Trabajo Final

Chantres Arrieta Nikole
Corona López José Luis

3 de diciembre de 2024

1. Introducción

1.1. Base de datos

Se ha tomado una base de datos que contiene información detallada sobre los montos reclamados en relación con seguros de automóviles. Estos datos fueron recopilados de varios estados del este de Estados Unidos, proporcionando una visión amplia de las reclamaciones realizadas en dicha región. La base de datos abarca diversos aspectos de las reclamaciones, como el límite de la cobertura general, tipo de incidente, deducible de la póliza, hora del día del incidente, reclamo por vehículo y demás variables. Este conjunto de datos resulta interesante para analizar el comportamiento de las reclamaciones de seguros. A la base hemos decidido cambiar el nombre de las variables, para que al realizar el análisis en software facilite el llamado de estas.

1.2. Variables

Las variables de la base de datos son:

- **X1 (Meses como cliente):** Representa la duración de la relación entre el asegurado y la compañía de seguros, lo cual podría influir en la confianza de la aseguradora en el cliente y en las decisiones sobre el monto reclamado.
- **X2 (Edad del asegurado):** La edad podría estar relacionada con el tipo de conductores que generan siniestros, ya que algunos grupos de edad tienen más probabilidades de estar involucrados en ciertos tipos de accidentes.
- **X3 (Deducible de la póliza):** Un deducible más alto puede reducir el monto reclamado por el asegurado, ya que este debe cubrir una mayor parte del costo de los daños.
- **X4 (Prima anual de la póliza):** La cantidad que se paga anualmente por el seguro puede estar vinculada al tipo de cobertura y, por lo tanto, al monto reclamado en caso de siniestro.
- **X5 (Límite de la cobertura general):** Este factor determina el máximo que la aseguradora pagará en caso de una reclamación, lo cual podría afectar el monto final reclamado.
- **X6 (Sexo del asegurado):** Las estadísticas de siniestros pueden variar según el sexo del asegurado, por lo que esta variable podría tener un efecto sobre el monto reclamado.
- **X7 (Tipo de incidente):** Diferentes tipos de incidentes (como colisiones de un solo vehículo o robos) suelen tener diferentes costos asociados, lo que influye directamente en el monto reclamado.
- **X8 (Tipo de colisión):** El tipo de colisión (lateral, trasera o frontal) también puede afectar el costo de los daños y, por ende, el monto reclamado.
- **X9 (Gravedad del incidente):** La gravedad del daño (daños importantes, menores, pérdida total, etc.) tiene un impacto directo sobre el monto de la reclamación, ya que los daños mayores requieren reparaciones más costosas o incluso la sustitución total del vehículo.
- **X10 (Hora del día del incidente):** El momento del incidente podría influir en factores como el tráfico y las condiciones de conducción, lo cual afectaría los daños y el monto de la reclamación.

- **X11 (Número de vehículos involucrados):** Un mayor número de vehículos involucrados en el incidente puede aumentar el monto reclamado debido a los daños adicionales y la complejidad de la resolución del siniestro.
- **X12 (Lesiones corporales):** Si el accidente resulta en lesiones, esto generalmente eleva el costo total de la reclamación debido a los gastos médicos y las posibles compensaciones.
- **X13 (Testigos):** La presencia de testigos podría influir en la rapidez y claridad con la que se resuelve el reclamo, afectando potencialmente el monto total reclamado.
- **X14 (Informe de póliza disponible):** La disponibilidad de un informe de la póliza podría facilitar la verificación de la cobertura y acelerar el proceso de pago de la reclamación, lo que podría influir en el monto reclamado.
- **X15 (Reclamo por lesiones), X16 (Reclamo por daños materiales), X17 (Reclamo por vehículo):** Estas variables representan los diferentes tipos de reclamos que pueden estar involucrados, lo cual tiene un impacto directo en el monto total reclamado, dependiendo de la naturaleza del incidente.
- **X18 (Año del automóvil):** El año del vehículo podría influir en el costo de las reparaciones o en la depreciación del automóvil, afectando el monto de la reclamación.

