DIAGNÓSTICOS GRÁFICOS Y FORMALES EN MODELOS DE REGRESIÓN LINEAL MÚLTIPLE

ISADORE NABI

I. CASO DE APLICACIÓN: VELOCIDAD DE NADO

Se realizó un estudio para analizar la velocidad de nado de las personas mayores de 18 años que son miembros regulares de un equipo de natación, y se tomaron en cuenta algunas variables que pueden estar relacionadas con esta velocidad. Se hizo

una prueba a los participantes y se tomó el tiempo que duraban en nadar 50m. Entonces como medida de la velocidad de nado se tiene el tiempo (en segundos) el cual se puede transformar en la variable velocidad dividiendo la distancia entre el tiempo. Esta variable se llama 'veloc'. Como variables predictoras se tienen las siguientes:

a. 'edad': la edad en años cumplidos.
b. 'sexo': el sexo codificado como 0 (mujeres) y 1 (hombres).
c. 'imc': el índice de masa corporal se calcula dividiendo el peso en kilogramos entre la altura al cuadrado (en metros), lo cual da una medida en kg/m^2.
d. 'pierna': la longitud promedio de ambas piernas (en centímetros).
e. 'brazo': la longitud promedio de ambos brazos (en centímetros).

```
load("velo.Rdata")
base$sexo=factor(base$sexo)
levels(base$sexo)=c("Mujer","Hombre")
attach(base)
```

I.I. Distribución Normal de los Errores: Obtención de Residuales del Modelo Clásico de Regresión Lineal Sin Interacciones

Puede construirse un modelo de regresión lineal empleando la totalidad de predictores (5) sin interacciones y obtener los residuales.

```
mod = lm(veloc ~ .,base)
mod

##
## Call:
## lm(formula = veloc ~ ., data = base)
##
## Coefficients:
```

```
## (Intercept)        edad   sexoHombre         imc      pierna       brazo
##    0.369735    -0.020189     0.105953     0.014021    -0.004222     0.021599

res = mod$res
res

##           1           2           3           4           5           6
##  0.059924767 -0.056329998 -0.121512457 -0.109358062  0.033936013 -0.029045764
##           7           8           9          10          11          12
## -0.180211206  0.044475772  0.269796920  0.168383939  0.200313351 -0.073747198
##          13          14          15          16          17          18
##  0.076747471 -0.344571772  0.096145727 -0.238829190  0.074175316 -0.109493134
##          19          20          21          22          23          24
##  0.068539673  0.114550726  0.052683533 -0.138445798  0.105789134 -0.123258343
##          25          26          27          28          29          30
##  0.005499153 -0.045992965  0.082257036 -0.180765874  0.263029172  0.035314058
```

A continuación, se detalla el procedimiento antes hecho relativo al cálculo de los valores residuales en el escenario en que existen variables categóricas, que permite verificar que es equivalente al mismo cálculo en el escenario en el que no existe tal clase de variables.

Para esto se realizará el cálculo del residual para un hombre específico y también para una mujer específica, por ejemplo, los individuos 1 (Mujer) y 6 (Hombre). Estos resultados pueden ser comparados con los obtenidos con la sintaxis anteriormente expuesta.

```
#HOMBRE
#Velocidad del individuo 6 (hombre)
y6=veloc[6]
#Predicción de la velocidad del individuo 6 (hombre) con las características de la
observación 6 (21 años, hombre, imc de 21.29529, longitud de pierna de 103.5 cm y
```

```
longitud de brazo de 74.75 cm)
yh6=predict(mod,data.frame(edad=21,sexo="Hombre",imc=
21.29529,pierna=103.5,
brazo=74.75))
#Diferencia entre el valor observado y el valor esperado o predicción
y6-yh6

##         1
## -0.0290457

#Comparación con el residual obtenido de forma automatizada
res[6]

##          6
## -0.02904576

#MUJER
#Velocidad del individuo 6 (hombre)
y1=veloc[1]
#Predicción de la velocidad del individuo 1 (mujer) con las características de la observación
1 (22 años, mujer, imc de 21.14769, longitud de pierna de 100.5 cm y longitud de brazo de
72 cm)
ym1=predict(mod,data.frame(edad=22,sexo="Mujer",imc= 21.14769,pierna=100.5,
brazo=72))
#Diferencia entre el valor observado y el valor esperado o predicción
y1-ym1

##         1
## 0.05992478

#Comparación con el residual obtenido de forma automatizada
res[1]
```

```
##         1
## 0.05992477
```

Si la variable de respuesta (velocidad) tiene una distribución normal condicional a las X, es decir, para cada combinación de los predictores debe haber una distribución normal resultante, ¿por qué basta revisar que los errores en conjunto tengan una distribución normal? La razón es porque los errores tienen también una distribución normal puesto que provienen del valor de Y menos su media condicional, con media 0[1] y variancia constante para cada X. Entonces al tener cada uno de ellos una distribución exactamente igual, es posible combinarlos y analizar la distribución de todos ellos en conjunto, es decir, la distribución global de tales errores. A causa de asumir el supuesto de homogeneidad de varianza (homocedasticidad), esta combinación de residuales sólo es válida cuando se cumpla dicho supuesto, de otra forma se podría estar incurriendo en un error al tratar de analizar un fenómeno como si se distribuyese normalmente cuando pudiese existir un problema de heteroscedasticidad.

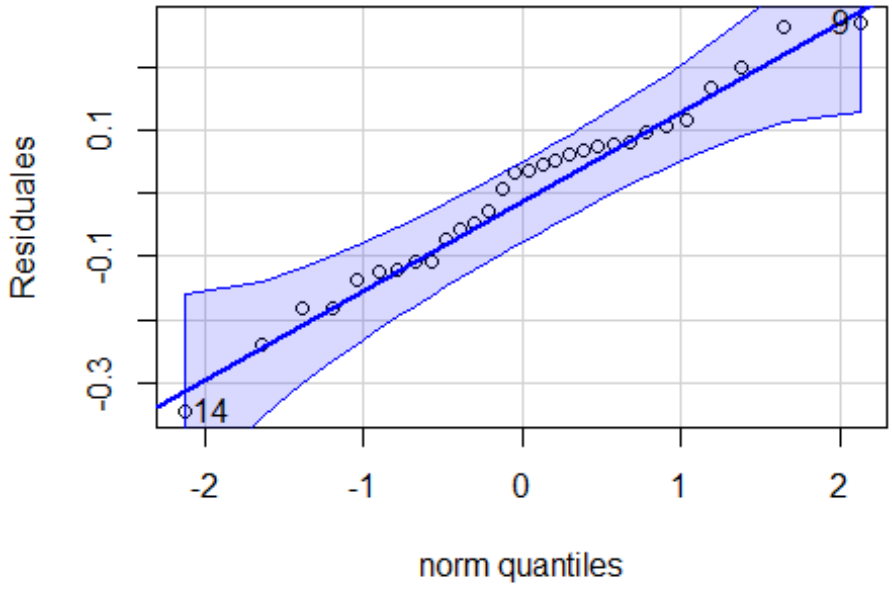
Es posible construir el gráfico cuantiles-cuantiles conocido como gráfico Q-Q para ver analizar gráficamente si los residuales de este modelo provienen de una distribución normal. La función qqPlot de la librería 'car' permite obtener bandas de confianza. Utilícese esta función como qqPlot(res).

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(res,ylab="Residuales")
```

[1] Que su media sea nula se explica por dos razones. La primera es debido a que la distribución normal es simétrica, entonces existen tantas observaciones hacia un lado como hacia el otro

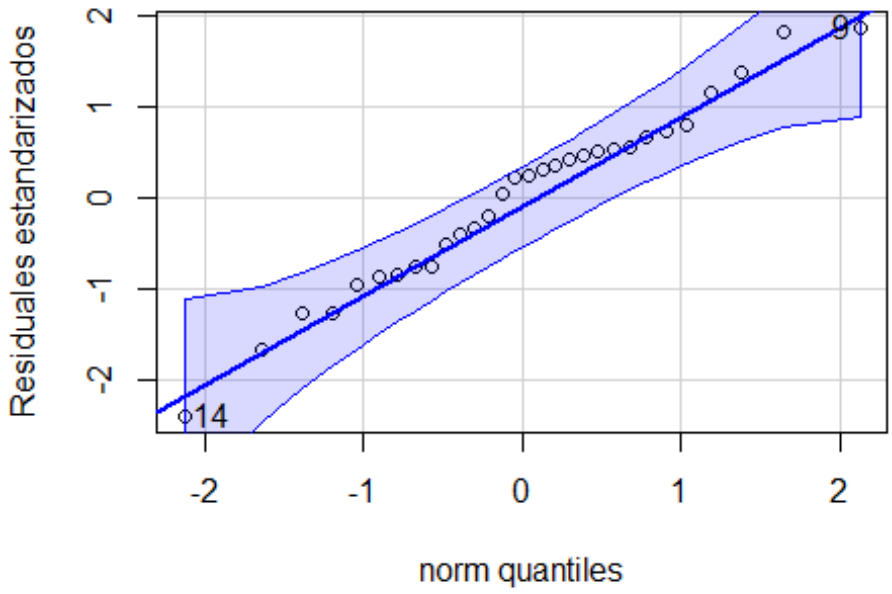



```
## [1] 14  9
```

Ya que al comparar los residuales con los cuantiles teóricos de una distribución normal todos ellos se colocan muy cerca de una línea recta, se observa preliminarmente evidencia de proporcionalidad entre ambos, con lo cual se puede asumir que la distribución que los generó es normal.

Adicionalmente, es posible construir el gráfico anterior comparando ahora los cuantiles teóricos con los residuales estandarizados. Para ello, estos deben estandarizarse con la sintaxis scale(res).

