Ciencia de Datos Aplicada al Transporte

Curso de Complementación

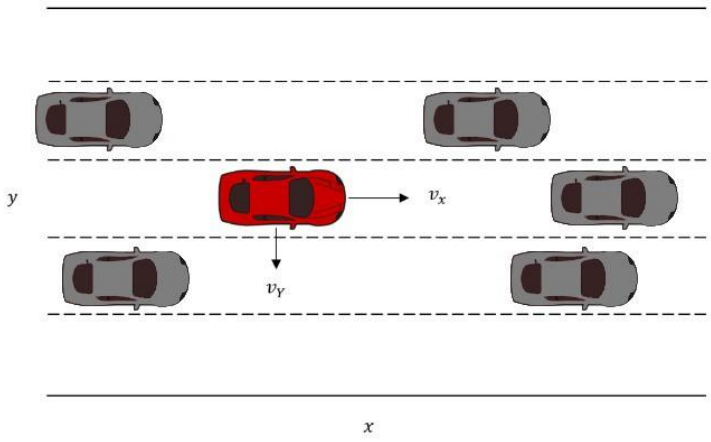
Trabajo Práctico N° 1: Modelo Lineal

Integrantes

Lucia Cebreiros
Mateo Mastelli

El dataset car-following trajectory.csv con el que se va a trabajar en el siguiente trabajo práctico cuenta con 120 observaciones de las siguientes variables:

- v Vel: velocidad en m/s del vehículo de referencia (ego) v Acc: aceleración del vehículo de referencia en m/s2
- Space Headway: distancia en metros entre el vehículo de referencia y el auto de enfrente
- Preceding Distance: Distancia en movimiento con respecto al auto de enfrente en la medición anterior
- Following: un indicador que vale 1 si el vehículo de referencia está siendo seguido por otro vehículo.
- Local Y diff: variación de distancia (en metros) del vehículo de referencia.



1) Cargar los datos del archivo. La variable Following es una variable categórica donde el 1 indica si el auto esta siguiendo y 0 si no. Transformarla en un factor. Finalmente revisar que todas las variables contenidas en el dataframe estén correctamente definidas.

```
#Cargamos los datos
data<- read.csv("car-following_trajectory.csv")

#transformamos la variable vs que es categorica
data$Following <- as.factor(data$Following)
```

```
Data
data                    120 obs. of 6 variables
    $ v_Vel             : num  4.33 4.34 3.75 3.17 3.58 ...
    $ v_Acc             : num  0.1692 -0.0302 -1.3567 0.2131 1.4841 ...
    $ Space_Headway     : num  19.3 18.3 17.4 17.3 18.1 ...
    $ Preceding_Distance: num  3.17 3.42 4.1 4.11 3.94 ...
    $ Following         : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 2 2 2 2 ...
    $ Local_Y_diff      : num  4.38 4.16 3.33 3.19 4.25 ...
```

Se comprueba que el resto de las variables están correctamente categorizadas.

2) Se desea ajustar un modelo de regresión múltiple para predecir la variable Local_Y _dif f en función del resto de las variables en el data set. Escribir el modelo propuesto, indicando los supuestos del mismo.

Se plantea el siguiente modelo lineal que relaciona a la variable Local_Y_dif con el resto de las variables del dataset:

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \beta_5 * x_5 + \varepsilon$$

Donde:

y: $Local\ Y\ diff$

$\beta_0, .., \beta_5$: $Parametros\ del\ modelo$

x_1: $v\ Vel$

x_2: $v\ Acc$

x_3: $Space\ Headway$

x_4: $Preceding\ Distance$

x_5: $Following$

ε: $Error\ no\ observable\ del\ modelo$

Considerando que se cuenta con 120 observaciones de la muestra tendremos

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4} + \beta_5 * x_{i5} + \varepsilon_i\ , \qquad i = 1, \dots, 120$$

Supuestos del modelo:

a) Los errores ε_i tienen media cero. Lo cual implica $E(\varepsilon)$= 0
b) Los errores ε_i tienen todos la misma varianza, $var(\varepsilon_i) = \sigma^2$ (supuesto de homocedasticidad)
c) Los errores ε_i tienen distribución normal
d) Los errores ε_i son independientes entre sí y no están correlacionados con las covariables X_i

