

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Análisis de Frecuencia en Tarificación de Seguros Bolivar

INFORME DE PASANTIA PARA OPTAR POR EL TÍTULO DE MATEMÁTICO
PROYECTO CURRICULAR DE MATEMÁTICAS

Michael Sebastián Diaz Silva
Dirigido por: Luís Fernando Villarraga Poveda
Profesional Designado: Carlos Andres Castillo Ramirez

Bogotá DC
Junio de 2024

Resumen

Este trabajo de grado fue hecho mientras realizaba pasantías en la empresa Seguros Bolívar, una entidad líder en el sector asegurador que forma parte del conglomerado empresarial llamado Grupo Bolívar. Durante mi estadía en Seguros Bolívar, tuve la oportunidad de sumergirme en el área de seguros patrimoniales, donde pude adquirir un profundo conocimiento sobre el funcionamiento y las necesidades del sector de seguros no vida, con un enfoque particular en los seguros de hogar.

El objetivo principal de este trabajo de grado es desarrollar un modelo predictivo para los seguros de vivienda. En un contexto empresarial cada vez más orientado hacia el análisis de datos y la toma de decisiones basada en evidencia, la capacidad de prever eventos o tendencias se ha vuelto fundamental. Este modelo se diseñará para predecir el conteo de reclamaciones de una póliza dada, teniendo en cuenta una variedad de factores como el tipo de vivienda, el costo del inmueble, entre otros. Al prever con mayor precisión las reclamaciones, la aseguradora puede optimizar sus procesos de tarificación, ofreciendo así una cobertura más personalizada y competitiva a sus asegurados. Además, este modelo contribuye a la gestión proactiva de riesgos, permitiendo a la empresa anticiparse a posibles eventos adversos y tomar medidas preventivas para mitigar su impacto.

En resumen, este trabajo no solo representa un esfuerzo académico, sino también una contribución tangible al sector asegurador, facilitando la toma de decisiones informadas y mejorando la calidad de los servicios ofrecidos por Seguros Bolívar.

Palabras clave: Seguros, Modelo, Riesgo, Base de datos, Regresión Poisson.

Agradecimientos

Quiero expresar mi profunda gratitud a mis padres, familiares y amigos por su constante aliento, comprensión y amor incondicional. Su apoyo moral y emocional fueron mi mayor fortaleza durante los momentos desafiantes de este proyecto.

Agradezco sinceramente a mi director de tesis, Luis Fernando Villarraga Poveda, por su orientación experta, paciencia y apoyo incondicional a lo largo de este proceso de investigación. Desde el inicio, su compromiso con mi crecimiento académico fue evidente. Recuerdo las numerosas reuniones en las que discutimos el enfoque de la investigación y las estrategias para abordar los desafíos metodológicos. Su sabiduría y consejos fueron fundamentales para el éxito de este trabajo. Además, su capacidad para motivarme incluso en los momentos más difíciles fue invaluable.

Por último, agradezco sinceramente a mis profesores y compañeros de clase por su inspiración, estímulo y valiosas contribuciones a mi formación académica. Las discusiones en el aula, los comentarios sobre mis avances y las sugerencias para mejorar mi trabajo fueron esenciales para mi crecimiento como estudiante. Cada interacción con ellos amplió mi perspectiva y me ayudó a fortalecer mis habilidades. Su apoyo constante y su compromiso con la excelencia académica me impulsaron a esforzarme aún más en este proyecto.

Índice

1. Objetivos **4**

1.1. Objetivo General: 4

1.2. Objetivos Específicos: 4

2. Marco Teórico **4**

2.1. Grupo Bolivar 4

2.2. Probabilidad y estadística 6

2.2.1. Distribución Poisson 6

2.2.2. Modelos lineales y Modelos lineales Generalizados 7

2.2.3. Regresión Poisson 8

2.2.4. Maxima Verosimilitud 8

2.2.5. Métricas de rendimiento 9

3. Creación y desarrollo del modelo **11**

3.1. Base de datos 11

3.2. Análisis exploratorio y limpieza de datos 13

3.3. Regresión Poisson con R 18

4. Conclusiones **24**

5. Recomendaciones **25**

5.1. A la empresa: 25

5.2. A la universidad: 25

1. Objetivos

1.1. Objetivo General:

- Desarrollar un modelo predictivo mediante regresión Poisson para estimar el número de reclamaciones de pólizas de seguros de vivienda con el fin de fortalecer la comprensión teórica de la modelización de datos en el ámbito académico.

1.2. Objetivos Específicos:

- Explorar y aplicar los principios matemáticos de la regresión Poisson en el contexto de la modelización de datos, profundizando en la comprensión de la función de verosimilitud y su maximización para la estimación de parámetros.
- Emplear el lenguaje de programación R para realizar la limpieza de datos, selección de variables y creación del modelo de regresión Poisson.

2. Marco Teórico

2.1. Grupo Bolivar

El Grupo Bolívar es un grupo empresarial con un enfoque multilatino con más de 70 años de existencia, conformado por empresas de varios sectores: sector financiero, sector asegurador, sector de construcción, entre otros. El Grupo Bolívar es una sociedad que se dedica a la inversión de sus fondos o disponibilidades en la conformación, administración y manejo de un portafolio de inversiones propias, constituido por acciones, cuotas sociales y parte de interés en sociedades nacionales o extranjeras [1].

El Grupo Bolívar inició sus actividades con la constitución de la Compañía de Seguros Bolívar S.A. el de diciembre del 1939. En sus inicios la Compañía de Seguros Bolívar se dedicó a atender los seguros colectivos o de grupo de vida y los seguros de vida individual, incursionando en un mercado conformado principalmente por aseguradoras internacionales [2].

Actualmente Seguros Bolívar ofrece una variedad de servicios como los son los seguros de vida, salud, automóviles, hogar, mascotas y Pymes siempre teniendo en cuenta su propósito principal “Enriquecer la vida con integridad ” y al mismo tiempo creando una cultura de prevención mediante sus seguros.