```
qqPlot(scale(res),ylab="Residuales estandarizados")
```

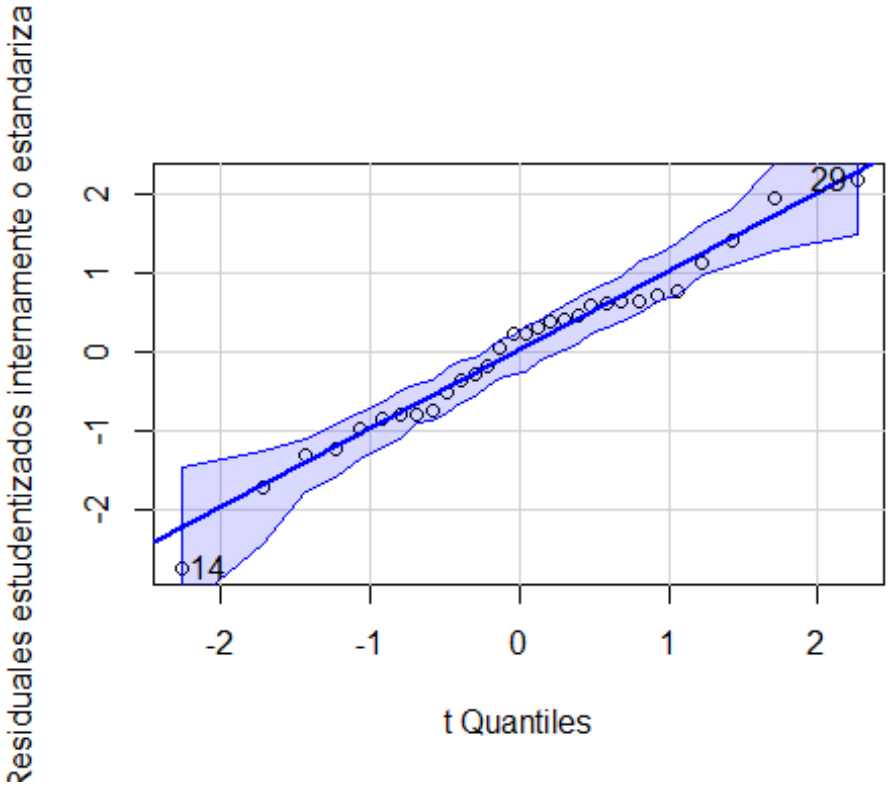



```
## [1] 14  9
```

Se observa que al estandarizar los residuales (conocido este proceso también como studentización interna de los residuales) se presenta el mismo comportamiento, salvo la obviedad del cambio de escala en el eje Y.

Es posible graficar automáticamente los residuos estandarizados. Para ello, basta con utilizar la sintaxis anterior estableciendo como único argumento el modelo mod e indicando mediante "ylab" el título del gráfico a generar.

```
qqPlot(mod,ylab="Residuales estudentizados internamente o estandarizados")
```




```
## [1] 14 29
```

Al estudentizar los residuales se observa un patrón lineal, aunque al hacer esto cambia la relación observada en el sentido de que ahora se asegura que hay una varianza condicional constante en los residuales. Se debe recordar que la estudentización sólo es válida si hay homoscedasticidad.

I.II. Verificación de Homocedasticidad

```
knitr::include_graphics("FIG1.JPG")
```

FIGURA 3.4
Homoscedasticidad.

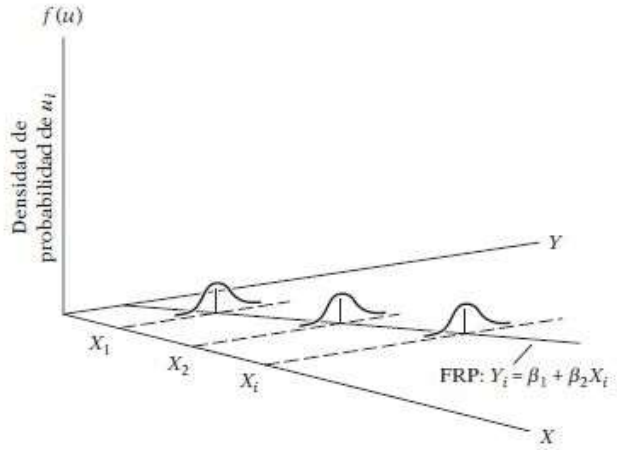
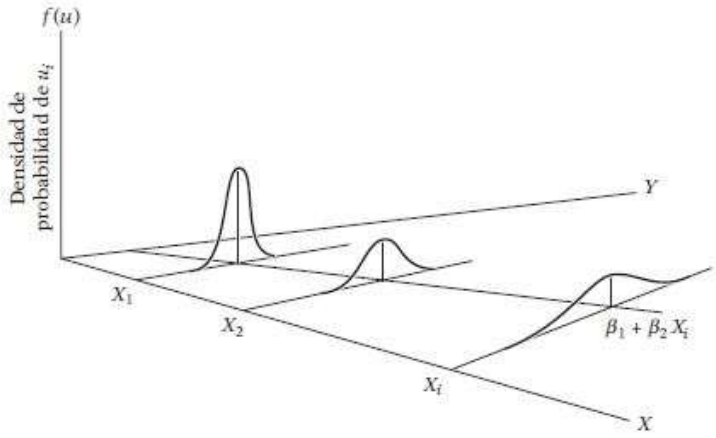


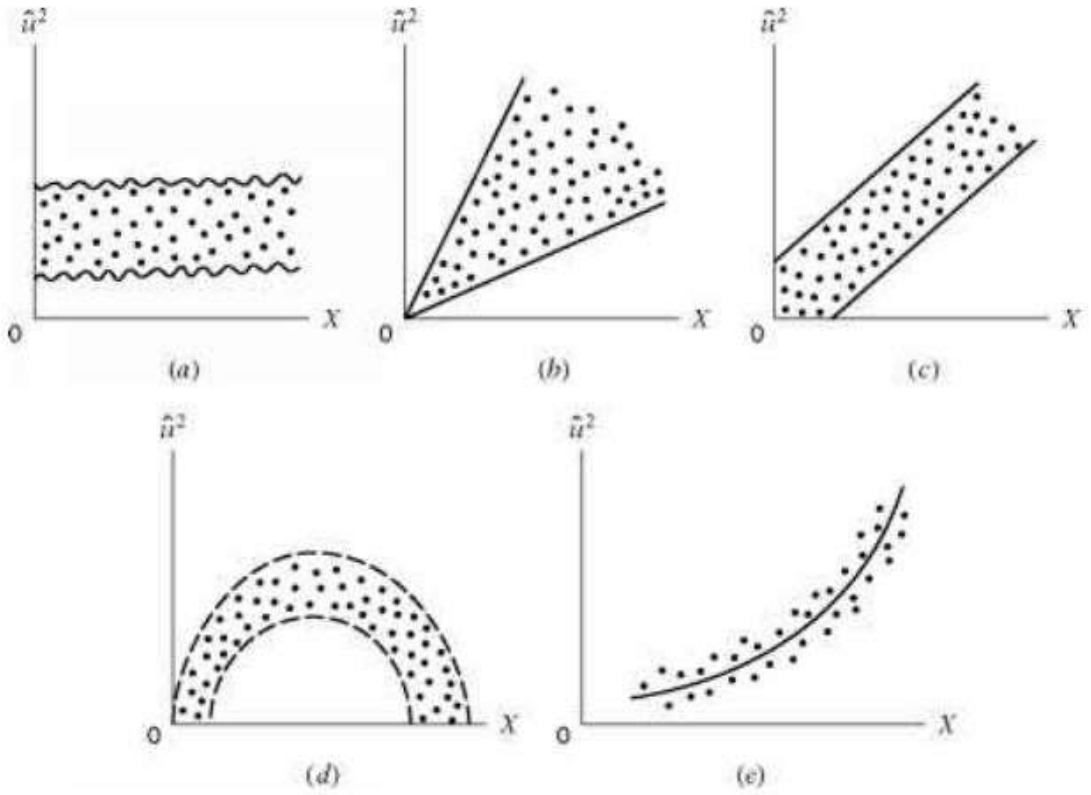
FIGURA 3.5
Heteroscedasticidad.



#*Figura 1: Homoscedasticidad vs Heteroscedasticidad*

#*Fuente:* (Gujarati & Porter, 2010, pág. 65).