Esto implica $\varepsilon_i \sim N(0, \sigma^2)$ independientes para cada $i = 1, \dots, 120$

Es imposible estimar el valor de Y dado que nunca se podrán conocer los errores no observables, por lo tanto, lo que se busca estimar es la esperanza de Y

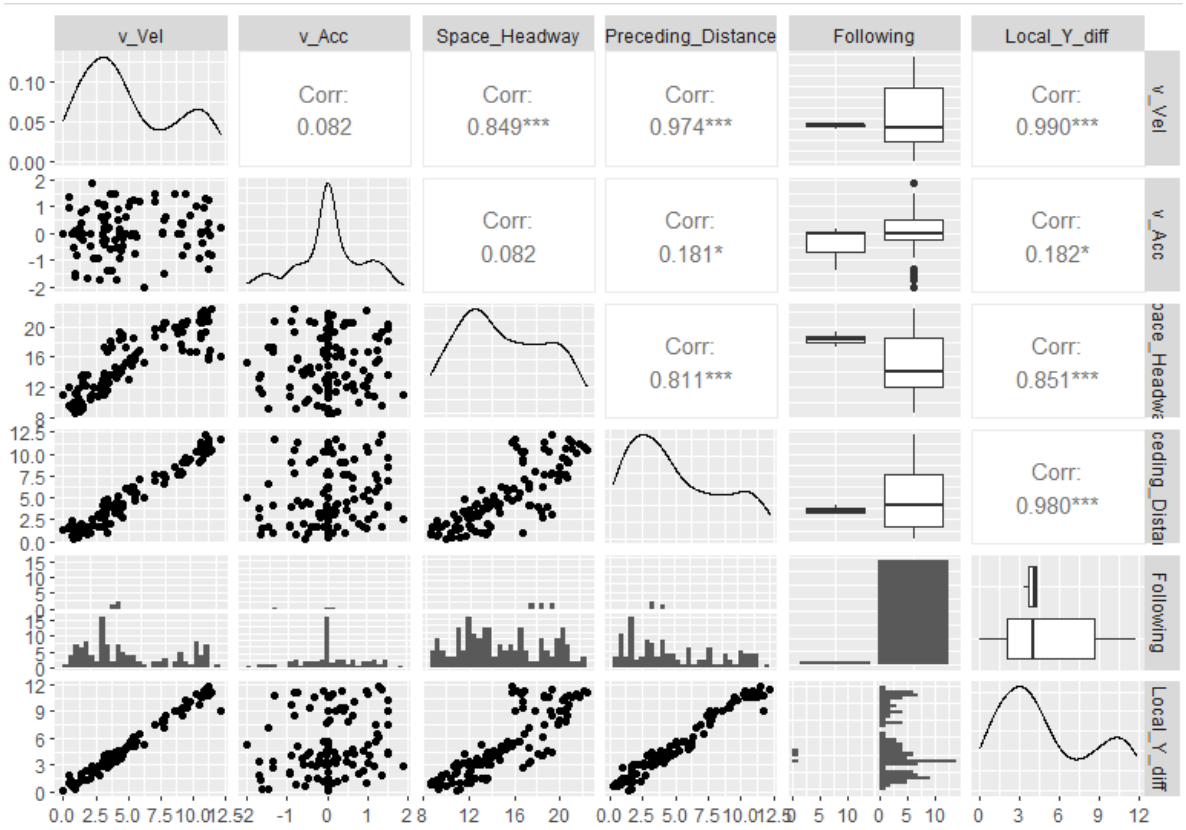
$$\hat{Y} = E(\widehat{y|X = x}) = \widehat{\beta_0} + \widehat{\beta_1} * x_1 + \widehat{\beta_2} * x_2 + \widehat{\beta_3} * x_3 + \widehat{\beta_4} * x_4 + \widehat{\beta_5} * x_5$$

Donde:

$\widehat{\beta_0}, .., \widehat{\beta_5}$: $Estimaciones\ de\ los\ parametros\ del\ modelo$

3) Realizar un scatterplot de las variables con la función ggpairs.

Utilizando la función GGpair, perteneciente a la librería de R "GGally", se obtiene el siguiente gráfico:



Se observa que las variables V_vel, Preceding_Distance y Space_Headway guardan una fuerte correlación lineal con la variable objetivo Local_Y_diff.
Por otro lado, también se observa que las variables V_vel, Preceding_Distance y Space_Headway guardan un alto grado de linealidad entre sí.