Los siguientes conceptos básicos de seguros son tomados de [3] y son relevantes para el presente trabajo

- **Agente de seguros**: Persona física o moral independiente que desempeña una labor de intermediación entre las compañías aseguradoras y los consumidores.
- **Asegurado**: Es la persona natural o jurídica a quien el siniestro va a afectarle directamente. Es aquel sobre cuya cabeza o bienes van a recaer las consecuencias del siniestro. La figura del asegurado es esencial dentro del contrato de seguro. Porque no cabe concebir un contrato sin la existencia de un riesgo que cubrir o tampoco pensar en un negocio jurídico sin que haya una persona o destinatario final de la garantía que se pacta.
- **Asegurador**: El asegurador es la persona jurídica que asume el riesgo. Está debidamente autorizado conforme a las leyes y reglamentos de una nación.
- **Prima**: La prima es el costo del seguro que establece una compañía de seguros para el pago del asegurado. Esta es calculada teniendo en cuenta la frecuencia y severidad en la ocurrencia de eventos similares, la historia misma de eventos ocurridos al cliente, y los gastos internos o externos que tenga dicha aseguradora.
- **Póliza**: La póliza es el instrumento escrito en el cual constan las condiciones del contrato. La póliza es el documento principal del contrato de seguro, en donde constan los derechos y obligaciones de las partes, es un documento privado redactado en varios folios.
- **Riesgo Asegurable**: En el lenguaje de los seguros se utiliza el término riesgo para expresar indistintamente dos ideas diferentes. De un lado, riesgo como el objeto asegurado; de otro lado, riesgo como posible ocurrencia por azar de un acontecimiento que produce una necesidad económica y cuya aparición real o existencia se previene y garantiza en la póliza.
- **Seguro**: Es un contrato por el cual una persona natural o jurídica adquiere protección y/o asistencia de una aseguradora en caso de un siniestro u eventualidad. Mediante este contrato, tanto el asegurado como la compañía aseguradora están obligados, mediante el cobro de una prima, a resarcir un daño o a pagar una suma de dinero al verificarse la eventualidad prevista en el contrato. El contrato de seguro puede tener por objeto toda clase de riesgos si existe interés asegurable, salvo prohibición expresa de la ley. El tomador del seguro se obliga a efectuar el pago de esa prima, a cambio de la cobertura otorgada por el asegurador, la cual le evita afrontar un perjuicio económico mayor, en caso de que el siniestro se produzca. El contrato de seguro es consensual; los derechos y obligaciones recíprocos de asegurador y tomador, empiezan desde que se firma el contrato, aun antes de emitirse la «póliza» o documento que refleja datos y condiciones del contrato de seguro. Al realizar un contrato de seguro, se intenta obtener una protección económica de bienes o personas que pudieran en un futuro sufrir daños.
- **Siniestro**: Es la manifestación concreta o realización del riesgo asegurado, que produce unos daños amparados en la póliza hasta determinada cuantía.

2.2. Probabilidad y estadística

En esta sección se va a mostrar algunos conceptos de teoría de probabilidad, donde solo se hablará de variables aleatorias discretas, ya que el interés es la distribución Poisson. Por otro lado, se trabajará con vectores aleatorios, ya que, estos proporcionaran claridad, tanto en terminología, como en simbología, en el momento de hablar sobre regresiones lineales.

2.2.1. Distribución Poisson

La distribución Poisson se usa generalmente para representar datos de conteo, es decir, la frecuencia de un determinado suceso. Por lo tanto, toma valores positivos, incluido el cero [4]. Lo cual ayuda a la hora de estimar cuantos siniestros puede ocurrir con un cliente dado en un periodo de tiempo y características dadas [5].

Definición 1. Se dice que una variable aleatoria X tiene distribución Poisson con parámetro $\lambda > 0$, si su función de densidad está dada por

$$f(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{si} \quad x = 0, 1, \cdots \\ 0 & \text{en} \quad \text{otro caso} \end{cases}$$

De lo anterior se denota como $X \sim P(\lambda)$.

Las características principales que tiene la distribución Poisson son las siguientes

Teorema 1. *Sea X una variable aleatoria de parámetro λ. Entonces:*

(1) $E(X) = \lambda$

(2) $Var(X) = \lambda$

(3) $m_X(t) = e^{\lambda(e^t - 1)}$

Ejemplo: El número de personas que toma Seguros en la ciudad de Pereira en un día tiene una distribución Poisson de media 10. ¿A qué es igual la probabilidad de que, en un día en particular, el número de clientes afiliados en Bogotá sea menor o igual a 3?

Para solucionar esto, se dice que X ="número de personas que toma Seguros en la ciudad de Pereira en un día". Por otro lado, se sabe que $X \sim P(10)$ por lo tanto

$$\begin{aligned} P(X \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\ &= e^{-10} + 10e^{-10} + 50e^{-10} + \frac{1000}{6}e^{-10} \\ &= 1{,}03336 \times 10^{-2} \end{aligned}$$

2.2.2. Modelos lineales y Modelos lineales Generalizados

La regresión es un concepto estadístico que se refiere a la relación entre una o más variables independientes (también conocidas como variables predictoras) y una o más variables dependientes. El objetivo principal de la regresión es modelar y analizar la relación ente estas variables.

Hay muchos tipos de regresiones, el más común y famoso es la regresión lineal, la cual se cubre en la siguiente sección.

Se hará una breve explicación de que es un Modelo lineal:

Definición 2. Un modelo estadístico lineal que relaciona una respuesta aleatoria Y con un conjunto de variables aleatorias $x_1, x_2, \cdots, x_n$ es de la forma

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

donde $\beta_0, \beta_1, \cdots, \beta_n$ son parámetros desconocidos, ϵ es una variable aleatoria y las variables $x_1, x_2, \cdots, x_n$ toman valores conocidos. Se supone que $E(\epsilon) = 0$ y, por tanto, [6]

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

De la definición anterior, se puede decir que un modelo lineal debe verificar las siguientes tres condiciones:

1. Los errores se distribuyen normalmente.
2. La varianza es constante
3. Las variables independientes están relacionadas con la variable dependiente de manera lineal [4]

El problema es calcular $\beta_0, \beta_1, \cdots, \beta_n$, por ejemplo, cuando el modelo es $y = \beta_0 + \beta_1 x_1 + \epsilon$ entonces por lo general los estimadores de β_0 y β_1 son

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Ahora bien, puede pasar que un modelo no cumpla una o más características anteriormente mencionadas, ya sea que las variables no sean continúas, sino del tipo discreto o categorías, o cuya distribución no sea normal, dado este caso los Modelos Lineales Generalizados (GML) son una extensión de un modelo lineal.[4].

Para permitir relaciones no lineales entre la variable respuesta y las variables explicativas, se introduce la función enlace g la cual relaciona el predictor lineal y la media de la variable repuesta como sigue

$$g(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

2.2.3. Regresión Poisson

Un modelo de regresión de Poisson es un modelo lineal generalizado, en el que la función de enlace es el logaritmo de la regresión Poisson [7]. Es decir

$$log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Donde x_i representa cada variable independiente o predictora y los β_i los coeficientes o parámetros con $i = 1, \cdots, n$. Este modelo es ***lineal***, ya que cada y está escrito como una combinación lineal de los x_i.

