Maestría Oficial en Big Data y Data Science

ACTIVIDAD 1

06MBID – Estadística Avanzada

Alumno: ZOLEIDA MORALES ARISTIZABAL

Fecha de entrega: 30/08/2023

Índice

1. Contexto y/o Motivación

Por medio de este trabajo se espera afianzar los conocimientos adquiridos durante la asignatura, de modo que todos aquellos conceptos que fueron revisados en clase puedan ser puestos en práctica.

El trabajo consiste en realizar un análisis estadístico completo desarrollado en R, seleccionar un dataset o base de datos para completar el análisis descriptivo univariante y bivariante de la misma, se debe estimar y validar un modelo de regresión.

Además de interpretar los resultados, se debería discutir el potencial del modelo y describir posibles limitaciones y/o mejoras.

2. Objetivos

- Realizar un análisis estadístico completo programando con R.
- Estimar y validar un modelo de regresión y realizar predicciones.
- Extraer conclusiones a partir del análisis estadístico.
- Transmitir las conclusiones obtenidos de manera clara y concisa a través de herramientas de visualización de datos.

3. Base de datos

Se cuenta con una base de datos del sector automotriz colombiano en donde, se recopila las pólizas emitidas del seguro de automóviles para el periodo 01/10/2011 a 31/12/2013, y cada una de las **120.930** entradas representan datos agregados de **23** características de vehículos en diferentes estados de Colombia.

Las características que contiene nuestro dataset, son:

CARACTERÍSTICA	DESCRIPCIÓN	TIPO DE DATO	% DE COMPLETITUD
ID	Identificador de la fila	Numérico	100
Tipo póliza	Indicador si las pólizas son nuevas o renovadas	String	100
Valor Asegurado	Suma total acumulada de los valores asegurados de cada una de las coberturas de la póliza.	Numérico	100
Fecha Emisión	Fecha en la que se emite la póliza.	Date	100
Fecha Inicio	Fecha en que inicia la cobertura del vehículo.	Date	100
Fecha fin	Fecha en que termina la cobertura del vehículo.	Date	100
Valor prima Anual	Costo total anualizado de las coberturas otorgadas por la aseguradora de cada póliza expedida.	Numérico	100
Valor asegurado Vehículo	Valor asegurado del casco del vehículo de acuerdo con la Guía de valores de Fasecolda.	Numérico	100
Ciudad	Ciudad de circulación del vehículo.	String	99,7
Departamento	Estado de circulación del vehículo.	String	99,7
Ocupación	Ocupación del asegurado.	String	100
Edad	Edad del Asegurado.	String	90,3
MARCA/TIPO (CODIGO FASECOLDA)	Tipo de vehículo y marca según la organización FASECOLDA.	String	99,9
Fasecolda-MARCA	Marca del vehículo según la organización FASECOLDA.	String	100
Fasecolda-REF1	Referencia1 según la organización FASECOLDA.	String	100
Fasecolda-REF2	Referencia2 según la organización FASECOLDA.	String	96
Fasecolda-REF3	Referencia3 según la organización FASECOLDA.	String	96
Fasecolda-CLASE	Clase del vehículo según FASECOLDA	Numérico	100
Modelo del vehículo	Año del modelo del vehículo de acuerdo con la tarjeta de propiedad.	String	99,8

