e ir un paso más allá cuando estos no se cumplen, como es la aplicación de la regresión de Cox estratificado, que es

una de sus extensiones y que resultan muy prácticas y útiles, así también ha permitido descubrir nuevas librerías para el análisis de datos con el software R.

4.5. Nivel de beneficio obtenido por el centro laboral

Esta y otras medidas adoptadas tanto por el negocio, como por el proceso de reclutamiento y selección han permitido una reducción de 30% promedio mensual en el cese voluntario del personal para esta posición masiva comercial, cabe destacar que el conocimiento de las variables que se desprenden a partir del análisis fueron adoptados por otras iniciativas que han permitido mejoras de cara a mantener controlado los indicadores de rotación y otros indicadores comerciales cuyo objeto de estudio no son parte del presente trabajo pero podrían ser considerados también en un posterior trabajo de investigación.

El beneficio ha sido significativo no solo en términos de la reducción de la rotación en este tipo de posiciones masivas comerciales, sino también en la optimización de procesos generados por extraer información de la base de datos de personal, así como en despertar la curiosidad a los principales directivos por los beneficios de la aplicación de técnicas estadísticas a variables demográficas del empleado.

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones:

Las conclusiones del presente trabajo son:

1. Usando la regresión de Cox estratificado sobre 24 covariables demográficas pertenecientes al seguimiento de 1023 colaboradores de las principales posiciones masivas de una institución financiera, solo 6 variables tienen impacto en la probabilidad de riesgo de renuncia siendo estas su relación con el riesgo de renuncia como se indica en la tabla 9 y ordenadas de mayor a menor impacto.

Tabla 8: Resumen de los efectos de las variables sobre el riesgo de renuncia.
Fuente: elaboración propia

Estado de la variable	Riesgo de Renuncia
Si el colaborador pertenece a la región Sur del Perú	Aumenta respecto al contraste
Si el colaborador tiene jornada a tiempo completo	Disminuye respecto al contraste
Si el género del Colaborador es hombre	Aumenta respecto al contraste
A más número de hijos de un colaborador	Disminuye
A más Edad del Colaborador	Aumenta
A más edad del líder	Disminuye

2. Para llegar al modelo final se tuvo que hacer uso del modelo de regresión estratificado de Cox, convirtiendo a estrato las variables Edd_Category, Leader_Gender, Cat_Leader_Edd, lo que significa que estas variables no cumplen con el supuesto de riesgos proporcionales, es decir son variables dependientes de tiempo y podrían agregar mucha información al análisis, pero para ser incluidas en el modelo tendría que hacerse uso de un modelo extendido de Cox para covariables dependientes del tiempo.

3. Teniendo identificado el modelo de Cox estratificado se procedió a validar uno a uno los supuestos de riesgos proporcionales tanto a través de la prueba de hipótesis de riesgos proporcionales, así como por los métodos gráficos de residuales de Schoenfeld, dfbetas, y deviance, siendo satisfactorio el cumplimiento de los supuestos de riesgos proporcionales.

5.2. Recomendaciones:

Se recomienda en próximos estudios adicionar otros tipos de variables que no fueron incluidas en el presente estudio por su disponibilidad. Por ejemplo, se pueden considerar variables salariales, como el sueldo básico, el promedio de los últimos 6 meses de incentivos, el promedio de los últimos 6 meses de comisiones, entre otras. También podemos considerar variables del tipo de riesgo crediticio, variables salud laboral, así como de consumos de tarjeta de crédito o de débito, variables geográficas, entre otras variables relacionadas al riesgo de renuncia del colaborador o colaboradora. Si bien en este estudio se usó una variable de tipo geográfica (Relation_location), esta no resultó ser no significativa, tal vez la naturaleza descentralizada de las sedes de trabajo de los colaboradores no constituye un factor que lleve al colaborador al riesgo de renuncia. Considerar la posibilidad de que las variables que se deseen añadir tal vez no cumplan con el supuesto de riesgos proporcionales, como lo fue en nuestro caso para las

variables Cat_Leader_Edd1, Leader_Gender1 y EDD_Category1. Pero si en caso se desee conocer su efecto, se recomienda usar el modelo extendido de Cox que incluye variables dependientes del tiempo.