El método mayormente utilizado para estimar al vector de coeficientes o parámetros β de un modelo Poisson es al igual que en los Modelos Lineales Generalizados el de Máxima Verosimilitud iterativo [8].

2.2.4. Maxima Verosimilitud

La máxima verosimilitud es el estimador que se obtiene maximizando la funcion de verosimilitud de la muestra, esto tomando el logaritmo de la funcion de verosimilutud, para n observaciones independientes, la función de log-verosimilud para el estimador μ de la función de probabilidad es:

$$LogL(\lambda; y) = \sum_{i=1}^{n} (y_i ln(\lambda_i) - \lambda_i - log(y_i)!)$$

Luego derivando con respecto a λ e igualando a cero

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \lambda_i) X_{ij} = 0$$

Para finalmente obtener un sistema de ecuaciones que encuentra los estimadores.

$$\sum_{i=1}^{n} \left(y_i - e^{X_i \hat{\beta}} \right) X_{ij} = 0, \quad \text{para cada } j$$

2.2.5. Métricas de rendimiento

Un aspecto muy crucial al momento de construir modelos de regresión es evaluar la calidad del ajuste de estos modelos, que ayudan a entender su comportamiento y a saber que tan eficiente es el modelo a la hora de predecir. Ahora bien, hay muchas formas de saber que tan bueno es un modelo.

El desarrollo del modelo se dividirá en 2 partes:

- **Entrenamiento**: Esta parte es donde se entrena el modelo, se recomienda utilizar el 80 % de los datos del total de datos para este uso, aunque dependiendo de la calidad y la cantidad de datos se puede variar el porcentaje.
- **Prueba**: Esta parte se conforma con los datos restantes que no fueron tomados en cuenta para la parte de entrenamiento. El fin de esto es evaluar el modelo ya entrenado.

En cierta medida, hay muchas formas de hacer métricas de rendimiento para saber si un modelos es bueno o no. Algunas de ellas son:

Métrica	Definición	Formula	Descripción
MAE	Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{m}\|O_i - P_i\|^2$	Mide tanto la falta de exactitud como la precisión en escala absoluta. Mantiene las mismas unidades que la variable de respuesta. Menos sensible a valores atípicos que el MSE o el RMSE.
R2	R-Squared	$1 - \frac{\sum_{i=1}^{m}(O_i - P_i)^2}{\sum_{i=1}^{m}(O_i - \hat{O})^2}$	Mide la fuerza de la asociación lineal entre previstos y observados. Sin embargo, mide "precisión"pero no exactitud
RSE	Relative Squared Error	$\frac{\sum_{i=1}^{m}(O_i - P_i)^2}{\sum_{i=1}^{m}(O_i - \hat{O})^2}$	Proporción de la suma total de cuadrados que corresponde a las diferencias entre predicciones y observaciones
RMSE	Root Mean Squared Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{m}(O_i - P_i)^2}$	Comprende tanto precisión como exactitud, tiene las mismas unidades que la variable de interés. Muy sensible a los valores atípicos

Cuadro 1: Métricas de rendimiento para un modelo ya entrenado, donde O_i son los datos observados, P_I los datos predecidos y $\bar{O}$ es el valor medio de los datos observados [9]

Una vez que el modelo ya ha sido entrenado, se probara con el resultado de su predicción con los datos del conjunto de prueba con las métricas presentadas en el Cuadro 1.

3. Creación y desarrollo del modelo

En esta sección se va a realizar el modelo, que va a ser una regresión Poisson, para ello se tomará una base de datos sobre seguros de vivienda, con esta, se adecuan los datos y se hace un breve análisis de estos. A continuación se creará la regresión de Poisson, para ello, se toma un conjunto de datos que se entrena y otro que se prueba. Al finalizar se tomará otra regresión similar a Poisson donde se mirará que tanto rendimiento tienen ambas.

3.1. Base de datos

Debido a normas políticas y de privacidad de la compañía, en este proyecto no se va a trabajar con una base de datos asociada a ellos. Por lo que se va a manejar una base de datos obtenida de Data in Brief [10] la cual tiene una licencia de Creative Commons que se puede compartir, copiar o modificar, siempre que se dé un apropiado crédito y se indique los cambios que fueron hechos.

El análisis y los procedimientos que se hacen con esta base de datos, son muy parecidos a lo que se hacía en la pasantía, con lo cual, no se debería tener ningún problema con trabajar datos ajenos a la empresa.

La base de datos contiene 122935 registros y 22 variables, donde se tomó un periodo de tiempo que va desde año 2010 hasta el año 2014 de una cartera de seguros de España. La base de datos contiene tanto información sobre seguros automotriz como seguros de vivienda, ya que el objetivo solo es con seguros de vivienda, se hace omisión a los datos relacionados con la parte automotriz. A continuación se tiene en cuenta las siguientes 15 variables:

- **(PollID)**: Es un número único, es el cual identifica cada póliza.
- **(gender)**: Es el género del cliente (1=hombre, 0=mujer).
- **(Age_client)**: La edad del cliente.
- **(Client_Seniority)**: El número de años que el cliente lleva con la compañía.
- **(year)**: El año de registrado (va del 1 al 5).
- **(num_policiesC)**: El número total de pólizas que tiene el mismo cliente en la compañía.
- **(metro_code)**: En que zona está (1=urbano o metropolitano, 0=rural).
- **(Policy_PaymentMethodH)**: Indica si la frecuencia del pago (1=anual, 0=mensual).
- **(Retention)**: Indica si la póliza fue renovada o no (1=renovada, 0=en otro caso).

- **(appartment)**: Indica si es un apartamento o no (1=apartamento, 0=en otro caso).
- **(Insuredcapital_content_re)**: El valor asegurado del contenido.
- **(Insuredcapital_continent_re)**: El valor asegurado de la construcción.
- **(Types)**: Es el tipo de reclamo si es que hay (1=no hay reclamos de vivienda ni de automóvil, 2=reclamo de automóvil, pero no de vivienda, 3=reclamo de vivienda, pero no de automóvil, 4=hay reclamos tanto de automóvil y vivienda).
- **(NClaims2)**: Es el número de reclamos en la póliza de seguros de vivienda en el año correspondiente.
- **(Claims2)**: Es la suma del costo de las reclamaciones en la póliza para el año correspondiente.

Esta base de datos, sus unidades monetarias, están expresadas en euros. Este archivo está guardado en el formato CSV ("Comma Separated Values").