```
knitr::include_graphics("FIG2.JPG")
```

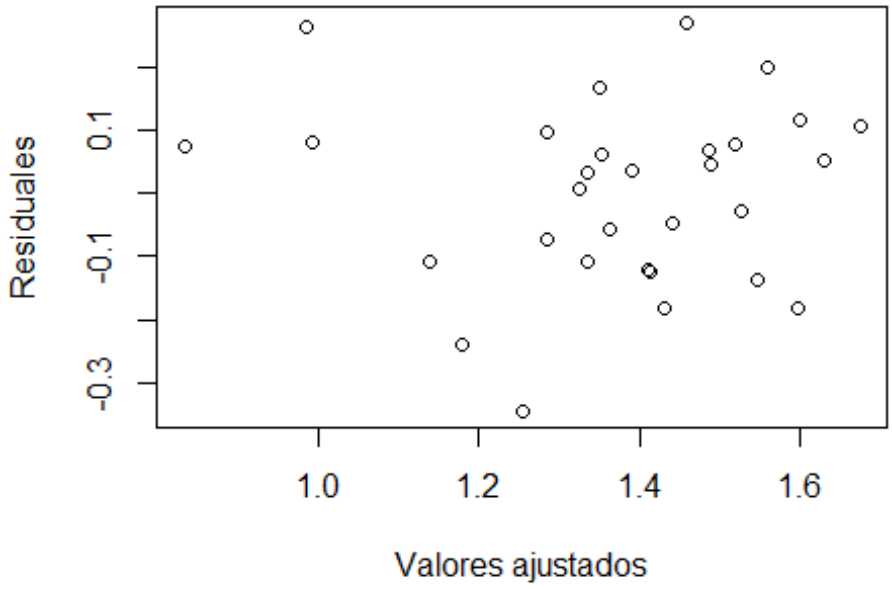


#*Figura 2: Patrones Heteroscedásticos de los Residuos*

#*Fuente:* (Hayden Economics, 2022).

Es posible indagar gráficamente en el conjunto de datos, a través de los valores ajustados contra los residuales, para determinar si existe algún patrón en el comportamiento de los residuos que indique presencia de heterocedasticidad.

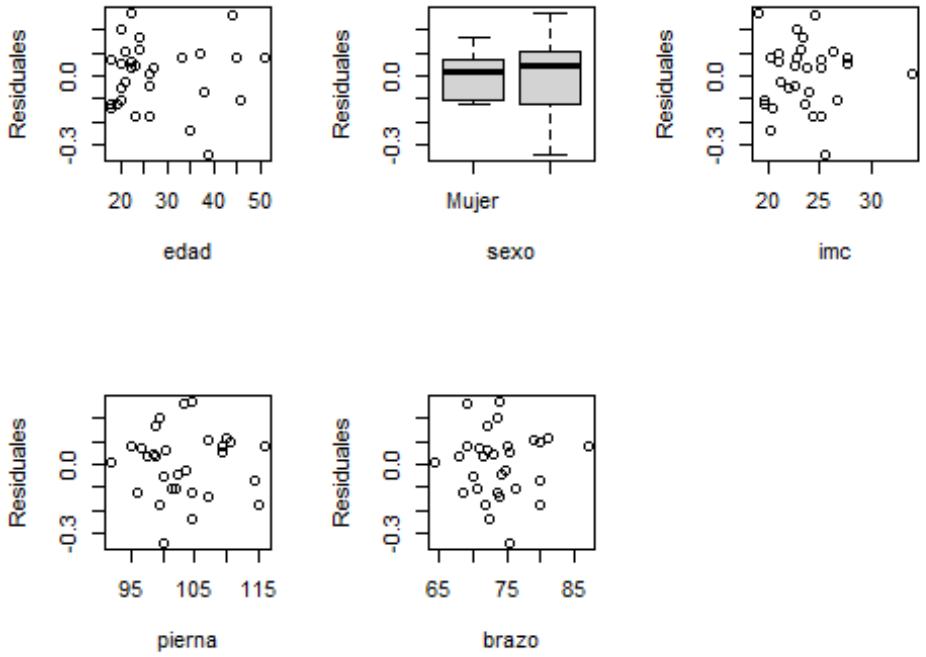
```
plot(mod$fit,mod$res,xlab="Valores ajustados",ylab="Residuales")
```



Bajo la lógica de la figura 2, es posible afirmar que los residuales parecen seguir un comportamiento aleatorio sin presentar ningún patrón, por lo que se puede esperar que se cumpla el supuesto de homoscedasticidad.

Adicionalmente, puede elaborarse un gráfico de los residuales contra cada variable para tratar de observar si hay algún patrón. En el caso de variables categóricas, debe hacerse un boxplot y observar si la amplitud de todas las cajas difiere.

```
par(mfrow=c(2,3))
for(i in 2:6) plot(base[,i],mod$res,xlab=names(base)[i],ylab="Residuales")
par(mfrow=c(1,1))
```



En todos los gráficos con variables continuas se observa un comportamiento de los residuales que podría considerarse como aleatorio, *i.e.*, sin presencia de patrones. Sin embargo, en el boxplot residuales-sexo se ven cajas con amplitudes que claramente no son equivalentes, aunque guardan cierta similitud. En general, lo anterior refuerza la posibilidad, sugerida por los otros tipos de gráficas de residuales, de que se cumple el supuesto de homoscedasticidad.

También pueden realizarse pruebas formales para detección de heteroscedasticidad. Las pruebas usualmente utilizadas son la prueba de Breusch-Pagan y la prueba de White (conocida también como prueba de Breusch-Pagan con residuos studentizados internamente[2]). Para realizar la prueba de Breusch-Pagan, debe crearse una variable llamada 'r2' que contenga los residuales al cuadrado y construir un nuevo modelo llamado 'modres' con 'r2' como respuesta y la

[2] La diferencia en si los residuos son studentizados interna o externamente radica en si se considera o no la i-ésima observación, respectivamente (Cross Validated, 2014).

totalidad de los predictores del modelo original. Se extrae la suma de cuadrados de regresión de 'modres' y la suma de cuadrados residual del modelo original. Finalmente se calcula el valor:

$$\chi^2 = \frac{SCReg^2/2}{(SCRes/n)^2}$$

```
r2 = res^2
modres = lm(r2 ~ .-veloc,base)
SCReg=sum(anova(modres)[-6,2])
SCRes = anova(mod)[6,2]
n=nrow(base)
chi= (SCReg/2)/((SCRes/n)^2)
chi
```

```
## [1] 5.44904
```

```
1-pchisq(chi,5)
```

```
## [1] 0.3635662
```

Lo anterior puede realizarse de forma automatizada usando la función 'bptest' de la librería lmtest. Para realizar la prueba de Breusch-Pagan debe elegirse que no se estudenticen los residuales (en caso contrario se realizaría la prueba de White) quedando de la siguiente forma: bptest(mod,studentize=F).

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

bptest(mod,studentize=F)

##
##  Breusch-Pagan test
##
## data:  mod
## BP = 5.449, df = 5, p-value = 0.3636

bptest(mod,studentize=T)

##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 6.1381, df = 5, p-value = 0.293
```

Bajo la hipótesis nula de que la variancia condicional de la respuesta es la misma para cada valor de los predictores, se obtiene una probabilidad asociada de 0.36 para la prueba de Breusch-Pagan y de 0.293 para la de White. Con esto se falla en rechazar la hipótesis nula con un alpha de 0.05 y se asume que hay homoscedasticidad.

II. CASO DE APLICACIÓN: GRASA CORPORAL

Se quiere estudiar la relación de grasa del cuerpo con algunos predictores entre 20 mujeres sanas de 25 a 34 años. Se cuenta con medidas de tres partes del cuerpo como predictores de grasa: grosor promedio de los tríceps, circunferencia promedio de los muslos y circunferencia promedio de los brazos.

II.I. Verificación de Multicolinealidad

Lo primero que debe hacerse es obtener las correlaciones entre los predictores.

```
load("grasa.Rdata")
attach(base)
cor(base[,-4])
```

```
##             triceps     muslo antebrazo
## triceps   1.0000000 0.9238425 0.4577772
## muslo     0.9238425 1.0000000 0.0846675
## antebrazo 0.4577772 0.0846675 1.0000000
```

Puede observarse una alta correlación entre muslo y tríceps, una correlación intermedia entre antebrazo y tríceps, mientras que existe muy baja correlación entre muslo y antebrazo.

Pueden ajustarse diferentes modelos con el fin de observar cómo cambian los coeficientes y sus errores estándar al tener una sola variable o varias variables en el modelo.

```
mod1 = lm(grasa ~ triceps)
mod2 = lm(grasa ~ muslo)
mod3 = lm(grasa ~ antebrazo)
mod12 = lm(grasa ~ triceps+muslo)
mod13 = lm(grasa ~ triceps+antebrazo)
mod23 = lm(grasa ~ muslo+antebrazo)
mod123 = lm(grasa ~ triceps+muslo+antebrazo)
summary(mod1)$coef
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -1.4961046  3.3192346 -0.4507378 6.575609e-01
## triceps      0.8571865  0.1287808  6.6561675 3.024349e-06
```

```
summary(mod2)$coef

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -23.6344891  5.6574137 -4.177614 5.656662e-04
## muslo         0.8565466  0.1100156  7.785681 3.599996e-07

summary(mod3)$coef

##               Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 14.6867809  9.0959259 1.6146548 0.1237780
## antebrazo    0.1994286  0.3266297 0.6105649 0.5491202

summary(mod12)$coef

##               Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -19.1742456  8.3606407 -2.2933943 0.03484327
## triceps       0.2223526  0.3034389  0.7327755 0.47367898
## muslo         0.6594218  0.2911873  2.2645969 0.03689872

summary(mod13)$coef

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  6.791627  4.4882871  1.513189 1.486003e-01
## triceps      1.000585  0.1282321  7.802921 5.117394e-07
## antebrazo   -0.431442  0.1766156 -2.442831 2.578645e-02

summary(mod23)$coef

##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -25.99695164  6.9973208 -3.7152723 1.719844e-03
## muslo         0.85088172  0.1124482  7.5668743 7.722182e-07
## antebrazo     0.09602947  0.1613927  0.5950052 5.596775e-01

summary(mod123)$coef
```

```
##             Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) 117.084695  99.782403  1.173400 0.2578078
## triceps       4.334092   3.015511  1.437266 0.1699111
## muslo        -2.856848   2.582015 -1.106441 0.2848944
## antebrazo    -2.186060   1.595499 -1.370142 0.1895628
```

Se ve un gran cambio en el coeficiente de tríceps cuando está solo (0.85) con respecto al modelo que tiene también las otras dos variables (4.3). Lo mismo ocurre con el error estándar, pues pasa de 0.13 a 3.0. También se nota que cuando la variable tríceps está sola su coeficiente es significativo, mientras que en el otro modelo pierde su significancia. De forma similar el coeficiente de la variable muslo cambia cuando entra con las otras dos variables con respecto a cuando la variable muslo entra sola.