CARACTERÍSTICA	DESCRIPCIÓN	TIPO DE DATO	% DE COMPLETITUD
Color	Color del vehículo de acuerdo con la tarjeta de propiedad.	String	99,9
Años de no reclamación	Cantidad de años en que el asegurado no ha presentado siniestros.	String	62,3
Género	Genero del asegurado.	String	100
Deducibles	Plan de deducible escogido por el asegurado en caso de sufrir un accidente o una pérdida.	String	4,4

```
> dim(Proyecto)
[1] 120930     23
>
```

```
> glimpse(Proyecto)
Rows: 120,930
Columns: 23
$ ID                      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, …
$ TipoPoliza              <chr> "Renovacion", "Nuevo", "Nuevo",…
$ ValorAsegurado          <dbl> 684689000, 965205600, 640605600…
$ FechaEmision            <chr> "10/10/2011", "10/10/2011", "10…
$ FechaInicio             <chr> "11/21/2011", "10/5/2011", "10/…
$ FechaFin                <chr> "11/21/2012", "10/5/2012", "10/…
$ ValorPrimaAnual         <int> 782949, 715824, 740816, 745224,…
$ ValorAseguradoVehiculo  <int> 18800000, 24900000, 12600000, 2…
$ Ciudad                  <chr> "CARTAGENA", "BOGOTA D.C.", "ME…
$ Departamento            <chr> "BOLIVAR", "CUNDINAMARCA", "ANT…
$ Ocupacion               <chr> "EMPLEADO(A)", "EMPLEADO(A)", "…
$ Edad                    <chr> "45", "41", "33", "50", "38", "…
$ Marca.Tipo              <chr> "CHEVROLET AVEO FAMILY MT 1500C…
$ Marca                   <chr> "CHEVROLET", "CHEVROLET", "RENA…
$ REF1                    <chr> "AVEO", "AVEO EMOTION", "CLIO I…
$ REF2                    <chr> "FAMILY", "1.6L", "F.II EXPRESS…
$ REF3                    <chr> "MT 1500CC 4P AA", "MT 1600CC A…
$ Clase                   <chr> "AUTOMOVIL", "AUTOMOVIL", "AUTO…
$ ModeloVehiculo          <chr> "2010", "2009", "2002", "2006",…
$ Color                   <chr> "NEGRO EBONY", "BLANCO ARCO BIC…
$ AnosDeNoReclamación     <chr> "DOS AÑOS CONTINUOS", "CUATRO O…
$ Genero                  <chr> "MASCULINO", "MASCULINO", "FEME…
$ Deducibles              <chr> "10% del Valor de Reclamación M…
```

Para efectos de la práctica sólo se van a tener en cuenta las variables numéricas y se divide sobre 100.000 con el objetivo de visualizar de manera más sencilla, los datos.

```
Proyecto1 <- select(Proyecto, ValorAsegurado, ValorPrimaAnual,
                    ValorAseguradoVehiculo)/100000
glimpse(Proyecto1)
```

```
> glimpse(Proyecto1)
Rows: 120,930
Columns: 3
$ ValorAsegurado         <dbl> 6846.890, 9652.056, 6406.056, 6556.056, 12701.690, 2…
$ ValorPrimaAnual        <dbl> 7.82949, 7.15824, 7.40816, 7.45224, 7.71465, 8.06297…
$ ValorAseguradoVehiculo <dbl> 188.0000, 249.0000, 126.0000, 201.0000, 146.0000, 22…
```

Aunque el modelo del vehículo es un dato numérico, no haría parte de este análisis de correlación porque a pesar de los números, no representa un valor como tal.

La variable de salida o sujeta de predicción sería ValorPrimaAnual, que estaría sujeta a varios acontecimientos, por ejemplo: al modelo de vehículo, a si el vehículo ha tenido deducibles anteriormente, al valor asegurado, entre otros.