El modelo y el análisis se va a desarrollar con el lenguaje de programación R, trabajar con datos utilizando R ofrece varias ventajas, la cuales algunas son:

- R es un lenguaje de programación de alto nivel y entorno computacional para realizar análisis estadístico y gráficos de alta calidad.
- R es software libre y de código abierto, en otras palabras, se puede usar, modificarlo y distribuirlo libremente. Lo cual da la ventaja de usarlo gratis.
- R tiene miles de paquetes complementarios disponibles, además de tener una amplia comunidad.
- R es multiplataforma, de manera que puede ser ejecutado en Windows, Mac o Linux [11].

Los entornos integrados de desarrollo ("Integrated Development Enviroment" o mejor conocidos como IDEs) tiene como objetivo es la de facilitar la realización de proyectos. En este caso se va a usar RStudio, ya que cuenta con una licencia gratuita y además ayuda a que sea más manejable el código, así como las visualizaciones a diferencia de la interfaz que tiene por estándar R.

3.2. Análisis exploratorio y limpieza de datos

La limpieza de datos también conocida como *DataCleansing* o *DataScrubbing* engloba varios procesos destinados a mejorar la calidad de los datos. Estos procesos se utilizan para corregir o eliminar registros inexactos en una base de datos o conjunto de datos. En general, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos o irrelevantes[12]. Para realizar esta tarea existen varios items a tener en cuenta:

- **Eliminación de valores duplicados**: Identificar y eliminar registros duplicados en el conjunto de datos.
- **Manejo de valores perdidos**: Tratar con valores faltantes en el conjunto de datos, ya sea eliminándolos, imputándolos con valores estimados, o usando técnicas más avanzadas como la imputación multivariable.
- **Detección y eliminación de valores atípicos**: Identificar y manejar valores atípicos que puedan distorsionar el análisis.
- **Corrección de errores de formato**: Asegurarse de que los datos estén en el formato correcto, como fechas en el formato adecuado, números en el formato correcto, etc.
- **Eliminación de información irrelevante o redundante**: Identificar y eliminar características que no contribuyan significativamente al análisis o que estén altamente correlacionadas con otras características.
- **Verificación de la consistencia de datos**: Revisar la coherencia de los datos en diferentes campos para asegurar su integridad.

Teniendo esto en cuenta podemos empezar. En primer lugar, se leeran los datos utilizando el comando `read.csv` el cual permite leer este tipo de archivos

```
library(dplyr)
df<-read.csv(file="C:\\Users\\Downloads\\data_ex.csv",
         header=TRUE,
         sep=",")
df1<-select(df,"Columnas citadas anteriormente")
```

este comando contiene varios argumentos, en este caso solo se toman 3, `file` es la ubicación del archivo, `header` indica si el archivo tiene cabeceras, en este momento el archivo si tiene cabeceras y por último `sep` indica como están separados los archivos, en este caso, es por comas.

Debido a que se quiere seleccionar las variables de interes, se creara un nuevo dataframe, donde se tenga solo esas columnas. por lo que se carga la librería *dplyr*, el cual tiene el método `select`, que permite fácilmente seleccionar las columnas que se necesitan.

Iniciamos con laexploración de los datos, para eso, tenemos los siguientes dos comandos `str` y `summary`.

```
str(df1)
df1$gender<-as.factor(df1$gender)
df1$year<-as.factor(df1$year)
df1$metro_code<-as.factor(df1$metro_code)
df1$Policy_PaymentMethodH<-as.factor(df1$Policy_PaymentMethodH)
df1$Retention<-as.factor(df1$Retention)
df1$appartment<-as.factor(df1$appartment)
df1$Types<-as.factor(df1$Types)
summary(df1)
```

El método `str` da una consulta acerca de la estructura de la base de dato:

- Las variables como PolID, gender, Age_client, num_policiesC, metro_code, Retention, apparment, policy_PaymentMethodH, y Types son datos que toman valores numéricos del tipo entero (*integer*).
- Las variables Client_Seniority, year, Insuredcapital_content_re, Insuredcapital_continent_re, NClaims2, Claims2 son datos que toman valores numéricos del tipo num (*float*).

Cabe notar que a pesar de que todas las variables son numéricas, no se operará con ellas, ya que, por ejemplo, *apparment*, solo nos indica si la vivienda es un apartamento o no, con lo cual conviene "transformar" el dato a uno categórico. En R para forzar que un tipo de dato sea otro tipo de dato, es lo que se conoce como coerción, el cual se usa el comando `as.factor` que es con lo que se transforma un tipo de dato a categórico.[13]

Una vez que ya se hizo la coerción, al final `summary` da una información breve sobre que tipo de dato es cada campo, por ejemplo:

- Cuando las variables que son numéricas, como por ejemplo, Claims o Age_client, muestra un pequeño resumen estadístico de estos campos, como es el mínimo, el máximo, la media, mediana, etc.

- Cuando las variables son de tipo categóricas, como por ejemplo apparment, en este caso solo cuenta cuantos son de cada nivel.

Cabe notar que el método `summary` dice cuantos valores son "NA" (valores perdidos), en este caso no arrojo ninguno, con lo cual para confirmar que efectivamente no hay NA, podemos utilizar los dos comandos.

La primera línea, el método `is.na` verifica si un dato es NA o no, dado un arrojando `TRUE` si lo es o `FALSE` en caso contrario, al final se suman todos los NA del todo el dataframe, mientras que la segunda línea utiliza `apply`, esta función aplica funciones a un dataframe, en este caso, verifica que dato es NA y luego lo suma, a diferencia de la primera línea, este método lo hace por columnas. El primer método es sencillo, pero el segundo, es más específico y se da la idea de qué columna es la que tiene datos faltantes. En este caso, ambas líneas dio 0, con lo cual, no es necesario tratar con NA.

Por último, se va a eliminar los datos donde el campo *Types* es igual a 2, ya que este representa solo los reclamos de automóviles, el cual no es necesario para el propósito de este texto.

```
table(df1$Types)
df1<-droplevels(df1[!df1$Types=='2',])
```

El comando `table()` permite crear rápidamente una tabla de frecuencias, donde se muestra que hay 1926 registros del tipo 2, que representan el 1.6% de los datos, lo cual no indica una cantidad considerable para el número de datos que se tendría en cuenta, por lo tanto, se va a usar el comando `droplevels()` que permite eliminar datos que son del tipo factores.

Se puede decir que los datos están listos, por lo que conviene hacer un análisis un poco más profundo para ver a detalle como se relacionan las variables. Se va a omitir la variable de *PolId*, ya que este solo se toma como un identificador.