Puede estudiarse el aporte de 'triceps' cuando entra sola y cuando entra después de las otras variables.

```
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: grasa
##          Df Sum Sq Mean Sq F value    Pr(>F)
## triceps    1 352.27  352.27  44.305 3.024e-06 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod3,mod13)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: grasa ~ antebrazo
## Model 2: grasa ~ triceps + antebrazo
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 485.34
## 2     17 105.93  1     379.4 60.886 5.117e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod2,mod12)
```

```
## Analysis of Variance Table
##
## Model 1: grasa ~ muslo
## Model 2: grasa ~ triceps + muslo
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     18 113.42
## 2     17 109.95  1    3.4729 0.537 0.4737
```

```
anova(mod23,mod123)
```

```
## Analysis of Variance Table
##
## Model 1: grasa ~ muslo + antebrazo
## Model 2: grasa ~ triceps + muslo + antebrazo
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 111.110
## 2     16  98.405  1    12.705 2.0657 0.1699
```

Se observa que cuando la variable tríceps está sola explica 352.27, cuando entra después de antebrazo explica 379.4, cuando entra después de muslo explica 3.47 y cuando entra después de ambas explica 12.7. Entonces parece que la cantidad que logra explicar se reduce enormemente cuando ha entrado previamente muslo.

Puede estimarse el Factor de Inflación de la Variancia (VIF) de 'triceps' en presencia de las otras dos variables. Para esto, se construye un modelo con 'triceps' como respuesta y las otras variables como predictores. Si se denota como R_j^2 el coeficiente de determinación del modelo que toma la j-ésima variable como respuesta, el VIF se estima como

$$VIF = \frac{1}{1 - R_j^2}$$

.

```
modtriceps=lm(triceps~muslo+antebrazo)
r2trip=summary(modtriceps)$r.sq
viftrip=1/(1-r2trip)
viftrip

## [1] 708.8429
```

Al igual que antes, pero automatizando el proceso, puede encontrarse el VIF para todas las variables usando la función vif en la librería car. Para ello simplemente debe escribirse 'vif(mod)', donde el modelo debe contener todas la variables.

```
library(car)
vif(mod123)

##   triceps     muslo antebrazo
##  708.8429  564.3434  104.6060

vif(lm(grasa ~ triceps + muslo + antebrazo, data=base))

##   triceps     muslo antebrazo
##  708.8429  564.3434  104.6060
```

Se puede esperar que la varianza del coeficiente de triceps sea más de 700 veces cuando está con muslo y antebrazo con respecto a la varianza que tendría si esas

variables no estuvieran correlacionadas con triceps. Similarmente la varianza del coeficiente de muslo es 564 veces la que tendría si no existiese multicolinealidad, mientras que la de antebrazo es de 105 veces. Se observa que el coeficiente que menos se altera es el de antebrazo, sin embargo, de igual manera aumenta mucho su varianza en presencia de esas otras variables.

¿Qué pasaría si solo se consideran dos predictoras a la vez?

```
vif(mod12)

##  triceps    muslo
## 6.825239 6.825239

vif(mod13)

##   triceps antebrazo
##  1.265118  1.265118

vif(mod23)

##     muslo antebrazo
##   1.00722   1.00722
```

Cuando hay sólo dos variables juntas a la vez parece que las varianzas no se incrementan notoriamente, a pesar de haberse visto que triceps y muslo tienen una correlación muy alta.

Un valor del VIF menor a 10 es aceptable en el marco del supuesto de no-multicolinealidad (Bhandari, 2020), (ResearchGate, 2016).

III. CASO DE APLICACIÓN: PRESTIGIO PERCIBIDO

Se analiza el puntaje sobre el prestigio percibido de una selección de tipos de empleos en Canadá (método de Pineo-Porter). Los puntajes fueron determinados

en una encuesta social conducida a mediados de los 60´s. Se tienen datos de algunos predictores que fueron recolectados en el censo:

- **censo**: código de la ocupación en el censo canadiense (es un identificador, no es una variable)
- **presti**: puntaje de prestigio de la ocupación.
- **edu**: promedio de años de educación en 1971.
- **ingre**: promedio de ingreso en 1971 ($).
- **mujer**: porcentaje de mujeres ocupadas.
- **tipo**: tipo de ocupación (3 categorías).

III.I. Análisis Gráfico de Linealidad

```
load("prestigio.Rdata")
attach(base)
```

Puede ajustarse un modelo que contenga los cuatro predictores antes señalados. Nótese que la variable tipo es categórica por lo que no tiene sentido que exista linealidad entre la respuesta promedio y este predictor. Además, hay tres ocupaciones que tienen NA en el tipo de ocupación, esto hace que la regresión se base solamente en los casos que tienen información en todas las variables.

```
mod=lm(presti~ingre + edu + mujer+tipo,base)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: presti
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## ingre      1 14021.6 14021.6 275.6982 < 2.2e-16 ***
## edu        1  9052.8  9052.8 177.9999 < 2.2e-16 ***
## mujer      1    10.4    10.4   0.2039  0.652634
```

```
## tipo       2   583.1   291.5   5.7324  0.004506 **
## Residuals 92  4679.0    50.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

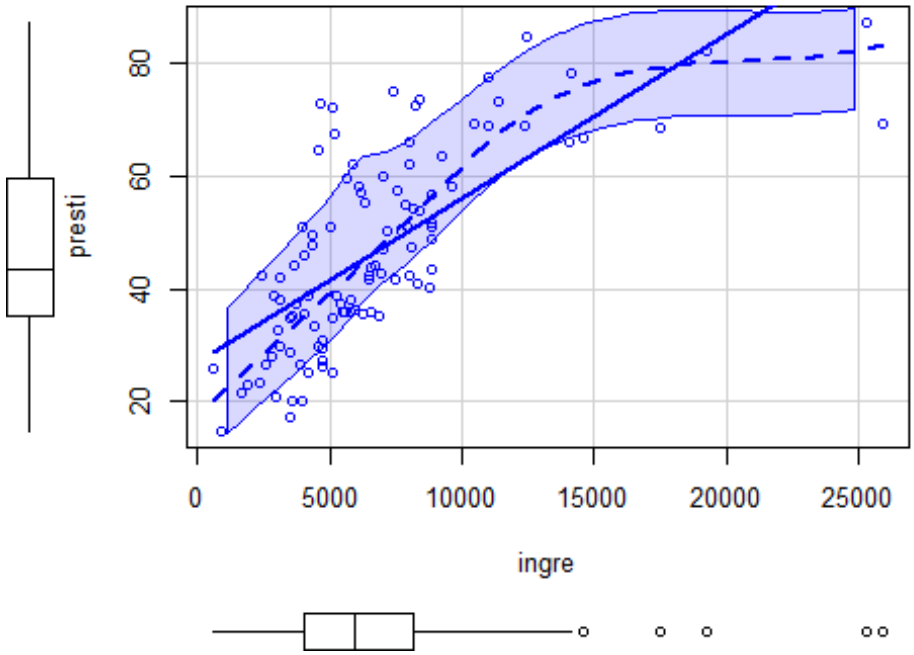
Pueden verificarse los grados de libertad de los residuales (n-p) para corroborar que no se están usando todos los datos.

```
nrow(base)

## [1] 102

summary(tipo)

##   bc prof   wc NA's
##   44   31   23    4
```

En la base de datos existen 102 ocupaciones, sin embargo, se dispone de 92 grados de libertad en los residuales, por lo que si $p = 6$ (un coeficiente-intercepto y cuatro coeficientes-pendiente, en donde el relativo a la variable dicotómica consume 2 gl) entonces $n = gl + p = 92 + 6 = 98$. Esto significa que se están usando solamente 98 casos (los cuatro valores restantes se usaron para calcular los cuatro coeficientes de regresión, por lo que $102 - 4 = 98$). Lo anterior se comprueba usando la sintaxis 'summary' en la variable "tipo", donde se ve que existen 4 valores NA.

Adicionalmente, puede obtenerse el scatterplot para ver estudiar gráficamente la relación lineal "pura" entre 'presti' e 'ingre'.