1.3. Presentación de modelos

Los modelos que se realizaron en el software fueron los siguientes:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_{18} X18 \tag{1}$$

$$Y = \beta_0 + \beta_7 X7 + \beta_{10} X10 + \beta_{11} X11^2 + \beta_{15} X15 + \beta_{16}\sqrt{X16} + \beta_{17}\sqrt{X17} \tag{2}$$

$$Y = \beta_0 + \gamma_1\sqrt{X3X15} + \gamma_2 X4X16 + \beta_{11} log(X11) + \beta_{12} exp(X12) + \beta_{17} log(X17) \tag{3}$$

$$log(Y) = \beta_3 X3 + \beta_4 X4 + \beta_6 X6 + \beta_{15} X15 + \beta_{16} X16 \tag{4}$$

2. Desarrollo

2.1. Modelo 1

2.1.1. Hipótesis Conceptual

La hipótesis conceptual de este modelo es explorar si todas las variables disponibles en la base de datos pueden explicar adecuadamente el monto total reclamado (Y) en función de diversas características del asegurado y del incidente.

2.1.2. Proceso de selección de variables

La selección de variables se basó en la disponibilidad de información en la base de datos. Se incluyeron todas las variables que estaban disponibles y que podrían tener una relación directa o indirecta con el monto reclamado. Estas variables fueron:

- Características del asegurado (edad, sexo, tiempo como cliente, tipo de póliza, etc.).
- Características del incidente (tipo de colisión, gravedad, hora del día, número de vehículos involucrados, etc.).
- Elementos adicionales como si hubo lesiones, testigos, y la disponibilidad del informe de la póliza.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18, data = insurance_claims)

Residuals:
      Min         1Q     Median         3Q        Max
-7.310e-10 -3.000e-12  1.450e-12  5.570e-12  8.171e-11