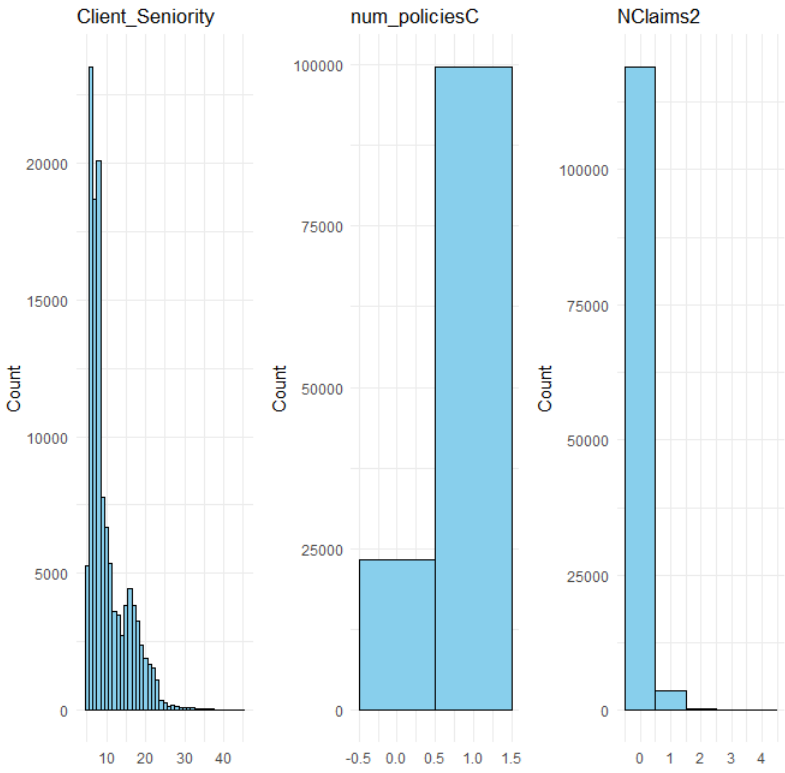


Figura 1: Histograma Variables Tipo Numérico

Al representar algunas variables en forma de histograma se pueden notar lo siguiente:

- El número de pólizas que tiene un cliente es a lo máximo 1
- Tanto el valor de asegurado del contenido como el valor asegurado de la construcción tiene una distribución similar, por lo que se puede tener cierta correlacion entre ambas variables. De forma similar están el número de reclamos con la suma de costos de reclamaciones.
- Se puede ver que la suma de reclamos la mayoría son cero, esto es muy habitual en el mundo de seguros, donde hay muchas pólizas sin reclamar.

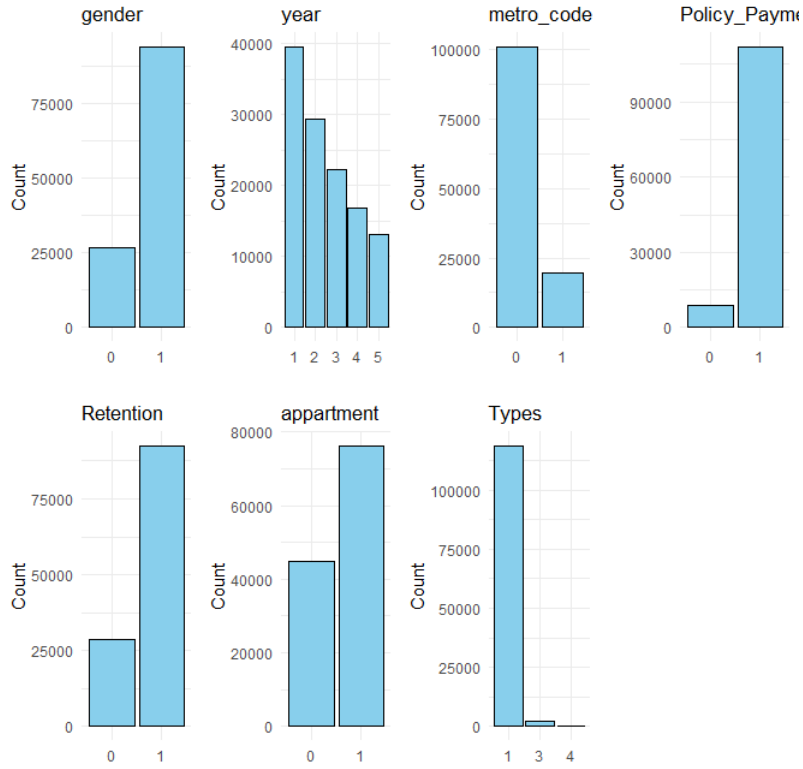


Figura 2: Histogramas Variables Tipo Categórico

Luego, al representar en digramas de caja las variables de tipo categorico se puede dar un cierto examen:

- Hay más pólizas que están aseguradas en la zona rural que en una zona urbana.
- La mayoría de pólizas son sobre apartamentos.
- Como paso en el caso anterior, hay muchas pólizas sin reclamaciones.

Se puede explorar como el paso de los años se relaciona con la suma del costo de las reclamaciones.

```
costo_anual<-df1 %>% group_by(year) %>% summarise(summary=sum(Claims2))
plot(costo_anual,xlab="año",ylab="suma de reclamaciones",
     main="Suma de Reclamaciones por año")
```

El operador %>% sirve para concatenar multiples operaciones [14], en este caso particular, primero se toma el dataframe, luego se quiere agrupar por año, utilizando el comando `group_by`, con ello se suma las cantidades agrupadas, tilizando el el comando `summarise`.