```
library(car)
scatterplot(presti~ingre)
```



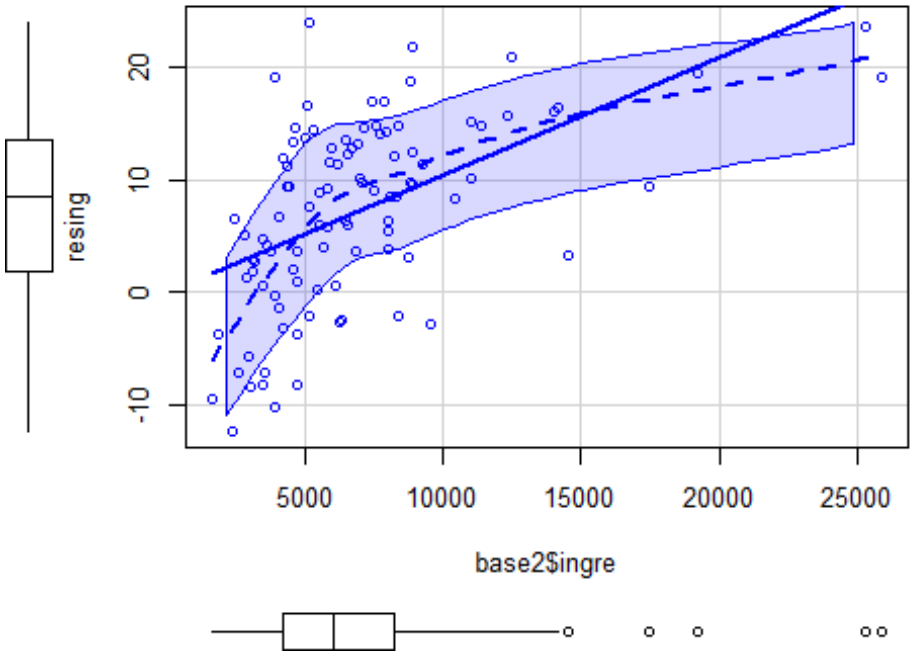
También es posible analizar gráficamente los residuales parciales relativos a la variable ingreso. Para hacerlo, deben obtenerse los residuales del modelo y sumárseles a estos un término igual a la variable ingreso multiplicada por su coeficiente de regresión, de la siguiente forma:

$$r_j^{(P)} = r + \beta_j x_j$$

El gráfico se construye con el scatterplot de los residuales parciales contra la variable 'ingre'. Hay que tomar en cuenta que se omitieron 4 casos debido a los NA, por lo que lo mejor es construir una nueva base de datos omitiendo esos 4 casos NA. Para esto puede utilizarse la sintaxis 'na.omit(base)' y posteriormente tomar la nueva variable 'ingre' a la que se le han eliminado los 4 casos.

```
base2=na.omit(base)
res=summary(mod)$res
beta=mod$coef
```

```
resing=res+base2$ingre*beta[2]
scatterplot(base2$ingre,resing)
```

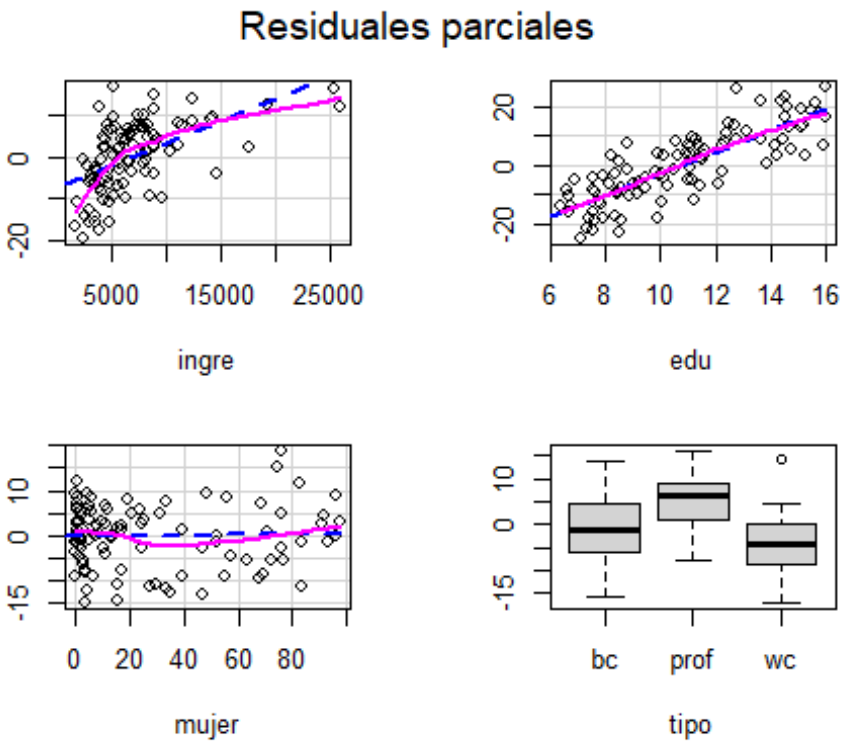


Es recomendable verificar gráficamente si el comportamiento de los residuales parciales es lineal con respecto al predictor y comparar este gráfico con el scatterplot anterior. El resultado de ello es que en ambos gráficos es notoria una curva que hace pensar que no hay una relación lineal entre la respuesta y el predictor 'ingre'. Complementariamente, es necesario matizar que, si bien este comportamiento se puede visualizar en ambos gráficos, éstos no son idénticos ya que al graficar los residuales parciales se toma en cuenta las otras variables, mientras que el primer gráfico se limita a capturar la relación directa entre la respuesta promedio y el predictor. Por lo tanto, para los fines establecidos es más adecuado que el estudio gráfico se realice con los residuales parciales.

Los gráficos de la totalidad de residuales parciales pueden obtenerse usando la función 'crPlots' de la librería 'car'. Basta poner el nombre del modelo dentro de la

función; nótese que esta función genera un boxplot para el caso de la variable categórica, sin embargo, no es útil para ver ninguna relación lineal por lo que, para tales fines, se puede obviar.

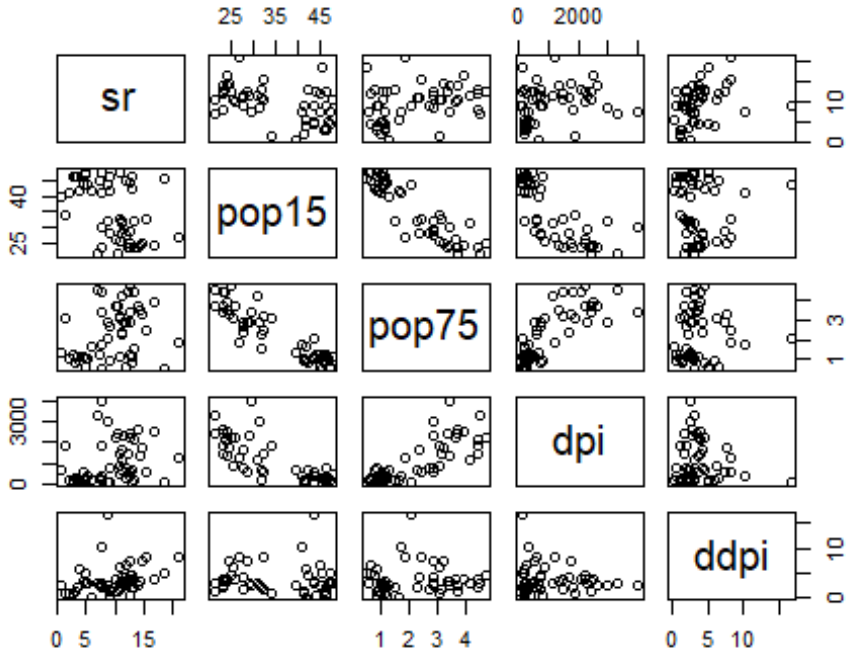
```
crPlots(mod,main="Residuales parciales",ylab="")
```



En el caso de 'edu' la relación es muy lineal, sin embargo, en el caso de 'mujer' se nota una curva que parecería describir una parábola y en el caso de *ingre* un tipo de curva más atípico.

IV. CASO DE APLICACIÓN: AHORRO DURANTE EL CICLO DE VIDA DE LAS PERSONAS ENTRE PAÍSES

```
load("ahorro.Rdata")
attach(ahorro)
pairs(ahorro) #Fuente: (The University of Auckland, 2022).
```



Se dispone de un conjunto datos con 50 observaciones sobre 5 variables.

- 'sr': Ahorro personal agregado.
- 'pop15': porcentaje de la población menor de 15 años.
- 'pop75': porcentaje de la población mayor de 75 años.
- 'dpi': renta real disponible per cápita.
- 'ddpi': tasa de crecimiento de la renta real disponible per cápita.

Como se señala en (ETCH Zürich, 2022), según la hipótesis de ahorro del ciclo de vida desarrollada por el economista neoclásico Franco Modigliani, la tasa de ahorro (ahorro personal agregado dividido por la renta disponible) se explica por la renta disponible per cápita, la tasa de variación porcentual de la renta disponible per cápita y dos variables demográficas: porcentaje de población menor de 15 años y porcentaje de población mayor de 75 años. Los datos se promedian durante la década 1960-1970 para eliminar el ciclo económico u otras fluctuaciones a corto

plazo. Se señala en el lugar citado que los datos se obtuvieron de Belsley, Kuh y Welsch (1980), quienes a su vez los obtuvieron de Sterling (1977).

IV.I. Análisis Gráfico de Valores Extremos

Al estudiar valores extremos debe considerarse que dos o más valores extremos cercanos pueden ocultarse el uno al otro, así como también que un valor extremo en un modelo puede no serlo en otro cuando las variables se han cambiado o transformado; en tal caso, será necesario re-investigar el asunto una vez cambiado el modelo. Cuando la distribución del error no es normal o se sospecha que pueda no serlo, es razonable esperar observar la presencia de residuales más grandes que otros. Valores extremos individuales no son un gran problema en conjuntos de datos grandes, sin embargo, sí hay que preocuparse ante aglomeraciones de valores extremos.

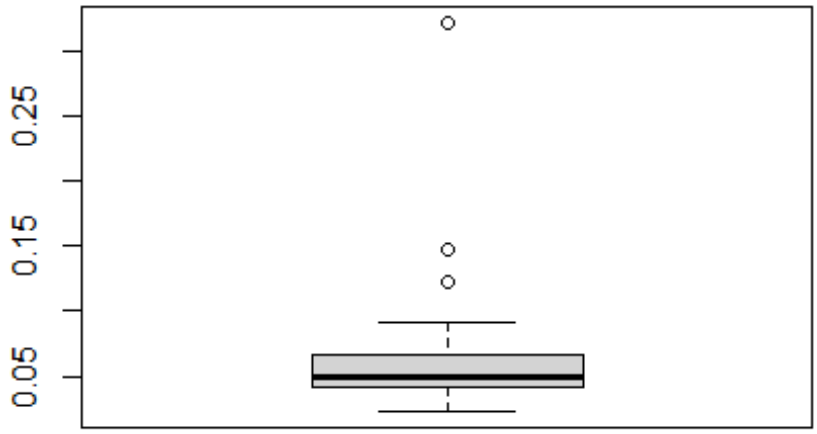
```
load("ahorro.Rdata")
attach(ahorro)