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept) -3.079e-11  5.023e-10 -6.100e-02   0.9512
X1          -3.866e-14  3.880e-14 -9.960e-01   0.3196
X2           2.895e-13  4.786e-13  6.050e-01   0.5456
X3           4.426e-16  2.473e-15  1.790e-01   0.8581
X4          -9.523e-15  6.034e-15 -1.578e+00   0.1151
X5           1.699e-18  7.193e-19  2.362e+00   0.0185 *
X6           6.559e-12  3.015e-12  2.175e+00   0.0301 *
X7          -1.375e-12  3.920e-12 -3.510e-01   0.7259
X8           1.137e-13  1.892e-12  6.000e-02   0.9521
X9           2.144e-12  1.830e-12  1.172e+00   0.2417
X10          1.562e-13  2.177e-13  7.180e-01   0.4733
X11          2.916e-12  5.601e-12  5.210e-01   0.6028
X12          2.547e-12  1.849e-12  1.377e+00   0.1690
X13          1.148e-13  1.383e-12  8.300e-02   0.9339
X14         -2.855e-12  3.004e-12 -9.500e-01   0.3424
X15          1.000e+00  4.119e-16  2.428e+15   <2e-16 ***
X16          1.000e+00  4.291e-16  2.331e+15   <2e-16 ***
X17          1.000e+00  1.631e-16  6.129e+15   <2e-16 ***
X18         -1.331e-14  2.503e-13 -5.300e-02   0.9576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.429e-11 on 516 degrees of freedom
  (465 observations deleted due to missingness)
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.598e+30 on 18 and 516 DF,  p-value: < 2.2e-16
```

Figura 1: Resumen del modelo 1

2.1.3. Validez del modelo

El modelo consiste en:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_{18} X18 \tag{5}$$

donde las variables X_i están mencionadas en la introducción.

Es importante mencionar que el modelo da un $R^2 = 1$, lo que parecería perfecto, sin embargo, no lo es, porque es un sobre ajuste debido a la carga excesiva de variables.

Pruebas de supuestos

Para la normalidad usaremos una prueba Jarque-Bera, tomando como juego de hipótesis:

$$H_0 : Los\ errores\ se\ distribuyen\ de\ forma\ normal$$

$$H_1 : Los\ errores\ no\ se\ distribuyen\ de\ forma\ normal$$

Tomando los resultados de R se tiene un $p - value <$ 2.2e-16, como es menor que 0.05, rechaza hipótesis nula y concluimos que los errores no siguen una distribución normal
Podemos también graficar la densidad de los errores y obtenemos:

Para la autocorrelación usaremos la prueba Durbin-Watson, donde:

$$H_0 : No\ hay\ autocorrelación\ enloserrores$$

$$H_1 : Hay\ autocorrelación\ en\ los\ errores$$

El resultado devuelto en R es un $p - value <$ 2.2e-16, más pequeño que 0.05, por lo que rechazamos hipótesis nula, o sea tenemos problemas de autocorrelación, es decir los errores no son independientes.

En cuanto la heterocedasticidad, utilizamos la prueba de Breusch-Pagan, en la cual:

$$H_0 : La\ varianza\ de\ los\ errores\ es\ constante\ (homocedasticidad)$$

$$H_1 : La\ varianza\ de\ los\ errores\ no\ es\ constante\ (homocedasticidad)$$

Obtenemos un $p - value = 0.852 > 0.05$, por lo que no rechazamos H_0, así que tenemos homocedastidad. Gráficamente tenemos:

Sin embargo notamos presencia de valores atípicos.

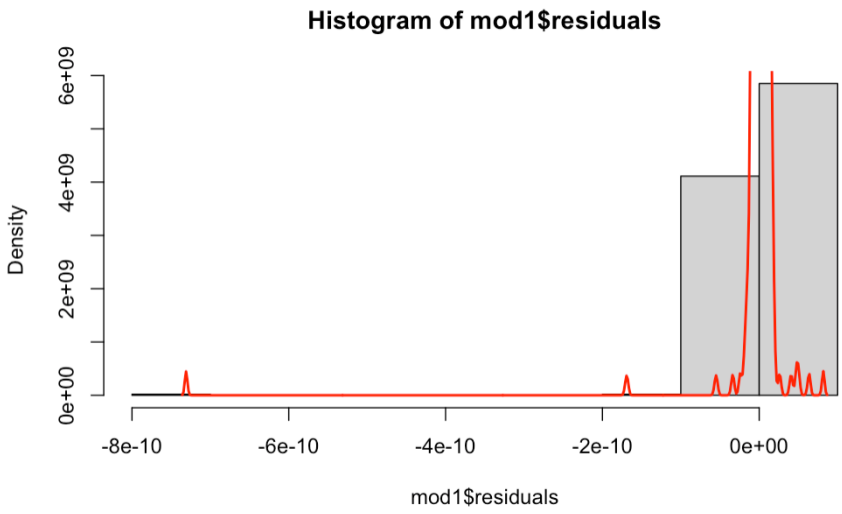

Figura 2:

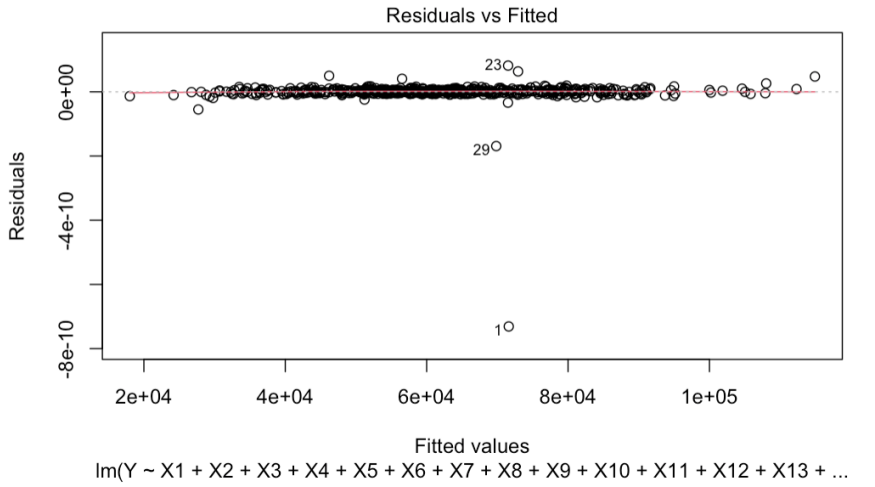

Figura 3:

A continuación, evaluaremos si existe un problema de colinealidad entre las variables independientes. Para ello, utilizaremos el Factor de Inflación de la Varianza (VIF), empleando la función correspondiente de la librería car. Recordemos que un valor de $VIF > 10$ indica la presencia de multicolinealidad severa. Después de ejecutar la función en el software, obtuvimos los siguientes resultados:

Podemos decir que sí hay problemas de colinealidad con las variables X1, X2 y en mayor medida con X7 y X11.

Prueba RESET

Ahora usaremos la prueba RESET, esta nos indica cuando un modelo está bien especificado, es decir, si estamos omitiendo variables relevantes, forma funcional incorrecta, entonces nuestras hipótesis son:

$$H_0 : El\ modelo\ está\ bien\ especificado$$

$$H_1 : El\ modelo\ está\ mal\ especificado$$

Aplicando la función necesaria en R nos da indica con un $p - value = 0.1922$ que el modelo con todas las variables está bien especificado, esto tiene sentido, pues al ser un modelo sobre ajustado no hay variables que

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
8.926881	8.963761	1.053000	1.025607	1.038192	1.028366	15.729422	1.045859	1.023049	1.030106	15.682175	1.037765	1.056979	1.026377	1.301310	1.370086	1.589038	1.029245

Figura 4:

se están omitiendo.

Nuestros AIC y BIC son -24243.92 y -24158.28 respectivamente.

2.1.4. Resultados

Observamos que a pesar de obtener un $R^2 = 1$, no podemos decir que es el mejor modelo porque se trata de un problema de sobreajuste, además no pasa las pruebas de los supuestos (normalidad, autocorrelación y colinealidad), para intentar solucionarlo en los siguientes modelos haremos algunas transformaciones y seleccionaremos solo algunas variables de la base, las más importantes.

2.2. Modelo 2

2.2.1. Hipótesis Conceptual

Nuestra hipótesis en este modelo se basa en considerar si la hora del día, el número de vehículos involucrados, el tipo de incidente así como el reclamo por lesiones, por daños materiales y por el vehículo explican adecuadamente el modelo o deberías considerar otras variables y discernir de algunas propuestas en este modelo.

2.2.2. Proceso de selección de variables

- Tipo de incidente (algunas reparaciones resultan más costosas que otras)
- Hora del día (las horas pico pueden influir en costos extra)
- Número de vehículos (elevamos al cuadrado la variable)
- Reclamo por lesiones (cuentas de hospital y demás)
- Reclamo por daños materiales (aplicamos raíz cuadrada)
- Reclamo por el vehículo (efecto directo sobre el monto)

2.2.3. Validez del modelo

El modelo consiste en:

$$Y = \beta_0 + \beta_7 X7 + \beta_{10} X10 + \beta_{11} X11^2 + \beta_{15} X15 + \beta_{16}\sqrt{X16} + \beta_{17}\sqrt{X17} \tag{6}$$

Este caso $R^2 = 0.9788$, por lo que tenemos un modelo que explica un buen porcentaje del monto total reclamado.

Pruebas de supuestos

Para la normalidad usaremos una prueba Jarque-Bera, tomando como juego de hipótesis:

$$H_0 : Los\ errores\ se\ distribuyen\ de\ forma\ normal$$

$$H_1 : Los\ errores\ no\ se\ distribuyen\ de\ forma\ normal$$

Tomando los resultados de R se tiene un $p - value <$ 2.2e-16, como es menor que 0.05, rechaza hipótesis nula y concluimos que los errores no siguen una distribución normal. Podemos también graficar la densidad de los errores y obtenemos:

```
Call:
lm(formula = Y ~ (X7) + (X15) + sqrt(X16) + sqrt(X17) + I(X11^2) +
    (X10), data = insurance_claims)

Residuals:
    Min      1Q  Median      3Q     Max
-8540.0 -2970.7  -447.5  2436.9 16292.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.379e+04  6.620e+02 -35.929  < 2e-16 ***
X7           1.815e+03  2.108e+02   8.612  < 2e-16 ***
X15          1.114e+00  3.620e-02  30.775  < 2e-16 ***
sqrt(X16)    1.474e+02  6.219e+00  23.693  < 2e-16 ***
sqrt(X17)    3.039e+02  4.117e+00  73.805  < 2e-16 ***
I(X11^2)    -7.641e+02  7.077e+01 -10.797  < 2e-16 ***
X10         -5.997e+01  1.810e+01  -3.314 0.000954 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3854 on 993 degrees of freedom
Multiple R-squared:  0.9788,    Adjusted R-squared:  0.9787
F-statistic:  7646 on 6 and 993 DF,  p-value: < 2.2e-16
```

Figura 5:

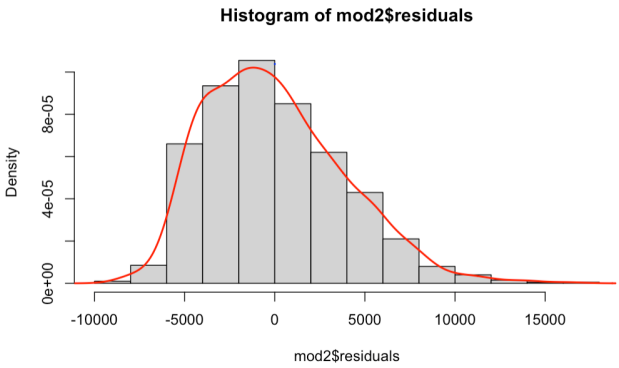

Figura 6:

Para la autocorrelación usaremos la prueba Durbin-Watson, donde:

$$H_0 : No\ hay\ autocorrelación\ en\ los\ errores$$

$$H_1 : Hay\ autocorrelación\ en\ los\ errores$$

El resultado devuelto en R es un $p-value = 0.05616$, mayor que 0.05, por lo que no rechazamos hipótesis nula, o sea no tenemos problemas de autocorrelación, es decir los errores son independientes y cumple con el supuesto.

En cuanto la heterocedasticidad, utilizamos la prueba de Breusch-Pagan, en la cual:

$$H_0 : La\ varianza\ de\ los\ errores\ es\ constante\ (homocedasticidad)$$

$$H_1 : La\ varianza\ de\ los\ errores\ no\ es\ constante\ (homocedasticidad)$$

Obtenemos un $p-value = 0{,}02551 < 0.05$, por lo que rechazamos H_0, así que tenemos heterocedasticidad. Gráficamente tenemos:

Notamos que hay una tendencia en los residuos, toman forma de U, además de notar también valores atípicos.

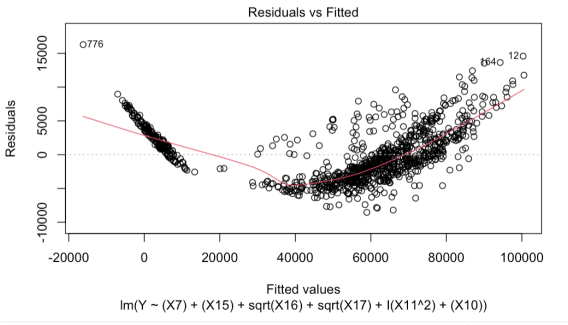

Figura 7:

X7	X15	sqrt(X16)	sqrt(X17)	I(X11^2)	X10
5.662376	2.098991	2.808228	4.443668	6.225167	1.064185

Figura 8:

A continuación, evaluaremos la colinealidad, por lo que los resultados arrojados por R:
Resulta que ninguno da mayor que 10 por lo que podemos decir que no tenemos problemas de colinealidad.

Prueba RESET

Ahora usaremos la prueba RESET, esta nos indica cuando un modelo está bien especificado, es decir, si estamos omitiendo variables relevantes, forma funcional incorrecta, entonces nuestras hipótesis son:

$$H_0 : El\ modelo\ está\ bien\ especificado$$

$$H_1 : El\ modelo\ está\ mal\ especificado$$

Aplicando la función necesaria en R nos da indica con un $p - value <$ 2.2e-16, lo que indica un modelo que no está bien especificado.

Nuestro AIC y BIC son 19360.81 y 19400.07 respectivamente.

2.2.4. Resultados

En comparación con el modelo 1 presenta mejoras, sin embargo, está lejos de ser el mejor u óptimo para poder explicar el monto total reclamado, presenta varios problemas como la independencia, heterocedasticidad, especificación, pero pasa la colinealidad y autocorrelación.

2.3. Modelo 3

2.3.1. Hipótesis Conceptual

La conjetura para este modelo son que el deducible de la póliza, la prima de póliza, el número de vehículos involucrados, las lesiones corporales, reclamo por lesiones, reclamo por daños materiales y reclamo por vehículo son importantes para explicar el monto total reclamado.

2.3.2. Proceso de selección de variables

- Deducible de la póliza (multiplicado por reclamo por lesiones)
- Prima anual de la póliza (multiplicado por reclamo por daños materiales)
- El número de vehículos involucrados
- Lesiones corporales (aplicamos función exponencial)

- Reclamo por lesiones (multiplicado por deducible por lesiones)
- Reclamos por daños materiales (multiplicado por prima)
- Reclamo por vehículo (aplicamos función logarítmica)

2.3.3. Validez del modelo

El modelo consiste en:

$$Y = \beta_0 + \gamma_1\sqrt{X3X15} + \gamma_2 X4X16 + \beta_{11}log(X11) + \beta_{12}exp(X12) + \beta_{17}log(X17) \tag{7}$$

```
Call:
lm(formula = Y ~ sqrt(X3 * X15) + I(X4 * X16) + log(X11) + exp(X12) +
    log(X17), data = insurance_claims)