4) A partir de la tabla de correlaciones estimadas entre las variables, si tuviera que elegir una sola variable para proponer un modelo de regresión simple, ¿Cuál elegiría y por qué?

En caso de tener que plantearse una regresión lineal con una única variable, se utilizaría la variable V_vel debido a que es la que posee el coeficiente de correlación lineal mas elevado respecto de la variable Local_y_Diff.

5) Realizar un ajuste de regresión lineal múltiple. ¿Es la regresión significativa? Especificar las hipótesis nula y alternativa de este test. ¿Como se calcula el p-valor en este caso? ¿Rechazaría a un nivel de significación de 0.05?.

Se realiza un ajuste de regresión lineal múltiple a través de la función "lm" de R, la cual minimiza los residuos a través de mínimos cuadrados.

```
reg <-lm(Local_Y_diff~., data=data)
```

Luego, se aplica la función “sumary” a la regresión para obtener toda la información que el R nos puede brindar sobre la misma

```
summary(reg)


Residuals:
     Min       1Q   Median       3Q      Max
-2.76555 -0.10690 -0.00464  0.11863  0.72084

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.50082    0.29905  -1.675 0.096727 .
v_Vel               0.79941    0.04575  17.472  < 2e-16 ***
v_Acc               0.38317    0.04296   8.918  9.1e-15 ***
Space_Headway       0.04273    0.01585   2.696 0.008084 **
Preceding_Distance  0.14552    0.04285   3.396 0.000942 ***
Following1          0.13903    0.20282   0.685 0.494427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3242 on 114 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9915
F-statistic:  2766 on 5 and 114 DF,  p-value: < 2.2e-16
```

Para saber si la regresión es significativa se plantea un test de significación en el cual se busca comprobar que el modelo elegido representa mejor a la variable objetivo que considerar el promedio de la misma.
De esta manera las hipótesis que se plantean en dicho test son las siguientes:

- H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- H_1: $Algun\ \beta\ es\ disntito\ de\ 0$

Nos interesa rechazar H_0.

Para poder afirmar si se rechaza o no la hipótesis, el R nos informa el p-valor. Este parámetro representa la probabilidad de que el estadístico del test sea más grande que el valor observado.

$$p - valor = P(F > f_{obs})$$

Comúnmente se busca poder rechazar H_0 con una seguridad de al menos el 95%. Dado que el p-valor es muy pequeño podemos rechazar la hipótesis y afirmar que la regresión es significativa.

6) A partir de la tabla de coeficientes estimados, ¿Qué variables resultan significativas? ¿A qué nivel? ¿Cuál es el valor de la estimación para σ^2? Especificar las hipótesis nulas y alternativas para alguno de los test t reportados en la tabla, el estadístico del test y la regla de decisión. ¿Cómo se calcula el p-valor para este test?

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.76555 -0.10690 -0.00464  0.11863  0.72084

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.50082    0.29905  -1.675 0.096727 .
V_Vel               0.79941    0.04575  17.472  < 2e-16 ***
V_Acc               0.38317    0.04296   8.918  9.1e-15 ***
Space_Headway       0.04273    0.01585   2.696 0.008084 **
Preceding_Distance  0.14552    0.04285   3.396 0.000942 ***
Following1          0.13903    0.20282   0.685 0.494427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.3242 on 114 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9915
F-statistic:  2766 on 5 and 114 DF,  p-value: < 2.2e-16
```

En azul se informa el valor de la estimación para σ^2 = 0,3242.
Al momento de hacer la regresión línea múltiple, R también realiza un test de hipótesis para cada variable por separado y nos informa su grado de significación. Pudiendo observar en el recuadro rojo que las variables V_vel, V_acc, Preceding_Disctance y Space_Headway resultan significativas.
Dichos p-valores se obtienen de plantear el siguientes test de hipótesis para cada variable:

- H_0: $\beta_i = 0$
- H_1: $\beta_i \neq 0$

El estadístico utilizado en el test es el siguiente:

$$T = \frac{\widehat{\beta}_i}{\sqrt{Var(\widehat{\beta}_i)}} \sim t_{n-p}$$

Donde:

n: *Cantidad de observaciones*

p: *Canitdad de variables*

Si H_0 es verdadero se puede asumir que el estadístico tiene distribución t student de n – p grados de libertad.

Finalmente se rechaza H_0 si $|T| > k$ siendo $k\ (nivel\ de\ confianza) = 1 - \alpha/2$ donde alfa es el nivel de significación.

7) Evaluar la bondad del ajuste realizado, a través del coeficiente de determinación. Indicar cuánto vale y que significa.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.76555 -0.10690 -0.00464  0.11863  0.72084

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.50082    0.29905  -1.675 0.096727 .
v_Vel               0.79941    0.04575  17.472  < 2e-16 ***
v_Acc               0.38317    0.04296   8.918  9.1e-15 ***
Space_Headway       0.04273    0.01585   2.696 0.008084 **
Preceding_Distance  0.14552    0.04285   3.396 0.000942 ***
Following1          0.13903    0.20282   0.685 0.494427
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3242 on 114 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9915
F-statistic:  2766 on 5 and 114 DF,  p-value: < 2.2e-16
```