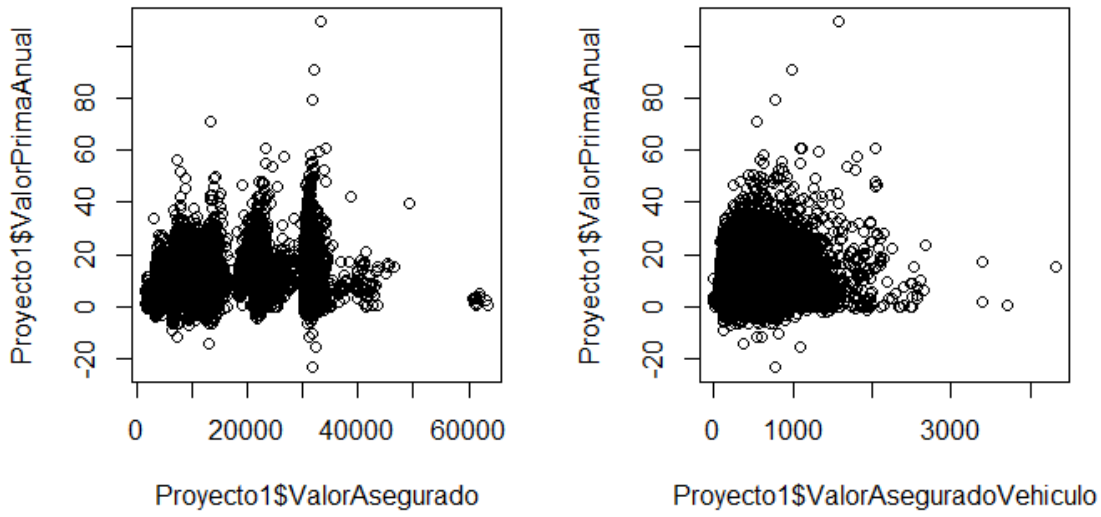
4. Análisis descriptivo

Matriz de correlación:

```
Proyecto1_cor <-cor(Proyecto1, method = 'pearson')
round_corr <-round(Proyecto1_cor,digits = 1)
round_corr
```

```
> round_corr
                       ValorAsegurado ValorPrimaAnual ValorAseguradoVehiculo
ValorAsegurado                    1.0            -0.1                    0.1
ValorPrimaAnual                  -0.1             1.0                    0.2
ValorAseguradoVehiculo            0.1             0.2                    1.0
>
```

```
#Divide la pantalla en 2 columnas
par(mfrow=c(1,2))

plot(x=Proyecto1$ValorAsegurado, y=Proyecto1$ValorPrimaAnual)
plot(x=Proyecto1$ValorAseguradoVehiculo, y=Proyecto1$ValorPrimaAnual)
```




```
> summary(Proyecto1)
 ValorAsegurado  ValorPrimaAnual   ValorAseguradoVehiculo
 Min.    : 1732   Min.    :-23.316   Min.    :   0.288
 1st Qu.:12719   1st Qu.:  1.801   1st Qu.: 194.000
 Median :21730   Median :  5.063   Median : 279.900
 Mean   :21357   Mean   :  5.744   Mean   : 322.601
 3rd Qu.:30687   3rd Qu.:  8.202   3rd Qu.: 402.000
 Max.   :63453   Max.   :108.968   Max.   :4320.000
Warning messages:
```

Se puede identificar que las variables están altamente correlacionadas entre sí, que el valor prima anual es mayor si tanto el valor asegurado, como el valor asegurado vehículos son mayores.

5. Modelo de regresión

Una vez conocidos las variables y con la teoría al respecto se procede a generar la correspondiente regresión, tal y como se evidencia continuación:

En este sentido el modelo será:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \varepsilon$$