## The following objects are masked from ahorro (pos = 3):
##
##     ddpi, dpi, pop15, pop75, sr

# Modelo con dos predictores: pop15 y dpi.
mod.a = lm(sr ~ pop15 + dpi)


#Dos formas de estimar la diagonal de la matriz H
#Los leverages (infuencias) ayudan a identificar valores extremos


#FORMA MANUAL
X = cbind(1, pop15,dpi)
Hsombrero = X %*% solve(t(X)%*%X) %*% t(X)
diag(Hsombrero)
```

```
##  [1] 0.05555273 0.06910743 0.05134744 0.03758993 0.03428606 0.14707959
##  [7] 0.02546323 0.04277347 0.05315642 0.06584132 0.06067541 0.05115931
## [13] 0.03282400 0.04847935 0.06104130 0.09149078 0.04958504 0.05723176
## [19] 0.04451057 0.04198237 0.02780425 0.06367230 0.04858191 0.03688057
## [25] 0.06618090 0.04636347 0.04721654 0.04851403 0.02303193 0.04417997
## [31] 0.03865343 0.03632983 0.04025141 0.05048273 0.08056661 0.04713181
## [37] 0.08283344 0.07602968 0.12151810 0.06880813 0.03696912 0.04975062
## [43] 0.05733637 0.32125421 0.07060138 0.04606004 0.03148665 0.07200469
## [49] 0.04180306 0.05652529
```

```
#FORMA AUTOMATIZADA
(lev = hatvalues(mod.a))
```

```
##          1          2          3          4          5          6          7
## 0.05555273 0.06910743 0.05134744 0.03758993 0.03428606 0.14707959 0.02546323
##          8          9         10         11         12         13         14
## 0.04277347 0.05315642 0.06584132 0.06067541 0.05115931 0.03282400 0.04847935
##         15         16         17         18         19         20         21
## 0.06104130 0.09149078 0.04958504 0.05723176 0.04451057 0.04198237 0.02780425
##         22         23         24         25         26         27         28
## 0.06367230 0.04858191 0.03688057 0.06618090 0.04636347 0.04721654 0.04851403
##         29         30         31         32         33         34         35
## 0.02303193 0.04417997 0.03865343 0.03632983 0.04025141 0.05048273 0.08056661
##         36         37         38         39         40         41         42
## 0.04713181 0.08283344 0.07602968 0.12151810 0.06880813 0.03696912 0.04975062
##         43         44         45         46         47         48         49
## 0.05733637 0.32125421 0.07060138 0.04606004 0.03148665 0.07200469 0.04180306
##         50
## 0.05652529
```

```
#Adicionalmente, se puede realizar un primer análisis gráfico.
boxplot(lev)
boxplot(lev)$out
```




```
##        6       39       44
## 0.1470796 0.1215181 0.3212542

ahorro[c(6,39,44),]

##                sr pop15 pop75     dpi ddpi
## Canada       8.79 31.72  2.85 2982.88 2.43
## Sweden       6.86 21.44  4.54 3299.49 3.01
## United States 7.56 29.81  3.43 4001.89 2.45
```

También pueden utilizarse los residuos estandarizados o studentizados internamente para estudiar valores extremos. Debe recordarse que para usar esta clase de residuos y poder realizar inferencias estadísticas válidas debe cumplirse el

supuesto de homogeneidad de varianza, excepto cuando las diferencias en varianzas se deban al muestreo probabilístico.

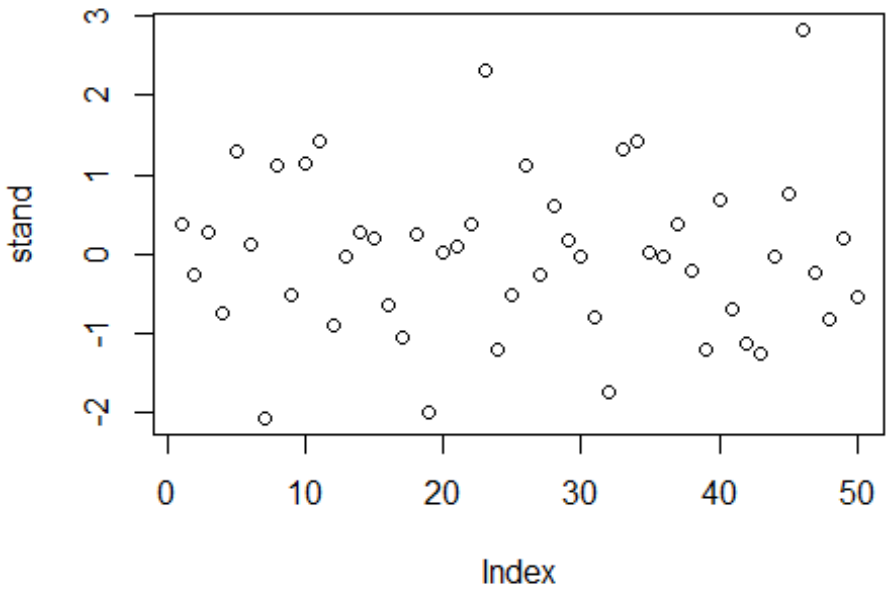
```
#FORMA MANUAL
(cme.a=anova(mod.a)$"Mean Sq"[3])

## [1] 15.83242

(raiz.cme.a=sqrt(cme.a))

## [1] 3.978997

res.a=mod.a$residuals


(stand=res.a/(raiz.cme.a*sqrt(1-lev)))

##          1          2          3          4          5          6
##  0.37909133 -0.25082580  0.27935996 -0.73712291  1.29370703  0.12651569
##          7          8          9         10         11         12
## -2.06635178  1.11713110 -0.50715272  1.14977521  1.42000010 -0.88961130
##         13         14         15         16         17         18
## -0.01859551  0.28519711  0.19653384 -0.64289817 -1.05987867  0.23892700
##         19         20         21         22         23         24
## -1.98260192  0.01277103  0.10563854  0.38668101  2.30792361 -1.19658671
##         25         26         27         28         29         30
## -0.50744389  1.10765661 -0.24884696  0.61351669  0.17263956 -0.02777234
##         31         32         33         34         35         36
## -0.80444221 -1.74106658  1.31060339  1.43078316  0.02704278 -0.04340881
##         37         38         39         40         41         42
##  0.38322275 -0.20177469 -1.19211231  0.68367200 -0.69892398 -1.11903488
##         43         44         45         46         47         48
## -1.25243505 -0.01843156  0.75616416  2.82428393 -0.23262454 -0.82854217
```

```
##         49         50
##  0.19803055 -0.54157110
```

```
#FORMA AUTOMATIZADA
rstandard(mod.a)
```

```
##          1          2          3          4          5          6
##  0.37909133 -0.25082580  0.27935996 -0.73712291  1.29370703  0.12651569
##          7          8          9         10         11         12
## -2.06635178  1.11713110 -0.50715272  1.14977521  1.42000010 -0.88961130
##         13         14         15         16         17         18
## -0.01859551  0.28519711  0.19653384 -0.64289817 -1.05987867  0.23892700
##         19         20         21         22         23         24
## -1.98260192  0.01277103  0.10563854  0.38668101  2.30792361 -1.19658671
##         25         26         27         28         29         30
## -0.50744389  1.10765661 -0.24884696  0.61351669  0.17263956 -0.02777234
##         31         32         33         34         35         36
## -0.80444221 -1.74106658  1.31060339  1.43078316  0.02704278 -0.04340881
##         37         38         39         40         41         42
##  0.38322275 -0.20177469 -1.19211231  0.68367200 -0.69892398 -1.11903488
##         43         44         45         46         47         48
## -1.25243505 -0.01843156  0.75616416  2.82428393 -0.23262454 -0.82854217
##         49         50
##  0.19803055 -0.54157110
```

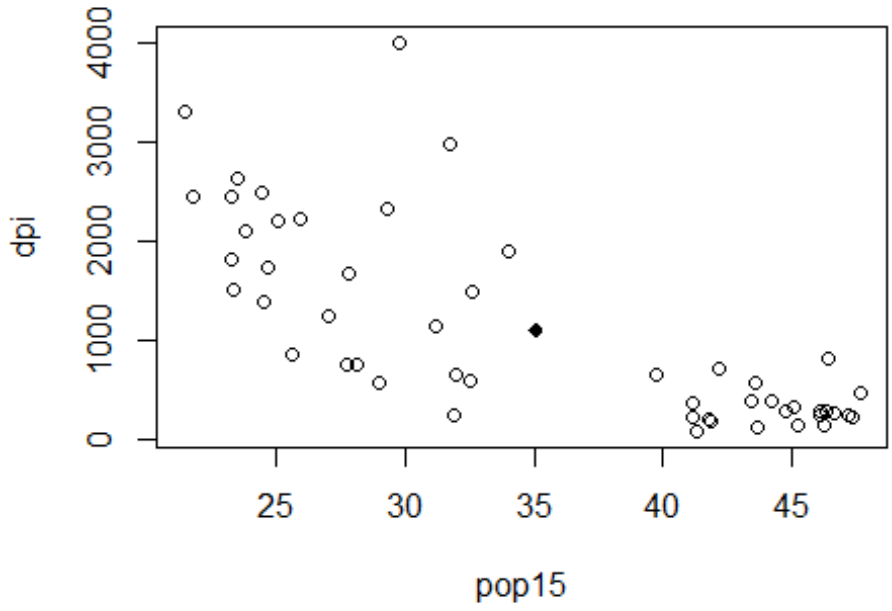
```
plot(stand)
```