$$R^2 = 0.9918\,; \quad {R_a}^2 = 0.9915 \qquad (valores\ casi\ perfectos)$$

El coeficiente de determinación (R^2) es una medida de la capacidad de ajuste del modelo. En otras palabras, mide cual es el porcentaje de la variabilidad de Y explicada por el modelo. Si R^2 está cerca de 1 significa que el modelo propuesto aporta para explicar dicha variabilidad.

$$R^2 = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} = \frac{SCR}{SCT}$$

Donde:

SCR: *Suma de Cuadrados de los Residuos*

SCT: *Suma de Cuadrados Totales*

Este coeficiente tiene un problema. A medida que se agregan covariables al modelo, el valor de R^2 se ve aumentado, por lo tanto es una medida que no sirve para comparar modelos que tengan diferente cantidad de variables.
Para solucionar este inconveniente se define $R^2\ ajustado$

$${R_a}^2 = 1 - \frac{n-1}{n-p}\frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} = 1 - (1 - R^2)\frac{n}{n-p}$$

Divide a cada suma de cuadrados por sus grados de libertad, y esta medida aumenta solamente si la variable que agregamos mejora el modelo.

8) Validar los supuestos expresados en el ítem 2 a partir del análisis de los residuos, para el modelo seleccionado. ¿Observa algo extraño en los gráficos? ¿Qué propone?

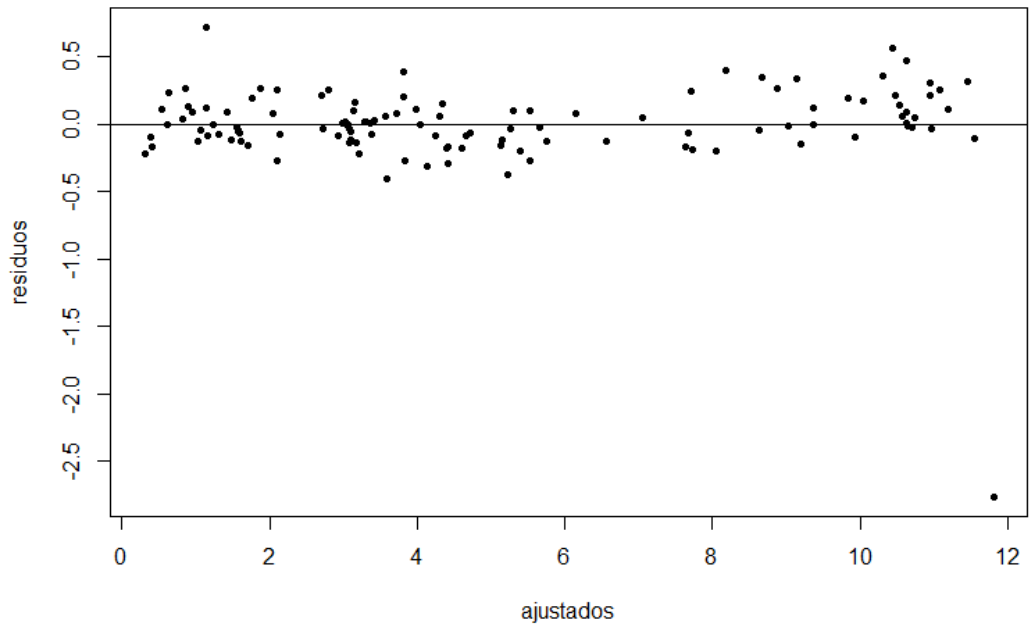
Supuestos del modelo:

a) Los errores ε_i tienen media cero. Lo cual implica $E(\varepsilon)= 0$
b) Los errores ε_i tienen todos la misma varianza, $var(\varepsilon_i) = \sigma^2$ (supuesto de homocedasticidad)
c) Los errores ε_i tienen distribución normal
d) Los errores ε_i son independientes entre sí y no están correlacionados con las covariables X_i

Para verificar estos supuestos en un modelo de regresión lineal múltiple se analizan los residuos.