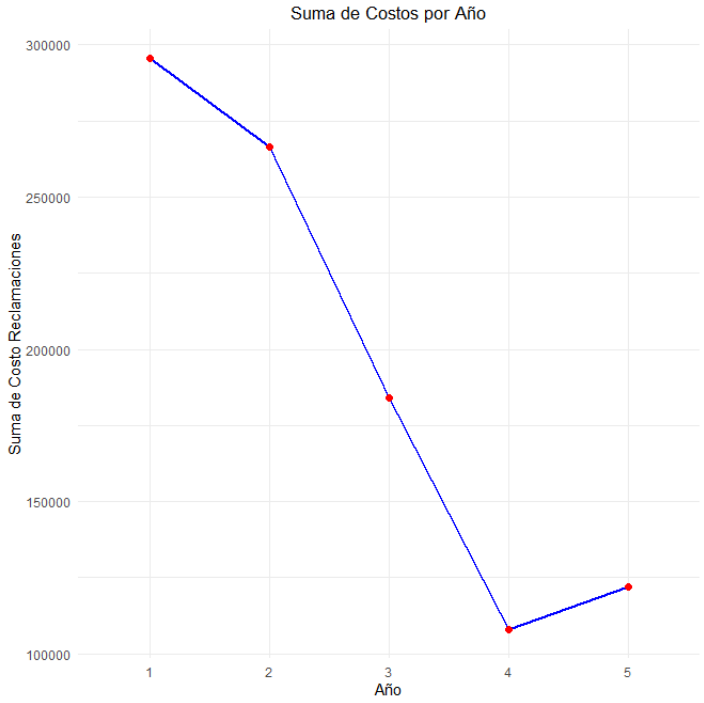


Figura 3: Histogramas Suma costos por año

Se puede notar en la Figura 3, al igual que con la variable *year*, la suma de los costos tiende a la baja. Con lo cual, se espera que al pasar los años, es menos probable que la póliza sea reclamada.

3.3. Regresión Poisson con R

Como los datos están listos para hacer la regresión. La variable objetiva es el número de reclamaciones (*NClaims2*) y las variables predictoras son las demás. Se carga el paquete *Caret* la cual integra un conjunto de funciones para la preparación de modelos predictivos, entre las cuales está `createDataPartition()` que ayuda a la separación de datos, donde se tendrá un conjunto de entrenamiento y uno de prueba.

```
library(caret)
set.seed(123)
```

```
training_ids<-createDataPartition(df1$Claims2,p=0.8,list=F)
df1_training<-df1[training_ids,]
df1_test<-df1[-training_ids,]
```

Ahora, el comando `set.seed()` configura una semilla inicial, esto es utíl a la hora de replicar un experimento, respetando la aleatoriedad. `df1_training` es el conjunto de entrenamiento, el cual tiene el 80% de los registros, mientras `df1_test` es el conjunto de prueba, el cual tiene el 20% restante de los datos

En R existe un comando que permite modelar GML's, el cual es `gml()`

```
formula="variables dependientes"~"variables independientes"
modelo_poisson<-glm(formula,df1_training,family=poisson(link="log"))
summary(modelo_poisson)
```

cuando se ejecuta el comando `summary(modelo.poisson)` este nos arroja un resumen muy sustancioso del modelo

- **Residuals**: Da un análisis descriptivo de los residuos
- **Coeffients**: Esta es una tabla de 4 columnas que nos indican
  * **Estimate**: Da los estimados de cada uno de los parámetros.
  * **Std Error**: Da la desviación estándar estimada de los estimadores de la regresión. [15]
  * **z-value**: Este valor se calcula tomando el estimador y diviendolo por el error estandar.
  * **Pr(>|t|)**: Habla de la probabilidad asociada a un valor z en particular. Cuando el valor es $< 0{,}05$, se dice que la variable es un predictor significativo en el modelo.
- **Signif. Codes**: Da los códigos para niveles de significancia (0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ')
- **Null deviance**: Dice que tan bien puede predecirse la variable objetivo en el modelo con solo un termino de intersección
- **Residual deviance**: Dice que tan bien puede predecirse la variable objetivo mediante el modelo especifico que se ajusta con p variables predictoras. Cuanto menor sea este número, mejor será el modelo al predecir la variable objetivo.

- **AIC (Akaike Information Criterion):** es una métrica que se utiliza para comparar el ajuste de diferentes modelos de regresión. Cuanto menor sea el valor, mejor podrá ajustarse el modelo de regresión a los datos. SU formula es $AIC = 2k - 2ln(L)$ donde k el numero de parametros del modelo y L es el maximo de la función de verosimilitud [16].
- **Number of Fisher Scoring iterations**: Es solo una salida detallada de mínimos cuadrados iterativos ponderados . Un gran número de iteraciones puede ser motivo de preocupación, lo que indica que el algoritmo no converge correctamente. [17]

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	$-5{,}714e+00$	$3{,}631e-01$	$-15{,}739$	$< 2e-16$	***
gender1	$-7{,}192e-02$	$4{,}112e-02$	$-1{,}749$	0,08026	.
Age_client	$-1{,}030e-06$	$1{,}458e-03$	$-0{,}001$	0,99944	
Client_Seniority	$-1{,}067e-03$	$3{,}655e-03$	$-0{,}292$	0,77034	
year2	$6{,}983e-02$	$4{,}746e-02$	1,471	0,14116	
year3	$1{,}040e-01$	$5{,}046e-02$	2,061	0,03929	*
year4	$5{,}782e-02$	$5{,}561e-02$	1,040	0,29852	
year5	$7{,}740e-02$	$6{,}201e-02$	1,248	0,21196	
num_policiesC	$-2{,}184e-02$	$4{,}558e-02$	$-0{,}574$	0,56625	
metro_code1	$2{,}184e-02$	$4{,}602e-02$	0,475	0,63504	
Policy_PaymentMethodH1	$-1{,}313e-01$	$6{,}282e-02$	$-2{,}090$	0,03660	*
Retention1	$-1{,}008e-01$	$3{,}920e-02$	$-2{,}571$	0,01014	*
appartment1	$9{,}169e-01$	$6{,}053e-02$	15,148	$< 2e-16$	***
Insuredcapital_content_re	$2{,}624e-02$	$2{,}364e-02$	1,110	0,26714	
Insuredcapital_continent_re	$6{,}807e-02$	$2{,}344e-02$	2,905	0,00368	**
Types3	$4{,}113e+00$	$3{,}793e-02$	108,430	$< 2e-16$	***
Types4	$4{,}182e+00$	$1{,}675e-01$	24,976	$< 2e-16$	***
Claims2	$3{,}452e-05$	$1{,}871e-05$	1,845	0,06499	.
Null deviance:	22905	on	96807	degrees of freedom	
Residual deviance:	11590	on	96790	degrees of freedom	
AIC:	17937				
Number of Fisher	Scoring iterations:	6			

Cuadro 2: Resultados del modelo de regresión de Poisson

Al ver los resultados en Cuadro 2, se puede notar que variables son mas significativas son el intercepto, si es apartemento o no y los tipos de reclamo, en menor medida estan, el valor del inmueble, si la poliza es renovada o no, la frecuencia con que se paga y el año que en este caso es del 2013. Por

otro lado, se pensaría que la suma de costos de los reclamos tendria mas relacion con el número de reclamos, pero vemos que no influye mucho. Otra que cabe destacar, es que si una variable factorial tiene k categorias entonces la salida de `gml()` devuelve $k-1$ categorias y la restante la toma como base, como ejemplo, no aparece *year1* pero si el resto de categorías [7].