```
stand[stand>2]

##       23       46
## 2.307924 2.824284

ahorro[c(23,46),]

##          sr pop15 pop75     dpi ddpi
## Japan  21.10 27.01  1.91 1257.28 8.21
## Zambia 18.56 45.25  0.56  138.33 5.14
```

Se puede observar que Zambia y Japón conforman los valores extremos, puesto que los extremos en este caso son tales con respecto a las variables que se incluyen en el estudio. Otra forma de identificar dichos valores es observando cuáles de los valores se alejan del centroide (que desempeña el papel de la media en una distribución), o en su defecto con los *leverages*.

```
# Un gráfico de la distribución bivariada de pop15 y dpi.
pais = row.names(ahorro)
plot(pop15,dpi)
points(mean(pop15),mean(dpi),pch=18)
identify(pop15,dpi,pais)
```



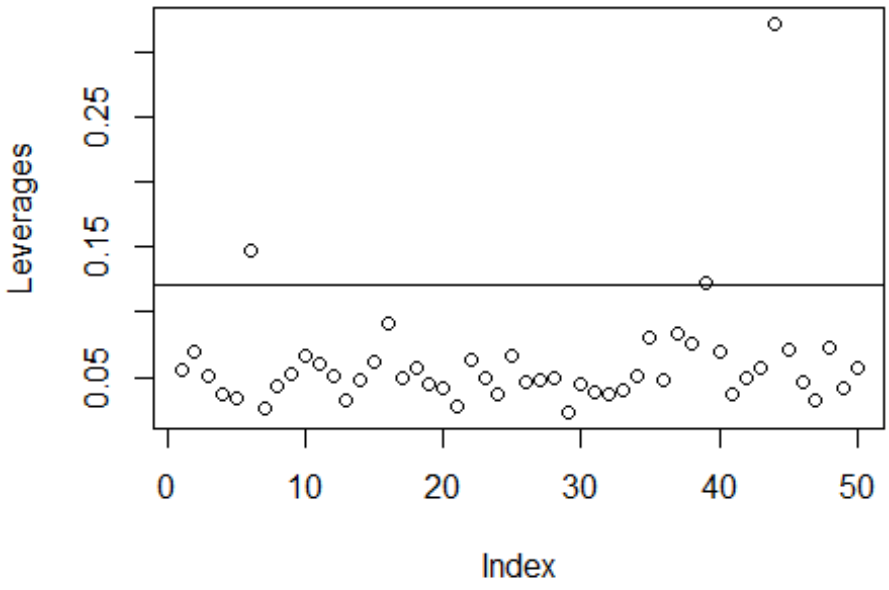

```
## integer(0)
```

Para la variable *dpi* se puede observar a Estados Unidos (USA) como valor extremo.

Además, pueden verificarse los valores extremos para el caso del primer modelo.

```
n = nrow(ahorro)
plot(lev,ylab="Leverages")
lim=2*mean(lev)
```

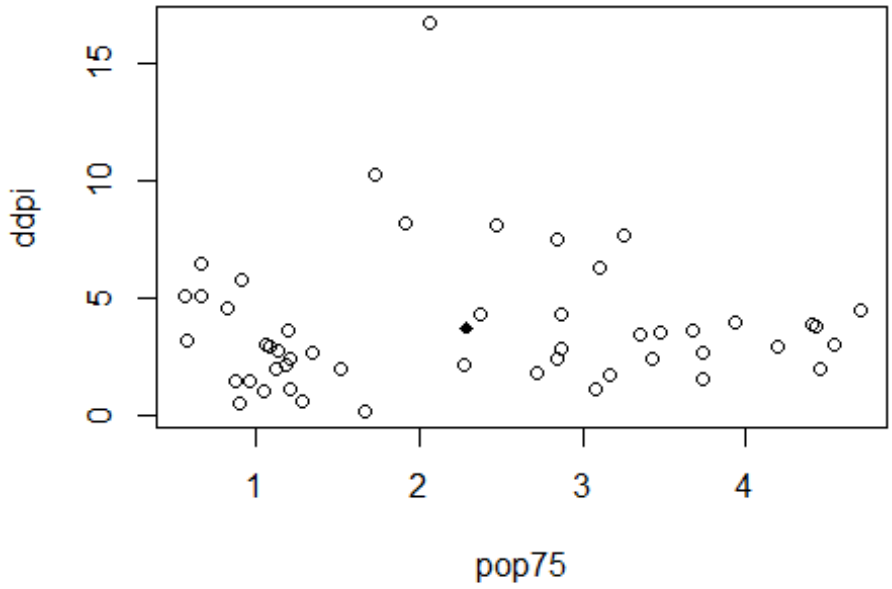
```
abline(h=lim)
identify(1:n,lev,pais)
```



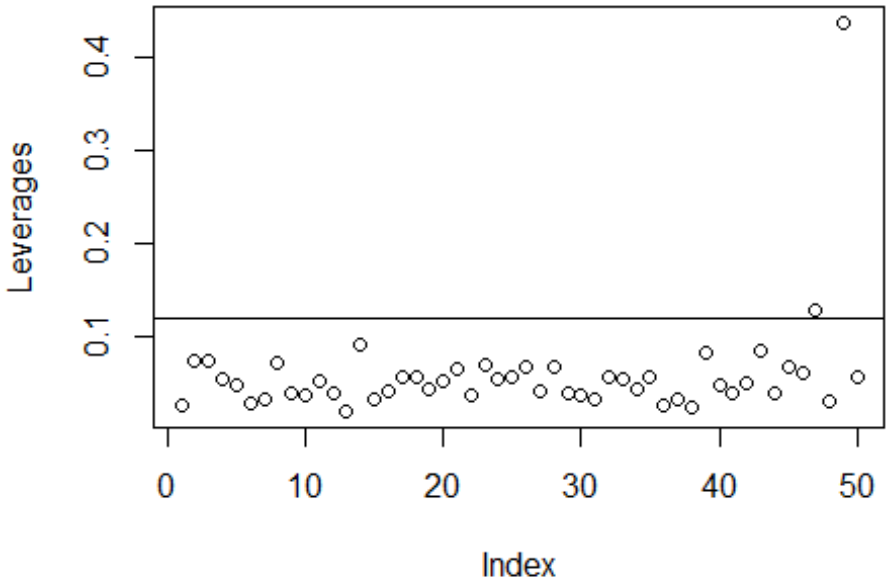

```
## integer(0)

#En este caso, los valores extremos son Estados Unidos y Canada.
```

También puede considerarse otro modelo mod.b con los predictores *pop75* y *ddpi* y verificar para este si existen valores extremos.

```
#mod.b
mod.b = lm(sr ~ pop75 + ddpi)
plot(pop75,ddpi)
points(mean(pop75),mean(ddpi),pch=18)
identify(pop75,ddpi,pais)
```




```
## integer(0)

# Identificación de valores extremos
lev2 = hatvalues(mod.b)
plot(lev2,ylab="Leverages")
lim2=2*mean(lev2)
abline(h=lim2)
n=nrow(ahorro)
identify(1:n,lev2,pais)
```