Si se realiza un gráfico de los residuos y el modelo es válido, no debería observarse ninguna estructura, deberían aparecer puntos distribuidos aleatoriamente al rededor del cero.

```
ajustados<- reg$fitted.values
residuos<- reg$residuals
plot(ajustados, residuos, pch=20)
abline(h=0)
```



Se observan en el gráfico dos problemas. Los residuos siguen una leve forma cuadrática y a demás se observa la existencia de un outlair hacia el final del gráfico. El hecho de que los residuos sigan una leve forma cuadrática es un indicativo de que al modelo planteado le falta un término cuadrático.

Verificamos normalidad. Si todos los supuestos fuesen validos los residuos deberían tener una distribución normal tal que

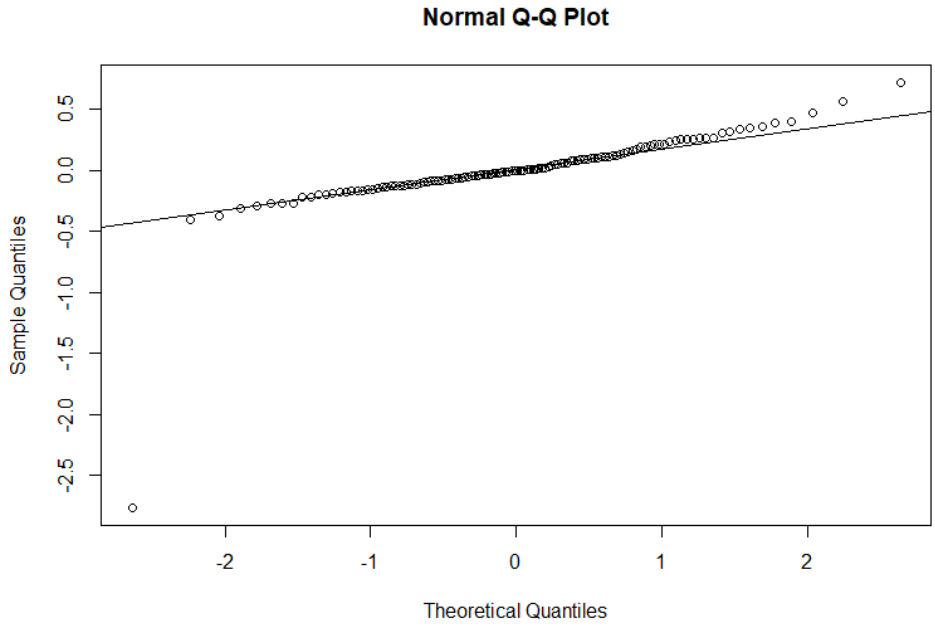
$$\frac{r_i}{\sqrt{\sigma^2(1-p_{ii})}} \sim N(0,1)$$

Donde:

$$P = X(X^T X)^{-1} X^T$$

Una forma de evaluar el supuesto de normalidad de los residuos es a través del método gráfico QQplots. Se gráfica el cuantil observado vs cuantil teórico.

```
qqnorm(residuos)
qqline(residuos)
```



Se observa en el gráfico, además del outlair antes mencionado, que los puntos se despegan en los extremos por lo cual no se cumpliría el supuesto de normalidad. Este supuesto es el menos grave de no cumplir ya que si la muestra fuese infinita tendería a la normalidad.

Para poder solucionar estos problemas se pueden aplicar las transformaciones Box y Cox. Proponen una familia de funciones de potencia para la variable respuesta con el objetivo de garantizar el cumplimiento de lo supuestos. Estas transformaciones combinan el objetivo de encontrar una relación simple con homogeneidad de varianzas, mejorando la normalidad.