Ahora una manera fácil para determinar si un modelo es útil, es calcular es estadistico χ^2 el cual es

$$x^2 = \text{Null deviance} - \text{Residual deviance}$$

con p grados de libertad. En este modelo, al calcular, se tiene que $\chi^2 = 22905 - 11590 = 11315$ con 17 grados de libertad. Ahora el valor que se calcula χ^2 da un número $< 0{,}05$ el cual nos sugiere que este modelo sí puede ser muy útil [16].

A continuación se manejará otro modelo, se escoge el modelo de regresión Quasi-Poisson, el cual es una generalización de la regresión de Poisson y se usa para modelar una variable de conteo con sobre-dispersión. La diferencia entre ambas es que el modelo de Poisson supone la hipotesis que la varianza y la media son iguales, en cambio en la Quasi-Poisson supone que la varianza es una función lineal de la media, es decir, $Var(X) = \phi \cdot E(X)$ [18].

Para implementar un modelo de regresión Quasi-Poisson, el comando `gml()` tiene esta opción.

```
modelo_quasi_poisson<-glm(formula,df1_training,family=quasipoisson(link="log"))
summary(modelo_quasi_poisson)
```

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	$-5{,}714e+00$	$3{,}523e-01$	$-16{,}221$	$< 2e-16$	***
gender1	$-7{,}192e-02$	$3{,}989e-02$	$-1{,}803$	0,07144	.
Age_client	$-1{,}030e-06$	$1{,}415e-03$	$-0{,}001$	0,99942	
Client_Seniority	$-1{,}067e-03$	$3{,}546e-03$	$-0{,}301$	0,76351	
year2	$6{,}983e-02$	$4{,}605e-02$	1,517	0,12939	
year3	$1{,}040e-01$	$5{,}396e-02$	2,124	0,03366	*
year4	$5{,}782e-02$	$5{,}396e-02$	1,071	0,28398	
year5	$7{,}740e-02$	$6{,}016e-02$	1,286	0,19830	
num_policiesC	$-2{,}614e-02$	$4{,}422e-02$	$-0{,}591$	0,55443	
metro_code1	$2{,}184e-02$	$4{,}465e-02$	0,489	0,62472	
Policy_PaymentMethodH1	$-1{,}313e-01$	$6{,}095e-02$	$-2{,}154$	0,03123	*
Retention1	$-1{,}008e-01$	$3{,}804e-02$	$-2{,}650$	0,00806	**
appartment1	$9{,}169e-01$	$5{,}873e-02$	15,61	$< 2e-16$	***
Insuredcapital_content_re	$2{,}624e-02$	$2{,}294e-02$	1,144	0,25278	
Insuredcapital_continent_re	$6{,}807e-02$	$2{,}274e-02$	2,993	0,00276	**
Types3	$4{,}113e+00$	$3{,}680e-02$	111,749	$< 2e-16$	***
Types4	$4{,}182e+00$	$1{,}625e-01$	25,741	$< 2e-16$	***
Claims2	$3{,}452e-05$	$1{,}815e-05$	1,902	0,05720	.
Null deviance:	22905	on	96807	degrees of freedom	
Residual deviance:	11590	on	96790	degrees of freedom	
AIC:	NA				

Cuadro 3: Resultados del modelo de regresión de Quasi-Poisson

Al ver el Cuadro 3 se nota que cambia ligeramente algunos valores, una cosa a remarcar es que la variable de sí póliza se renueva o no, tiene más significancia que en el modelo anterior.

Ahora vamos a evaluar ambos modelos con las métricas de rendimiento, para ello, se guardan las predicciones de ambos modelos en el conjunto de entrenamiento.

```
library(metrica)
predicción_modelo_poisson<-modelo_poisson %>% predict(df1_test)
predicción_modelo_Quasi_poisson<-modelo_quasi_poisson %>% predict(df1_test)

RSS(data=df1_test,obs=df1_test$NClaims2,pred=predicción_modelo_poisson)
MAE(data=df1_test,obs=df1_test$NClaims2,pred=predicción_modelo_poisson)
R2(data=df1_test,obs=df1_test$NClaims2,pred=predicción_modelo_poisson)
```

```
RSE(data=df1_test,obs=df1_test$NClaims2,pred=predicción_modelo_poisson)
RMSE(data=df1_test,obs=df1_test$NClaims2,pred=predicción_modelo_poisson)
```

La librería *metrica* tiene más de 80 funciones para medir el rendimiento de la regresiones [9]. Al hacer lo mismo con el modelo de regresion se obtiene el siguiente la siguiente tabla Cuadro 4.

Medidas de redimiento	Regresión Poisson	Regresión CuasiPoisson
MAE	4,282184	4,282184
R^2	0,3752264	0,3752264
RSE	471,8965	471,8965
RMSE	4,326421	4,326421

Cuadro 4: Rendimiento de la regresión Poisson y Quasi-Poisson

Se puede notar que ambas regresiones muestran rendimientos similares. Veamos que en las medidas de rendimiento de MAE, RMSE es muy bueno,ya que estos con valor 0 indica que el modelo tiene un ajuste perfecto y cuanto menor sea el RMSE, mejor será el modelo y sus predicciones. En cambio, para las métricas como RSE y R^2 no presentan una mejor situación, teniendo en cuenta que $R^2 < 0,5$ tenemos que presenta un ajuste debil, esto se debe a que agregar predictores al modelo puede aumentar el valor de R^2 debido al azar, lo que hace que los resultados sean engañosos. Agregar predictores al modelo puede causar un "sobreajuste", donde el modelo intenta predecir el ruido.[en] los datos. Esto disminuye su capacidad para funcionar mejor con datos 'nuevos' que no ha visto antes[19].

4. Conclusiones

Con base en el análisis realizado, se evidencia que al emplear un modelo fundamentado en la regresión de Poisson para la tarificación, determinadas variables adquieren mayor relevancia. Entre estas variables destacan el tipo de vivienda (por ejemplo, si se trata de un apartamento), el valor del inmueble a asegurar y la renovación de la póliza. Factores secundarios incluyen el género, el método de pago y el historial de reclamaciones.

En lo que respecta al desempeño del modelo, se observa una eficacia notable en la predicción. No obstante, se sugiere la exploración de otros enfoques, como la regresión binomial negativa o los árboles de decisión, combinando estos modelos con la experiencia de un experto para optimizar las decisiones tarifarias, no solo en el ámbito de vivienda, sino también en otras modalidades de pólizas.