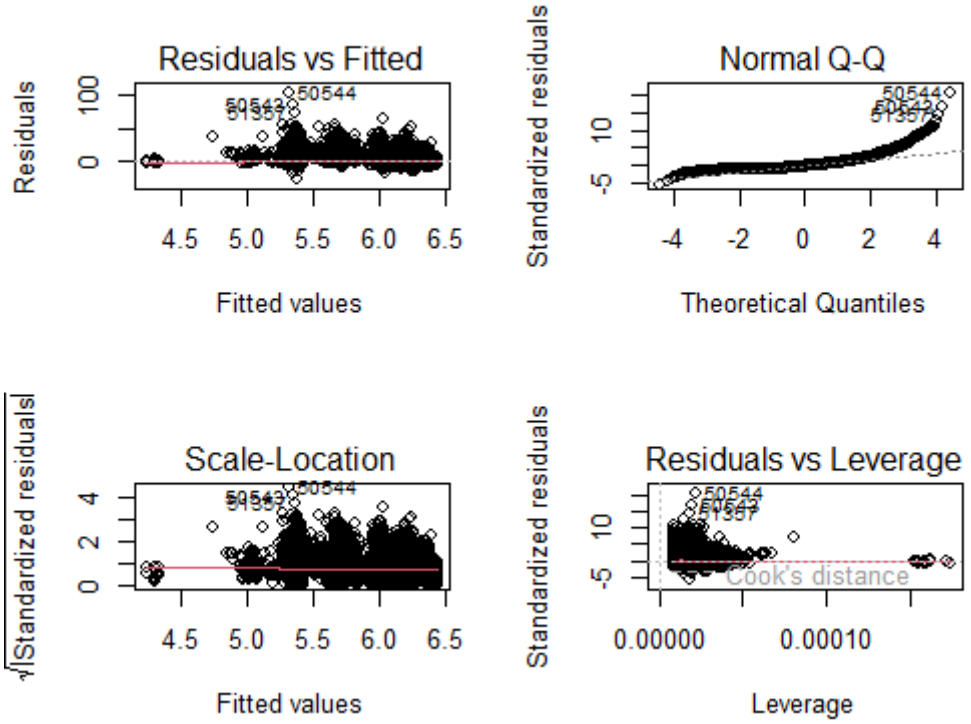
Aplicándolo en R, como se está realizando la regresión por una sola variable, generé dos modelos para encontrar el valor prima anual, de acuerdo con las otras dos variables de valor asegurado y valor asegurado vehículo.

```
lm.fit <- lm(ValorPrimaAnual~ValorAsegurado, data = Proyecto1)
summary(lm.fit)
(lm.fit)
coef(lm.fit)

# Calculamos intervalos
confint(lm.fit)
predict(lm.fit,data.frame(ValorAsegurado=(c(5,10,15))),
        interval="prediction")

# Validamos los supuestos
par(mfrow=c(2,2))
plot(lm.fit)
```

```
> summary(lm.fit)

Call:
lm(formula = ValorPrimaAnual ~ ValorAsegurado, data = Proyecto1)

Residuals:
    Min      1Q  Median      3Q     Max
-28.690  -3.843  -0.671   2.372 103.652

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.506e+00  3.644e-02  178.55   <2e-16 ***
ValorAsegurado -3.571e-05  1.562e-06  -22.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.103 on 120928 degrees of freedom
Multiple R-squared:  0.004305,  Adjusted R-squared:  0.004297
F-statistic: 522.9 on 1 and 120928 DF,  p-value: < 2.2e-16
```

```
> coef(lm.fit)
   (Intercept) ValorAsegurado
  6.506373e+00  -3.571379e-05
```

```
> predict(lm.fit,data.frame(ValorAsegurado=(c(5,10,15))),
+         interval="prediction")
       fit       lwr      upr
1 6.506194 -3.495539 16.50793
2 6.506016 -3.495717 16.50775
3 6.505837 -3.495896 16.50757
```



De acuerdo con la interpretación de los gráficos:

- Se encuentra que los residuos están distribuidos equitativamente alrededor de la línea, esto implica que no hay relaciones no lineales.
- Para el segundo gráfico los residuos se alinean bien a la línea discontinua recta.
- Es correcto porque los puntos de dispersión están de forma similar sobre la línea recta.
- Para el cuarto gráfico, al parecer se tienen algunos puntos influyectes.