Si bien en la revisión de la literatura del presente trabajo revisamos el trabajo de (Quispe Millones, 2014) que en lugar de usar la regresión de Cox aplicó un modelo de regresión logística y logística multinivel, también debemos considerar que existen técnicas estadísticas computacionales que si bien no aportan con una interpretación del efecto de las covariables complementan los resultados a través de predicciones sobre la decisión de renuncia del colaborador, como es el caso de las redes neuronales, las máquinas de soporte vectorial entre otras técnicas de la familia de las máquinas de aprendizaje.

6. REFERENCIAS BIBLIOGRÁFICAS

Avila Eyzaguirre, S. L., Guerra Del Carpio, R. F., & Mendoza Castro, K. R. (2017). La rotación laboral no deseada: causas y consecuencias en organizaciones empresariales. Anáisis de una empresa peruana de consumo masivo. *Tesis PUCP*.

Borges, R. E. (2005). Análisis de supervivencia de pacientes con diálisis peritoneal. *Revista Colombiana de Estadística*.

Chiavenato, I. (2007). *Administración de recursos humanos.* Mexico: McGraw-Hill Interamericana.

Collet, D. (1994). *Modelling Survival Data in Medical Research.* U.S.A: Springer.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society*.

Domínguez Olaya, M. K. (2015). Analisis de las causas de rotación de personal . *Tesis Universidad de Medellín*.

Enrico Antonio Colosimo, S. R. (2006). *Análise de Sobrevivência.* Brasil: ABE-Projeto.

Flores Flores, C. J. (31 de Enero de 2011). *Modelo de regresión de Cox con métodos flexible en pacientes con Linfoma No Hodgkin*. Cataluña, España. Obtenido de https://upcommons.upc.edu/bitstream/handle/2099.1/14538/Memoria.pdf?sequence=4&isAllowed=y

Harrel, F. E., & Lee, K. L. (1986). Verifying assumptions of the Cox proportional hazards model. *Proceedings of the Eleventh Annual SAWS User's Group International Conference* (págs. 823-828). Cary, North Carolina: SAS Institue, Inc.

Maurizio, R. (2017). La rotación laboral en América Latina. *Instituto Interdisciplinario de Economía Política*.

MTPE. (2016). Encuesta Nacional de Variación Mensual del Empleo. *Informe Estadístico Mensual*.

Quispe Millones, S. M. (2014). Rotación de personal: Predicción con modelo. *Tesis, Universidad Nacional Mayor de San Marcos*, 16.

R Core Team. (December de 2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*.

R Core Team. (2018). Obtenido de R: A language and environment for statistical computing. R Foundation for Statistical Computing: https://www.R-project.org/

Robbins, S., & Judge, T. (2013). *Comportamiento Organizacional.* Mexico: Pearson.