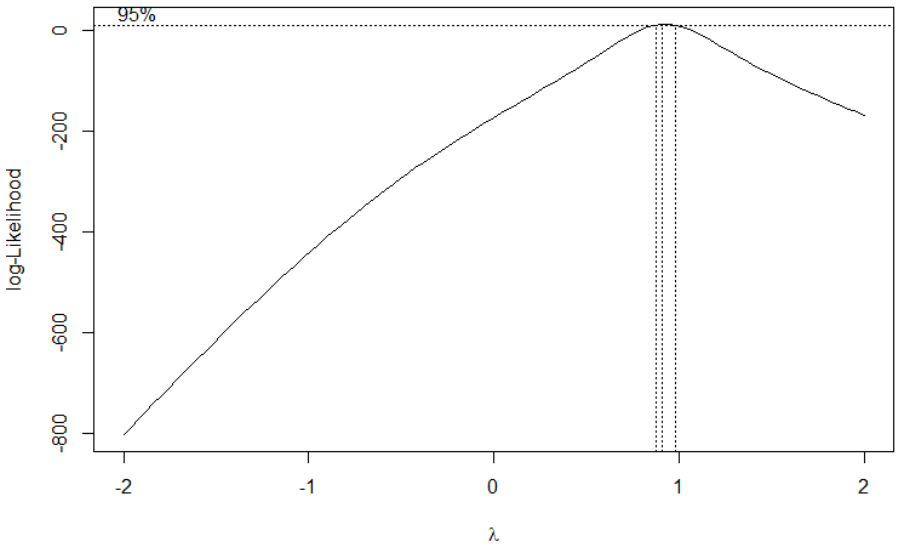
La transformación de Box y Cox está dada por:

$$y^{(\lambda)} \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & si\ \lambda \neq 0 \\ \log(y) & si\ \lambda = 0 \end{cases}$$

Utilizando la función “boxcox” de la Liberia MASS se calcula λ:

```
bc<- boxcox(Local_Y_diff~., data=data)
lamda<- bc$x[which.max(bc$y)]
lamda
```