Residuals:
   Min     1Q Median     3Q    Max
-18370  -5418  -1014   4571  69904

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.461e+05  3.548e+03 -41.183  < 2e-16 ***
sqrt(X3 * X15)  2.850e+00  2.435e-01  11.707  < 2e-16 ***
I(X4 * X16)     1.072e-03  5.531e-05  19.382  < 2e-16 ***
log(X11)       -2.016e+03  4.993e+02  -4.038 5.81e-05 ***
exp(X12)        1.794e+02  9.430e+01   1.902   0.0575 .
log(X17)        1.775e+04  3.983e+02  44.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8047 on 994 degrees of freedom
Multiple R-squared:  0.9076,    Adjusted R-squared:  0.9071
F-statistic:  1952 on 5 and 994 DF,  p-value: < 2.2e-16
```

Figura 9:

Con este modelo tenemos un R^2 =.9076, lo que es algo bueno, también no dista mucho del R^2 ajustada.

Pruebas de supuestos

Para la normalidad usaremos una prueba Jarque-Bera, tomando como juego de hipótesis:

$$H_0 : Los\ errores\ se\ distribuyen\ de\ forma\ normal$$

$$H_1 : Los\ errores\ no\ se\ distribuyen\ de\ forma\ normal$$

Tomando los resultados de R se tiene un $p-value <$ 2.2e-16, como es menor que 0.05, rechaza hipótesis nula y concluimos que los errores no siguen una distribución normal.
Podemos también graficar la densidad de los errores y obtenemos:

Para la autocorrelación usaremos la prueba Durbin-Watson, donde:

$$H_0 : No\ hay\ autocorrelación\ en\ los\ errores$$

$$H_1 : Hay\ autocorrelación\ en\ los\ errores$$

El resultado devuelto en R es un $p-value = 0.5774$, mayor que 0.05, por lo que no rechazamos hipótesis nula, o sea no tenemos problemas de autocorrelación, es decir los errores son independientes y cumple con el supuesto.

Para la heterocedasticidad, utilizamos la prueba de Breusch-Pagan, en la cual:

$$H_0 : Los\ varianza\ de\ los\ errores\ es\ constante\ (homocedasticidad)$$

$$H_1 : La\ varianza\ de\ los\ errores\ no\ es\ constante\ (heterocedasticidad)$$

Obtenemos un $p-value = 0.0003144 < 0.05$, por lo que rechazamos H_0, así que tenemos heterocedasticidad.
Gráficamente observamos:

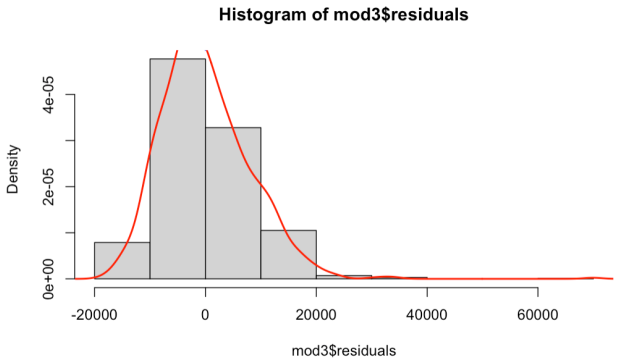

Figura 10: Caption

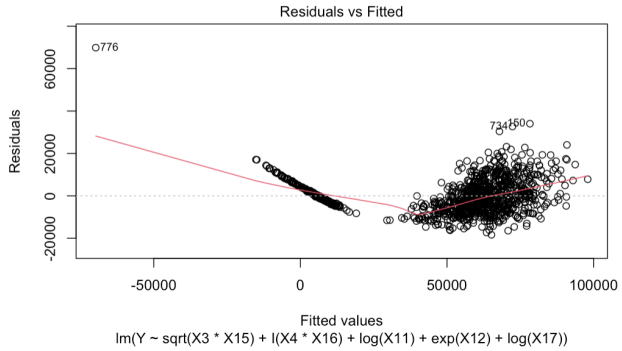

Figura 11:

Los errores siguen una tendencia evidente, por lo que podríamos buscar usar otras variables.
A continuación, evaluaremos la colinealidad, por lo que los resultados arrojados por R:
Resulta que ninguno da mayor que 10 por lo que podemos decir que no tenemos problemas de colinealidad.

Prueba RESET

Ahora usaremos la prueba RESET, esta nos indica cuando un modelo está bien especificado, es decir, si estamos omitiendo variables relevantes, forma funcional incorrecta, entonces nuestras hipótesis son:

$$H_0 : El\ modelo\ está\ bien\ especificado$$

$$H_1 : El\ modelo\ está\ mal\ especificado$$

Aplicando la función necesaria en R nos da indica con un $p - value <$ 2.2e-16, lo que indica un modelo que no está bien especificado.

Estadísticos

Nuestro AIC y BIC son 20831.97 y 20866.33 respectivamente, más altos que en el modelo 1 y 2.

2.3.4. Resultados

Este modelo tampoco es el mejor porque falla en los supuestos, a excepción de la autocorrelación, también sufre de problemas de especificación, podemos decir que un R2 cercana a 1 no siempre indica un buen modelo.