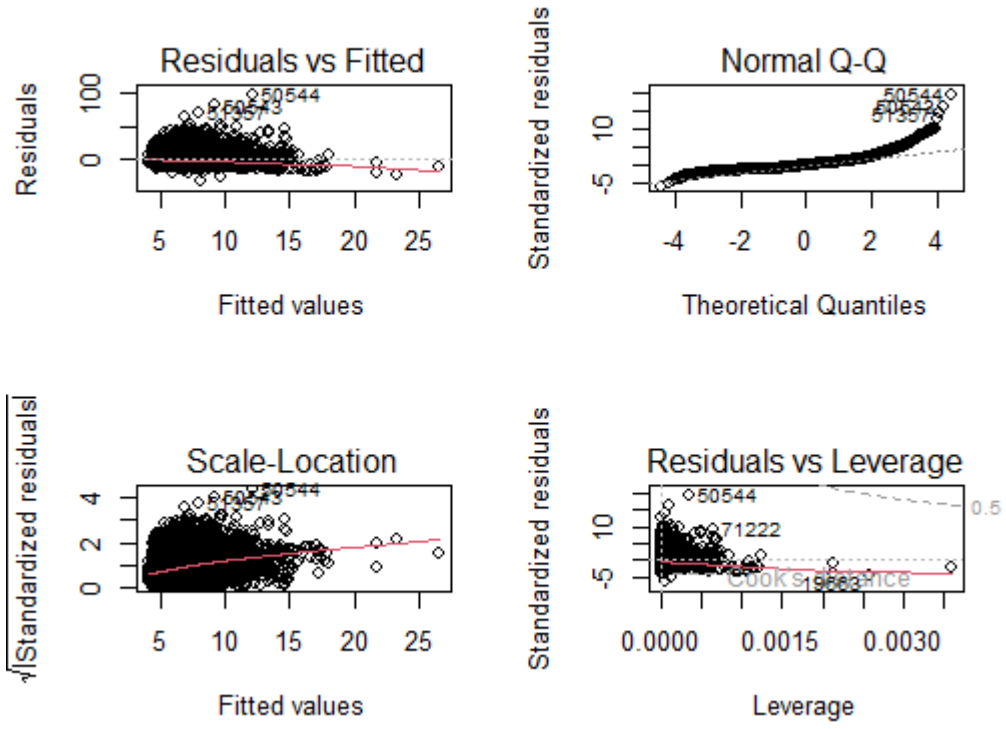
Con la otra variable:

```
lm.fit <- lm(ValorPrimaAnual~ValorAseguradoVehiculo, data = Proyecto1)
summary(lm.fit)
(lm.fit)
coef(lm.fit)

# Calculamos intervalos
confint(lm.fit)
predict(lm.fit,data.frame(ValorAseguradoVehiculo=(c(5,10,15))),
        interval="prediction")

# Validamos los supuestos
par(mfrow=c(2,2))
plot(lm.fit)
```

```
Call:
lm(formula = ValorPrimaAnual ~ ValorAseguradoVehiculo, data = Proyecto1)

Residuals:
    Min      1Q  Median      3Q     Max
-31.423  -3.895  -0.503   2.585  96.732

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.062e+00  2.820e-02  144.06   <2e-16 ***
ValorAseguradoVehiculo 5.213e-03  7.511e-05   69.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.015 on 120928 degrees of freedom
Multiple R-squared:  0.03831,   Adjusted R-squared:  0.0383
F-statistic:  4817 on 1 and 120928 DF,  p-value: < 2.2e-16
```

```
coef(lm.fit)
          (Intercept) ValorAseguradoVehiculo
           4.06197685              0.00521284
```

```
> predict(lm.fit,data.frame(ValorAseguradoVehiculo=(c(5,10,15))),
+         interval="prediction")
       fit       lwr      upr
1 4.088041 -5.741321 13.91740
2 4.114105 -5.715253 13.94346
3 4.140169 -5.689186 13.96952
```

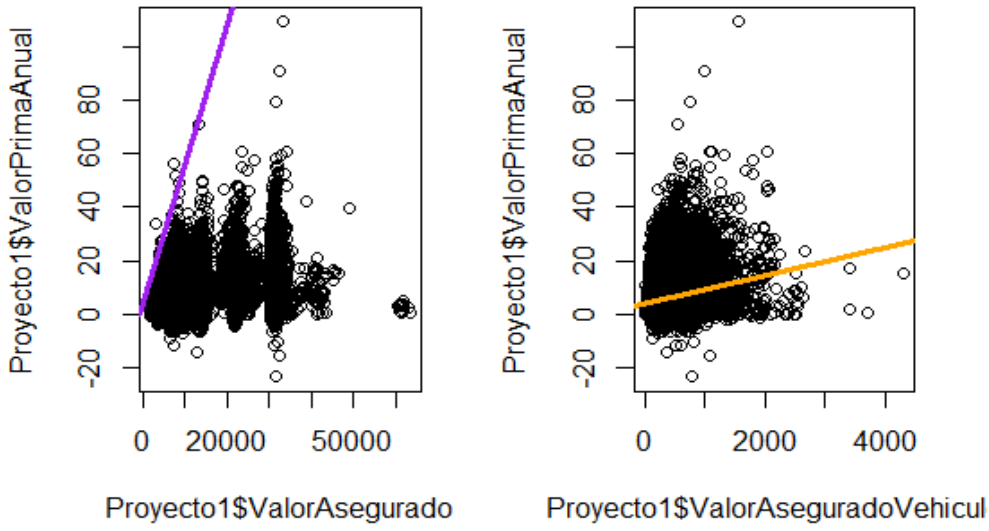


De acuerdo con la interpretación de los gráficos:

- Se encuentra que los residuos están distribuidos equitativamente alrededor de la línea, esto implica que no hay relaciones no lineales.
- Para el segundo gráfico los residuos se alinean bien a la línea discontinua recta.
- Es correcto porque los puntos de dispersión están de forma similar sobre la línea recta.
- Para el cuarto gráfico, al parecer se tienen algunos puntos influyectes.

Gráficos de los modelos

```
# Graficamos el modelo
par(mfrow=c(1,2))

plot(Proyecto1$ValorAsegurado,Proyecto1$ValorPrimaAnual)
abline(lm.fit)
abline(lm.fit,lwd=3)
abline(lm.fit,lwd=3,col="purple")

plot(Proyecto1$ValorAseguradoVehiculo,Proyecto1$ValorPrimaAnual)
abline(lm.fit)
abline(lm.fit,lwd=3)
abline(lm.fit,lwd=3,col="orange")
```




```
# REGRESION LINEAL MULTIPLE

lm.fit <- lm(ValorPrimaAnual~ValorAsegurado+
               ValorAseguradoVehiculo, data = Proyecto1)
summary(lm.fit)
lm.fit <- lm(ValorPrimaAnual~., data=Proyecto1)
summary(lm.fit)
```

```
> lm.fit <- lm(ValorPrimaAnual~., data=Proyecto1)
> summary(lm.fit)

Call:
lm(formula = ValorPrimaAnual ~ ., data = Proyecto1)

Residuals:
    Min      1Q  Median      3Q     Max
-31.044  -3.735  -0.446   2.494  96.967

Coefficients:
                       Estimate Std. Error t value
(Intercept)           5.017e+00  4.110e-02  122.07
ValorAsegurado       -4.897e-05  1.539e-06  -31.81
ValorAseguradoVehiculo  5.496e-03  7.532e-05   72.96
                      Pr(>|t|)
(Intercept)             <2e-16 ***
ValorAsegurado          <2e-16 ***
ValorAseguradoVehiculo  <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.994 on 120927 degrees of freedom
Multiple R-squared:  0.04629,   Adjusted R-squared:  0.04628
F-statistic:  2935 on 2 and 120927 DF,  p-value: < 2.2e-16
```

Se puede observar que las variables tienen un nivel de significación alto, ya que tienen un valor inferior al 0.01 %.

El valor asegurado no está dando un valor negativo.

En cuanto a los coeficientes R^2 nos estaría dando alrededor de un valor muy pequeño de un 0.046 y un R ajustado muy similar.

Vamos a revisar el modelo agregando una variable categórica que para este caso sería el tipo de póliza.

```
lm.fit <- lm(ValorPrimaAnual~ValorAsegurado
             +ValorAseguradoVehiculo+TipoPoliza,
             data = Proyecto)
summary(lm.fit)
```

```
> lm.fit <- lm(ValorPrimaAnual~ValorAsegurado
+              +ValorAseguradoVehiculo+TipoPoliza,
+              data = Proyecto)
> summary(lm.fit)

Call:
lm(formula = ValorPrimaAnual ~ ValorAsegurado + ValorAseguradoVehiculo +
    TipoPoliza, data = Proyecto)

Residuals:
     Min       1Q   Median       3Q      Max
-2528081  -293809   -50945   189513  9617663

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.165e+04  5.060e+03   10.21   <2e-16 ***
ValorAsegurado        -9.198e-05  1.498e-06  -61.40   <2e-16 ***
ValorAseguradoVehiculo 5.625e-03  6.855e-05   82.06   <2e-16 ***
TipoPolizaNuevo        6.521e+05  4.107e+03  158.76   <2e-16 ***
TipoPolizaRenovacion   5.027e+05  4.650e+03  108.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 454300 on 120925 degrees of freedom
Multiple R-squared:  0.2109,    Adjusted R-squared:  0.2109
F-statistic:  8081 on 4 and 120925 DF,  p-value: < 2.2e-16
```

Con la agregación de la variable al modelo, se sigue evidenciando que todas las variables son significativamente altas para el modelo.