$$\lambda = 0.9090909$$

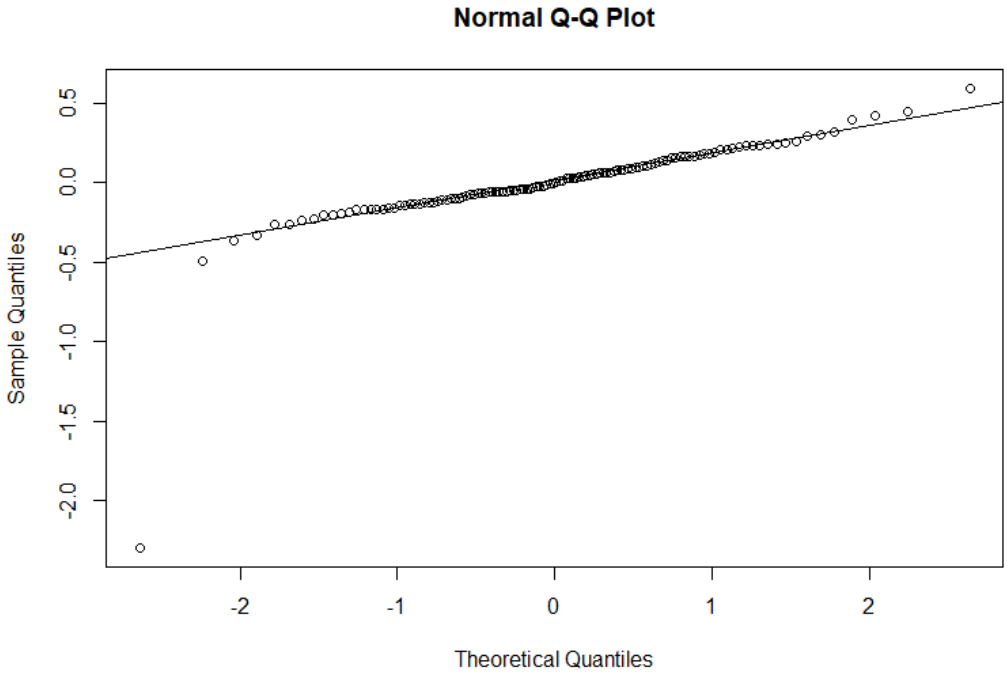


Se procede a realizar nuevamente la regresión lineal múltiple.

```
#nueva variable
y2<-(data$Local_Y_diff^lamda-1)/lamda

#nuevo data set con la variable objetivo reemplazada
data2<-data
data2$Local_Y_diff<-y2

reg2<-lm(Local_Y_diff~., data = data2)
summary(reg2)
residuos2<-reg2$residuals
qqnorm(residuos2)
qqline(residuos2)
points(residuos2,col="blue")
plot(reg2$fitted.values,residuos2$residuals,pch=20)
abline(h=0)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.3001 -0.1019  0.0008  0.1316  0.5972

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.27213    0.25793  -4.932 2.79e-06 ***
v_Vel               0.68508    0.03946  17.361  < 2e-16 ***
v_Acc               0.35167    0.03706   9.490 4.30e-16 ***
Space_Headway       0.04940    0.01367   3.613 0.000452 ***
Preceding_Distance  0.11520    0.03696   3.117 0.002312 **
Following1          0.08613    0.17493   0.492 0.623387
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2796 on 114 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9914
F-statistic:  2754 on 5 and 114 DF,  p-value: < 2.2e-16
```

Se observa una leve desmejora en el ajuste del modelo.



Tampoco se observa una mejora con respecto a la normalidad.

9) ¿Cuál seria la estimación de la esperanza de la variable a predecir para una observación con los siguientes valores: v vel = 4,06, v Acc = 0,0568, Space Headway = 8,575, P receding Distance = 7,62, F ollowing = 1

```
x0 = data.frame("v_Vel"=4.06, "v_Acc"=0.0568, "Space_Headway"=8.575, "Preceding_Distance"=7.62, "Following"= "1")
y0<- predict(reg, newdata=x0)
y0
```

$$y_0 = 4.380859\, m$$

10) Hallar un intervalo de confianza y de predicción de nivel 0.95 para la estimación hallada en el ítem anterior.

```
int<- predict(reg, newdata=x0, interval = "confidence", level = 0.95)
int
```

```
      fit      lwr      upr
4.380859 4.051004 4.710713
```

Por lo tanto, la esperanza de Y para el nuevo punto es 4.380859 m y se encuentra dentro del intervalo de confianza [4.051004 m; 4.710713 m] con un 95% de probabilidad.

```
intP<-predict(reg,, newdata=x0, interval = "prediction", level = 0.95)
intP
```

```
      fit      lwr      upr
4.380859 3.658903 5.102815
```

Por lo tanto, la estimación de Y para el nuevo punto es 4.380859 m y se encuentra dentro del intervalo de predicción [3.658903 m; 5.102815 m] con un 95% de probabilidad.

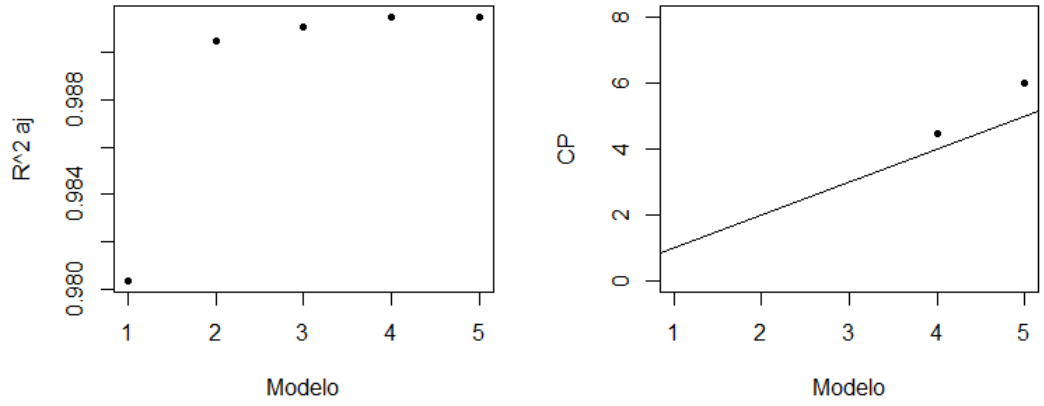
11) Selección de modelos. Plantear un nuevo modelo en el que intervengan aquellas variables que contribuyen significativamente y estimar los parámetros por mínimo cuadrados. ¿Qué modelo elegiría finalmente? Utilizar medidas de bondad de ajuste y de predicción, tal como la estimación del error cuadrático medio de validación cruzada.