```
sqrt(X3 * X15)    I(X4 * X16)       log(X11)       exp(X12)       log(X17)
      1.685598       1.867552       1.144195       1.004315       2.449613
```

Figura 12:

2.4. Modelo 4

2.4.1. Hipótesis Conceptual

Para este modelo suponemos que el deducible de la póliza, la prima anual de la póliza, el sexo (variable dummy), el reclamo por lesiones y el reclamo por daños materiales pueden justificar de modo adecuado el monto total de reclamo, sin perder significancia en las variables, así como en las pruebas.

2.4.2. Proceso de selección de variables

- Deducible de la póliza (puede reducir el monto reclamado porque el asegurado tendría que cubrir una mayor parte del costo)
- Prima anual de la póliza (estar asociada con una mayor cobertura, lo que podría resultar en mayores pagos)
- Sexo del asegurado (algunos estudios sugieren que los hombres tienden a tener un mayor riesgo de accidentes o reclamaciones por daños materiales)
- Reclamo por lesiones
- Reclamo por daños materiales

2.4.3. Validez del modelo

El último modelo consiste en:

$$log(Y) = \beta_3 X3 + \beta_4 X4 + \beta_6 X6 + \beta_{15} X15 + \beta_{16} X16 \tag{8}$$

```
Call:
lm(formula = log(Y) ~ X3 + X4 + X6 + X15 + X16 - 1, data = insurance_claims)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6777 -0.8333  0.2307  1.2925  5.3254

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
X3  6.737e-04  8.066e-05   8.353 2.22e-16 ***
X4  5.587e-03  1.015e-04  55.064  < 2e-16 ***
X6  6.485e-01  1.011e-01   6.413 2.21e-10 ***
X15 1.505e-04  1.265e-05  11.897  < 2e-16 ***
X16 1.441e-04  1.286e-05  11.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.629 on 995 degrees of freedom
Multiple R-squared:  0.9766,    Adjusted R-squared:  0.9765
F-statistic:  8299 on 5 and 995 DF,  p-value: < 2.2e-16
```

Figura 13:

En este caso aplicamos la función logaritmo principalmente por mejorar la adecuación del modelo a los supuestos de la regresión lineal, como la linealidad, la homoscedasticidad y la normalidad de los errores. Observamos que tenemos un R^2 =0.9766, lo cuál no difiere mucho de los demás modelos.

Pruebas de supuestos

Para la normalidad usaremos una prueba Jarque-Bera, tomando como juego de hipótesis:

$$H_0 : Los\ errores\ se\ distribuyen\ de\ forma\ normal$$

$$H_1 : Los\ errores\ no\ se\ distribuyen\ de\ forma\ normal$$

El resultado en R consistió en un $p - value$ =0.969, es decir, no rechazamos hipótesis nula, por lo que tenemos normalidad en nuestros residuos, hasta ahora ha sido el modelo que pasa esta prueba, además de que todas las variables son significativas.

También podemos observar un histograma:

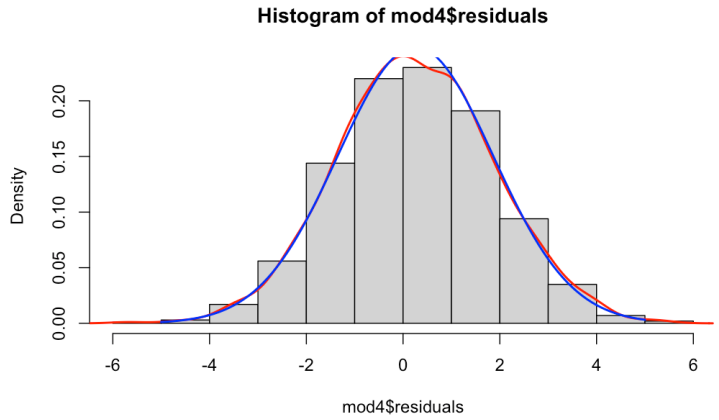

Figura 14:

Notamos que la densidad de los errores se acerca demasiado a la distribución normal con $\mu = 0$ y $\sigma^2 = 1$.
Para la autocorrelación usaremos la prueba Durbin-Watson, donde:

$$H_0 : No\ hay\ autocorrelación\ en\ los\ errores$$

$$H_1 : Hay\ autocorrelación\ en\ los\ errores$$

Este modelo nos indica con un $p - value = 0{,}6847$ que no tenemos problemas de autocorrelación, es decir lo errores sí son independientes.

Para la heterocedasticidad, utilizamos la prueba de Breusch-Pagan, en la cual:

$$H_0 : Los\ varianza\ de\ los\ errores\ es\ constante\ (homocedasticidad)$$

$$H_1 : La\ varianza\ de\ los\ errores\ no\ es\ constante\ (heterocedasticidad)$$

Obtenemos un $p - value = 0.775 > 0.05$, por lo que no rechazamos H_0, así que tenemos homocedasticidad.

Hasta aquí hemos notado que este cuarto modelo cumple con los supuestos de normalidad, autocorrelación y homocedasticidad, comparándolo con los anteriores, podemos decir que es el mejor, sin embargo, ahora veremos el VIF de este modelo.

El resultado para el VIF de este modelo es:

```
      X3       X4       X6      X15      X16
4.079185 6.351857 2.068772 4.766439 4.862515
```

Figura 15:

Ninguna de nuestras respuestas resulta mayor a 10, así que no presenta problemas de colinealidad.

Criterios de información

Nuestros AIC y BIC para este modelo son 3821.226 y 3850.672, respectivamente, notamos que son un poco más altos que los pasados.

2.4.4. Resultados

En comparación con los modelos pasados nos encontramos con una mejora evidente, esto se debe a la aplicación de la función logaritmo en la variable dependiente, pues ayudó a resolver los problemas de normalidad, homocedasticidad, autocorrelación, además de mantener un vif menor que 10 en las variables.

2.5. Elección de modelo

Los resultados de los modelos nos da una señal sobre cuál puede explicar de mejor forma el monto reclamado, hay que tomar en cuenta varias cosas, desde el cumplimiento de los supuestos hasta los criterios de información, de esta forma utilizamos la librería "performance". Los modelos seleccionados para comparar son el modelo 3 y 4, porque estos mostraron mejores resultados con los supuestos. La respuesta fue la siguiente:

```
# Comparison of Model Performance Indices

Name | Model |    R2 | R2 (adj.) |     RMSE |    Sigma | AIC weights | AICc weights | BIC weights | Performance-Score
-----------------------------------------------------------------------------------------------------------------------
mod4 |    lm | 0.977 |     0.976 |    1.625 |    1.629 |    0.00e+00 |     0.00e+00 |    0.00e+00 |            57.14%
mod3 |    lm | 0.908 |     0.907 | 8022.846 | 8047.023 |        1.00 |         1.00 |        1.00 |            42.86%
```

Figura 16:

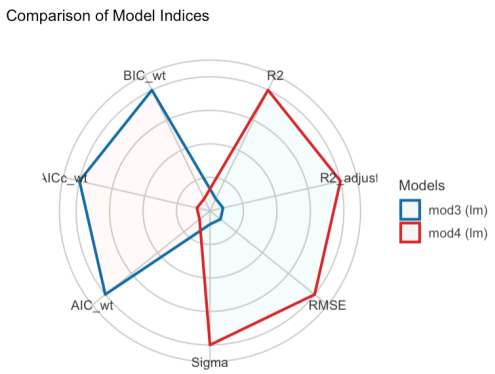

Figura 17: Caption

Los resultados indican que el mejor modelo para poder explicar el monto reclamado es el 4, esto se puede hacer evidente desde que pasó las pruebas de los supuestos, el 3 solo tuvo éxito para cumplir la autocorrelación.

3. Conclusión

El modelo estimado permite identificar y cuantificar los factores que influyen en el monto total reclamado (Y) en un seguro de auto. Los resultados indican que variables clave como el deducible de la póliza (X3), la prima anual (X4), el sexo del asegurado (X6), así como la presencia de reclamos por lesiones (X15) y reclamos por daños materiales (X16), tienen un impacto significativo en el monto reclamado.

El logaritmo del monto reclamado se explica adecuadamente a partir de estas variables, lo que proporciona un marco predictivo útil para evaluar el comportamiento financiero del seguro. Además, al pasar las pruebas de los supuestos, el modelo demuestra ser fiable para la toma de decisiones estratégicas, como la segmentación de clientes y el diseño de tarifas.

En este contexto, el modelo refuerza la importancia de personalizar las pólizas en función de características específicas del asegurado y del tipo de reclamo, optimizando así el equilibrio entre la cobertura ofrecida y la sostenibilidad financiera del producto.

4. Anexos

En este enlace se puede encontrar el archivo Rmd y la base datos:

https://drive.google.com/drive/folders/1HA5Qv_nqk__Y6y17PHa0GtkGQo8rKG0e?usp=drive_link

Referencias

R. Carter Hill, William E. Griffiths, and Guay C. Lim. *Principles of Econometrics*. John Wiley & Sons, Hoboken, NJ, 2017.

Bunty Shah. Auto insurance claims data. https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data, 2023.