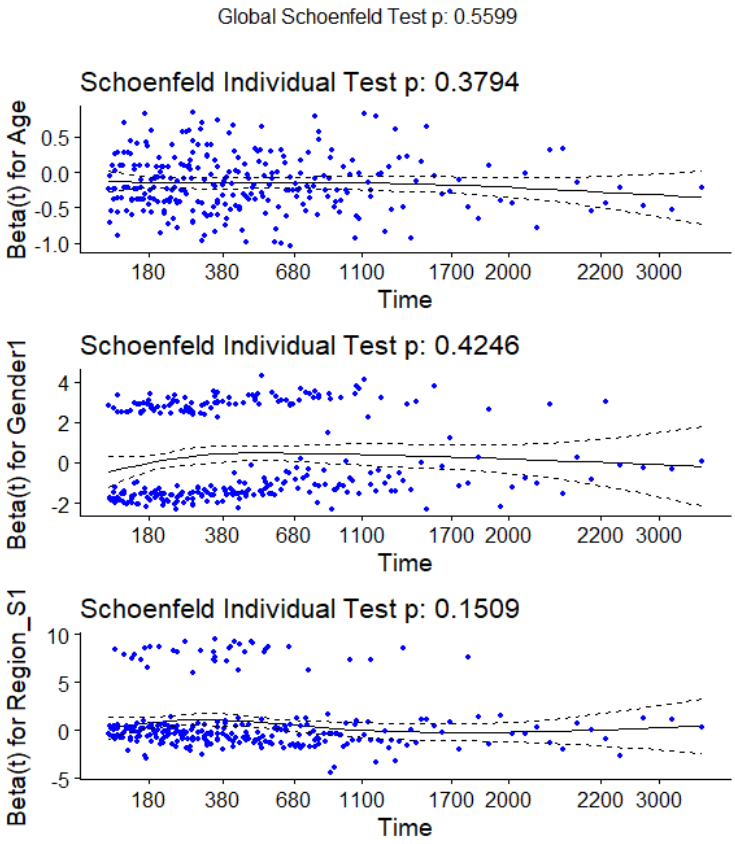
Roque, D. O. (2009). Forma funcional de covariables en el modelo de Cox . *Tesis Universidad Nacional Mayor de San Marcos*, 79-95.

Therneau, T., Grambsch, P., & Felimng, T. (1990). Martingale-based residuals for survival models. *Biometrika*, 147–160.

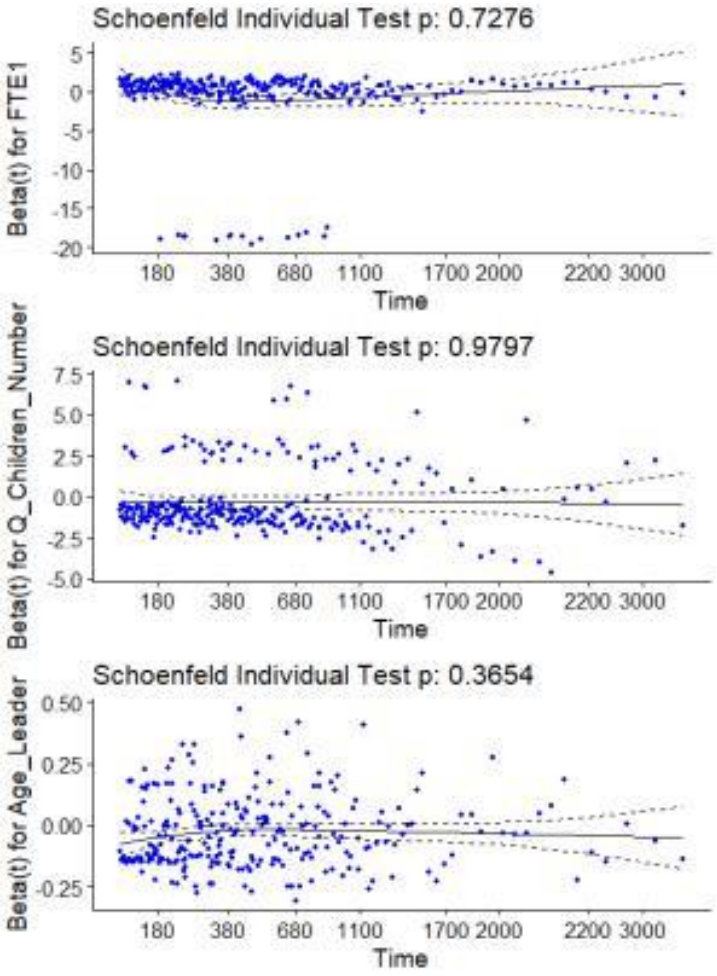
Velasco Álvarez, P. (2016). Modelo de Regresión de Cox y sus aplicaciones biosanitarias. *Tesis*.

6.1. ANEXOS

Anexo 1: Residuales de Schoenfeld para Age, Gender1 y Region_S1



Anexo 2: : Residuales de Schoenfeld para FTE1, Q-Children_Number y Age_Leader



Anexo 3: Verificación de los supuestos del modelo de Cox, desarrollado por (Borges, 2005)