Para determinar con cuantas variables nos quedamos para plantear el segundo modelo, utilizamos el método “Exhaustive”. Debido a que nuestro dataset no es muy grande, y que las variables del modelo son pocas podemos usarlo antes que los métodos forward y backward, los cuales son menos precisos pero computacionalmente más eficientes.

A través de la librería “Leaps” se utiliza la función “Exhaustive”.

```
attach(data)
x<-cbind(v_Vel,v_Acc,Space_Headway,Preceding_Distance,Following)

exhaustive<-regsubsets(Local_Y_diff~x,data = data, method = "exhaustive")
summary(exhaustive)
names(summary(exhaustive))

par(mfrow=c(1,2))
plot(summary(exhaustive)$adjr2,pch=20,xlab="Modelo", ylab= "R^2 aj")
plot(1:5,summary(exhaustive)$cp,pch=20,ylim=c(0,8),xlab="Modelo", ylab= "CP")
abline(0,1)
par(mfrow=c(1,1))
```

```
Call: regsubsets.formula(Local_Y_diff ~ x, data = data, method = "exhaustive")
5 Variables  (and intercept)
                     Forced in Forced out
xv_Vel                   FALSE      FALSE
xv_Acc                   FALSE      FALSE
xSpace_Headway           FALSE      FALSE
xPreceding_Distance      FALSE      FALSE
xFollowing               FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
          xv_Vel xv_Acc xSpace_Headway xPreceding_Distance xFollowing
1  ( 1 ) "*"    " "    " "             " "                 " "
2  ( 1 ) "*"    "*"    " "             " "                 " "
3  ( 1 ) "*"    "*"    " "             "*"                 " "
4  ( 1 ) "*"    "*"    "*"             "*"                 " "
5  ( 1 ) "*"    "*"    "*"             "*"                 "*"
```



En conclusión, el modelo que mejor ajusta es el que utiliza 4 variables ya que tiene el mayor coeficiente de determinación y el cp de Mallows que más cerca está de la recta identidad.

Se plantea un segundo modelo únicamente con las variables que contribuyen significativamente para estimar a la variable objetivo, las cuales son variables V_vel, V_acc, Preceding_Disctance y Space_Headway.

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_4 + \varepsilon$$

Donde:

y: *Local Y diff*

$\beta_0, .., \beta_4$: *Parametros del modelo*

x_1: *v Vel*

x_2: *v Acc*

x_3: *Space Headway*

x_4: *Preceding Distance*

ε: *Error no observable del modelo*

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.7834 -0.1141 -0.0059  0.1249  0.7171

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.32935    0.16349  -2.014  0.04629 *
v_Vel               0.80211    0.04548  17.637  < 2e-16 ***
v_Acc               0.38555    0.04273   9.024 4.86e-15 ***
Space_Headway       0.03911    0.01491   2.623  0.00989 **
Preceding_Distance  0.14641    0.04273   3.426  0.00085 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3234 on 115 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9915
F-statistic:  3473 on 4 and 115 DF,  p-value: < 2.2e-16
```

Dado que el $R_a{}^2 = 0.9915$ es igual que para el modelo de regresión con todas las variables y que , el modelo ganador resulta este último por contar con menos variables y ser menos complejo.

Otra medida de bondad de predicción es la estimación del ECM usando validación cruzada.

$$\widehat{ECM} = \frac{1}{n}\sum_{i=1}^{n} \frac{r_i{}^2}{(1-P_{ii})^2}$$

Donde:

$$P = X(X^T X)^{-1} X^T$$

```
P<- x%*%inv(t(x)%*%x)%*%t(x)
n = length(data$v_Vel)
r = reg$residuals

a = 0
for (i in n) {
a = a + r[i]^2/(1-P[i,i])
  }
ECM1 <- a/n
ECM1
```

$$\widehat{ECM}1 = 0.06997383$$

```
r2 = reg2$residuals
a = 0
for (i in n) {
  a = a + r2[i]^2/(1-P[i,i])
}

ECM2 <- a/n
ECM2
```

$$\widehat{ECM2} = 0.0484009$$

El error cuadrático medio comprueba que el modelo 2 se ajusta mejor a la variable objetivo.