Se puede evidenciar también que el R^2 estaría mejorando con un valor aproximado de un 21% de colinealidad y un R ajustado un poco más pequeño, castigado por la cantidad de variables utilizadas en el modelo.

6. Conclusiones

Con la elaboración de este trabajo práctico se ha podido afianzar más los conocimientos adquiridos en las clases, ya que al hacer uno mismo cada ejercicio y tratar de interpretarlo le lleva un gran esfuerzo que implica un mayor entendimiento de lo visto en clase.

En trabajos futuros sería interesante poder utilizar esta base de datos disponible para validar su eficiencia con varios modelos y establecer cuál sería el más adecuado, sin embargo, siento que requiero una mayor expertiz para realizar este tipo de análisis.

7. Anexos

Regresión Lineal

```
# Cargar dataset

setwd("/Users/zoleida.morales/Documents/Master
Bigdata/06_EstadisticaAvanzada/Practica1")

wd <- getwd(); file <- paste(wd,

                    "Base_proyecto.csv",sep="/")

# Para leer los datos anteriores

Proyecto<-read.csv(file, head=TRUE)[,-1]

#Exploremos los datos

library(readxl)

library(dplyr)

library(tidyr)

dim(Proyecto)

head(Proyecto, n=5)

glimpse(Proyecto)

Proyecto1 <- select(Proyecto, ValorAsegurado, ValorPrimaAnual,

          ValorAseguradoVehiculo)/100000

glimpse(Proyecto1)

Proyecto1_cor <-cor(Proyecto1, method = 'pearson')

round_corr <-round(Proyecto1_cor,digits = 1)

round_corr

summary(Proyecto1)

#Divide la pantalla en 2 columnas

par(mfrow=c(1,2))

plot(x=Proyecto1$ValorAsegurado, y=Proyecto1$ValorPrimaAnual)

plot(x=Proyecto1$ValorAseguradoVehiculo, y=Proyecto1$ValorPrimaAnual)
```

```
# Regresion lineal
library(MASS)
library(ISLR)
lm.fit <- lm(ValorPrimaAnual~ValorAsegurado, data = Proyecto1)
summary(lm.fit)
(lm.fit)
coef(lm.fit)

# Calculamos intervalos
confint(lm.fit)
predict(lm.fit,data.frame(ValorAsegurado=(c(5,10,15))),
        interval="prediction")

# Validamos los supuestos
par(mfrow=c(2,2))
plot(lm.fit)
lm.fit <- lm(ValorPrimaAnual~ValorAseguradoVehiculo, data = Proyecto1)
summary(lm.fit)
(lm.fit)
coef(lm.fit)

# Calculamos intervalos
confint(lm.fit)
predict(lm.fit,data.frame(ValorAseguradoVehiculo=(c(5,10,15))),
        interval="prediction")
# Validamos los supuestos
par(mfrow=c(2,2))
plot(lm.fit)
```

```
# Graficamos el modelo
par(mfrow=c(1,2))
plot(Proyecto1$ValorAsegurado,Proyecto1$ValorPrimaAnual)
abline(lm.fit)
abline(lm.fit,lwd=3)
abline(lm.fit,lwd=3,col="purple")
plot(Proyecto1$ValorAseguradoVehiculo,Proyecto1$ValorPrimaAnual)
abline(lm.fit)
abline(lm.fit,lwd=3)
abline(lm.fit,lwd=3,col="orange")


# REGRESION LINEAL MULTIPLE
lm.fit <- lm(ValorPrimaAnual~ValorAsegurado+
             ValorAseguradoVehiculo, data = Proyecto1)
summary(lm.fit)
lm.fit <- lm(ValorPrimaAnual~., data=Proyecto1)
summary(lm.fit)

lm.fit <- lm(ValorPrimaAnual~ValorAsegurado
             +ValorAseguradoVehiculo+TipoPoliza,
             data = Proyecto)
summary(lm.fit)
```