Se plantea la utilización de modelos de regresión de Poisson para examinar la frecuencia de siniestros en un inmueble durante un período prolongado, además de la incorporación de más variables predictivas relacionadas con riesgos que no se consideraron en este estudio.

En síntesis, al comparar los resultados obtenidos por el modelo con los criterios de evaluación de Seguros Bolívar, se constata que las variables determinantes para estimar el riesgo de siniestros en un inmueble y establecer primas son mayormente coincidentes, salvo por detalles como la ubicación, el material de construcción y la antigüedad de la edificación. Esto podría atribuirse al elevado riesgo asociado a ciertas zonas del país, que puede deberse a cuestiones de seguridad, normativas urbanísticas o condiciones del suelo.

5. Recomendaciones

5.1. A la empresa:

Con respecto al programa de contratación de jóvenes estudiantes para su primera experiencia laboral, recomiendo a Seguros Bolívar mantener y fortalecer esta iniciativa. La práctica de contratar jóvenes talentos no solo les brinda la oportunidad de adquirir experiencia práctica en el mundo laboral, sino que también enriquece la empresa con nuevas perspectivas, ideas innovadoras y energía fresca. Este enfoque contribuye al desarrollo profesional y personal de los estudiantes, fomenta la diversidad y la inclusión en el entorno laboral, y puede ser una estrategia efectiva para identificar y retener talentos prometedores a largo plazo. Seguir invirtiendo en este programa beneficiará tanto a la empresa como a la comunidad estudiantil, fortaleciendo la conexión entre el ámbito académico y el sector empresarial.

5.2. A la universidad:

Dada la creciente importancia de la Computación Matemática y sus aplicaciones en diversos campos, recomiendo a la Universidad Francisco José de Caldas ampliar sus programas académicos en esta área. La integración de cursos y electivas enfocadas en Computación Matemática permitirá a los estudiantes adquirir habilidades avanzadas en análisis de datos, modelado matemático y optimización computacional, preparándolos para enfrentar los desafíos actuales y futuros en campos como la ciencia de datos, la inteligencia artificial y la ingeniería de software. Esta iniciativa contribuirá al desarrollo de profesionales altamente capacitados y competitivos en el mercado laboral actual.

Referencias

[1] "¿quiénes somos?" https://n9.cl/t21kh5, accedido 02-03-2024.

[2] "Año 1939," https://n9.cl/iswllg, accedido 02-03-2024.

[3] "Glosario de seguros," https://n9.cl/lodhg, accedido 12-02-2024.

[4] L. M. Sarmiento, "Modelos lineales generalizados: Modelos con coeficiente de variación constante y otros." 2018.

[5] L. Blanco Castañeda, *Probabilidad.* Universidad Nacional de Colombia-Facultad de Ciencias, 2013.

[6] D. D. Wackerly, W. Mendenhall, and R. L. Scheaffer, *Estadística matemática con aplicaciones.* Australia; México, DF: Cengage Learning„ 2002.

[7] "Chapter 8 regresión de poisson | modelos lienales generalizados con r," https://bookdown.org/jaimeisaacp/bookglm/regresi%C3%B3n-de-poisson.html, accedido 04-03-2024.

[8] G. T. F. Arboccó, "La fecundidad y su relación con variables socioeconómicas, demográficas y educativas aplicando el modelo de regresión poisson," https://sisbib.unmsm.edu.pe/bibvirtualdata/monografias/basic/figueroa_ag/cap3.pdf, accedido 27-05-2024.

[9] "Regression performance metrics and indices," https://cran.r-project.org/web/packages/metrica/vignettes/available_metrics_regression.html, accedido 15-04-2024.

[10] M. Guillen, C. Bolancé, E. W. Frees, and E. A. Valdez, "Case study data for joint modeling of insurance claims and lapsation," *Data in Brief*, vol. 39, p. 107639, 2021.

[11] P. C. J. Nelson, *Introducción a la ciencia de datos en R: Un enfoque práctico.* Editorial Universidad Distrital Francisco José de Caldas, 2020.

[12] "Datacleaning limpieza de datos: definición, técnicas, importancia en data science," https://datascientest.com/es/datacleaning-limpieza-de-datos-definicion-tecnicas-importancia-en-data-science, accedido 27-05-2024.

[13] "4.7 coerción | r para principiantes," https://bookdown.org/jboscomendoza/r-principiantes4/coercion.html, accedido 26-03-2024.

[14] "4.7 coerción | r para principiantes," https://rsanchezs.gitbooks.io/rprogramming/content/chapter9/pipeline.html, accedido 15-04-2024.

[15] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet, *The R software: Fundamentals of programming and statistical analysis.* Springer, 2013.

[16] "How to interpret glm output in r (with example)," https://www.statology.org/interpret-glm-output-in-r/, accedido 15-04-2024.

[17] "Interpreting generalized linear models," https://www.r-bloggers.com/2018/11/interpreting-generalized-linear-models/, accedido 15-04-2024.

[18] M. García García, "Modelos de regresión de poisson," B.S. Tesis, 2021.

[19] "Conoce las métricas de precisión más comunes para modelos de regresión," https://www.growupcr.com/post/metricas-precision, accedido 28-05-2024.

SEGUROS COMERCIALES BOLÍVAR

Bogotá, 31 de Mayo del 2024

Señores
Universidad Distrital Francisco José de Caldas
Facultad de Ciencias Matemáticas y Naturales
Programa de Matemáticas
Carrera 4 # 26 D - 31
Bogotá

Asunto: Acta de finalización de pasantía

Por medio de la presente, certifico que el estudiante **Michael Sebastian Diaz Silva**, identificado con cédula de ciudadanía **1013659606**, realizó sus prácticas profesionales en **Seguros Comerciales Bolívar S.A.,** desde el 13 de Enero del 2023 hasta el 13 de Julio del 2023 bajo mi supervisión.

Michael Sebastian Diaz Silva cumplió **satisfactoriamente** con todas las actividades asignadas, mostrando un alto nivel de compromiso y profesionalismo, resaltando sus habilidades de analítica de datos en el desarrollo de la pasantía.

Nota de Evaluación: **Cuatro punto ocho (4.8)**

Atentamente,

CARLOS ANDRES CASTILLO RAMÍREZ
Supervisor de la pasantía
JEFE LÍNEA DE NEGOCIO HOGAR
SEGUROS COMERCIALES BOLÍVAR S.A.

Seguros Comerciales Bolívar S.A.
VIGILADO SUPERINTENDENCIA FINANCIERA DE COLOMBIA