```
## integer(0)
```

En este modelo se observa que los valores extremos son “Lybia” y tal vez “Jamaica”.

Complementariamente, pueden estimarse los residuales studentizados externamente de forma manual, así como también de forma automatizada. Para ello, se debe generar un criterio para construir un intervalo de referencia para realizar el análisis gráfico de los residuos. Este criterio puede ser, para esta u otra muestra (sin descartar posibles modificaciones requeridas), qt(1-0.05/(2*n),gl), donde 0.05 indica la significancia y gl los grados de libertad calculados como $n - p - 1$. Como se adelantó, los residuos studentizados externamente dejan de fuera una observación en sus cálculos. La relevancia de los residuos studentizados externamente se explica a continuación.

Como se señala en (Penn State University, 2022), al tratar de identificar valores atípicos, un problema que puede surgir es cuando existe un valor atípico potencial

que influye en el modelo de regresión hasta el punto de que la función de regresión estimada se ve "atraída" hacia el valor atípico potencial, de modo que no se marca como un valor atípico usando el criterio residual estandarizado.

Para abordar este problema, los residuos studentizados ofrecen un criterio alternativo para identificar valores atípicos. La idea básica es eliminar las observaciones de una en una, reajustando cada vez el modelo de regresión en las $n-1$ observaciones restantes. Luego, se comparan los valores de la respuesta observados con sus valores ajustados en función de los modelos con la i-ésima observación eliminada. Esto produce residuos eliminados. Si al proceso anterior se agrega la estandarización de los residuos eliminados, se producen entonces residuos sstudentizados externamente (porque, como se adelantó, se descarta la i-ésima observación), que es una forma de robustecer el método de residuos eliminados.

¿Por qué esta medida de residuos eliminados? El hecho de que la observación i-ésima sea influyente implica que dicha observación "tira" de la línea de regresión estimada hacia sí mismo. En ese caso, la respuesta observada estaría cerca de la respuesta predicha. Pero, si se elimina tal observación influyente del conjunto de datos, entonces la línea de regresión estimada "rebotaría" lejos de la respuesta observada, lo que resultaría en un gran residual eliminado[3]. Es decir, un punto de datos que tiene un gran residuo eliminado sugiere que el punto de datos es influyente.

Un residuo studentizado externamente es simplemente un residuo eliminado dividido por su desviación estándar estimada[4]. Esto es equivalente al residuo ordinario dividido por un factor que incluye el error cuadrático medio basado en el

---

[3] "Gran" en términos de su magnitud numérica.

[4] Véase la primera ecuación de (Penn State University, 2022).

modelo estimado con la i-ésima observación eliminada MSE_i, y el apalancamiento, h_{ii}[5].

```
#FORMA MANUAL
fit=mod.b$fit
res=mod.b$res
p=3
gl=n-p-1
(scres = anova(mod.b)[p,2])

## [1] 798.3953

(t = res * sqrt(gl/(scres*(1-lev2)- res^2)))

##          1          2          3          4          5          6
##  0.377999031  0.012085072  0.294418484 -0.398930404  1.098168929 -0.210235561
##          7          8          9         10         11         12
## -1.914308351  0.674627302 -0.752677981  0.686279794  1.336209641 -1.033952658
##         13         14         15         16         17         18
##  0.297602813  0.008525261  0.463279704 -0.254400967 -1.019401523  0.031995151
##         19         20         21         22         23         24
## -2.059774226  0.441710736 -0.002436402  0.847362820  2.608982483 -1.296745806
##         25         26         27         28         29         30
##  0.037570500  0.901818228 -0.204800728  0.534543126  0.241305253 -0.151057621
##         31         32         33         34         35         36
## -0.985054227 -1.270403905  1.398729185  1.295096413  0.122613425  0.539153888
##         37         38         39         40         41         42
##  1.312045718  0.293291915 -1.242680388  0.845114702 -0.706708183 -1.118907006
##         43         44         45         46         47         48
```

[5] Véase la segunda ecuación de (Penn State University, 2022).

```
## -0.854162190 -0.670835698  0.634276188  2.695799690 -1.133601567 -0.004513445
##           49           50
## -2.202742089 -0.954489975
```

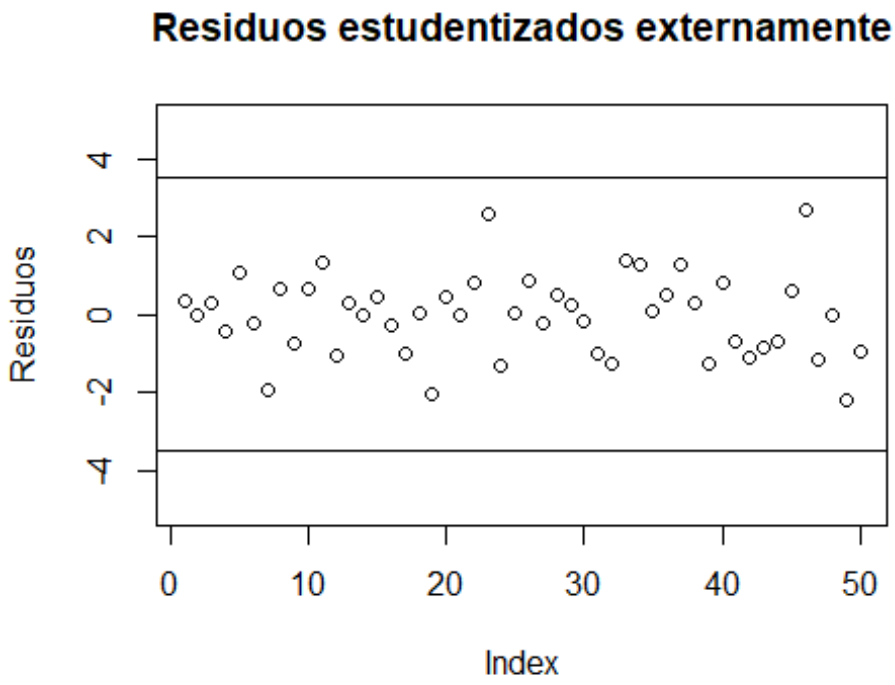
```
#FORMA AUTOMATIZADA
ts= rstudent(mod.b)
head(cbind(t,ts))
```

```
##            t         ts
## 1  0.37799903  0.37799903
## 2  0.01208507  0.01208507
## 3  0.29441848  0.29441848
## 4 -0.39893040 -0.39893040
## 5  1.09816893  1.09816893
## 6 -0.21023556 -0.21023556
```

```
##GRAFICANDO LOS RESIDUOS STUDENTIZADOS EXTERNAMENTE
plot(ts,ylim=c(-5,5),ylab="Residuos",main="Residuos estudentizados
externamente")
tcrit = qt(1-0.05/(2*n),gl)
tcrit
```

```
## [1] 3.514957
```

```
abline(h=c(-tcrit,tcrit))
identify(1:n,ts,pais)
```



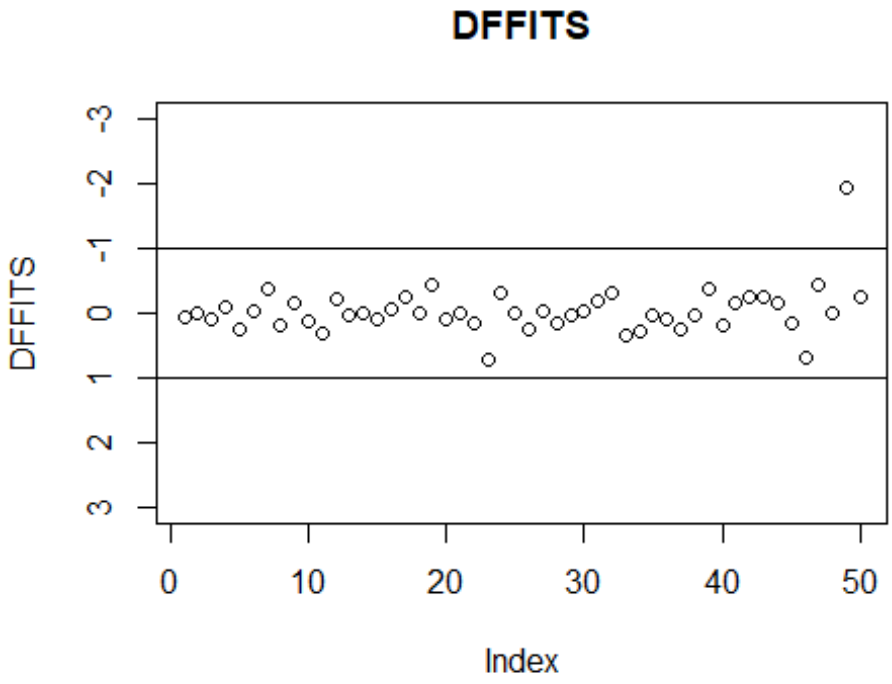

```
## integer(0)
```

IV.II. Casos de Influencia Individual: DFFITS.

El DFFIT fue propuesto por vez primera en David Belsley, A. David, Edwin Kuhn y Roy Welsh en 1980, es un diagnóstico diseñado para mostrar qué tan influyente es una observación dentro de una regresión estadística. El DFFIT es el cambio en el valor predicho para una observación, obtenido cuando esa observación se deja fuera de la regresión. Así, el DFFITS es el DFFIT Studentizado, donde la studentización se logra dividiendo el DFFIT por la desviación estándar del ajuste (el error) para esa estimación (Penn State University, 2022). Una observación se

considera influyente si el valor absoluto de su DFFITS es mayor que $2\sqrt{(k+2)/(n-k-2)}$, que en este caso sería $2\sqrt{(3+2)/(50-3-2)} \approx 0.67$[6].

```
#ANÁLISIS GRÁFICO
dffits = dffits(mod.b)
plot(dffits,ylab="DFFITS",main="DFFITS",ylim=c(3,-3)) #estos límites son ajustables según el caso
abline(h=c(-1,1))  #criterio para construir los extremos del intervalo
identify(1:n,dffits,pais)
```




```
## integer(0)
```

[6] Donde k es el número de parámetros estimados. Nótese que aún cuando no se utilice en el análisis, un modelo donde se calculó el intercepto siempre consume su respectivo grado de libertad (Cross Validated, 2017).

```
#ANÁLISIS FORMAL
dffits <- as.data.frame(dffits(mod.b))
dffits

##    dffits(mod.b)
## 1   0.0619742694
## 2   0.0034388066
## 3   0.0844058907
## 4  -0.0964002524
## 5   0.2470152793
## 6  -0.0359301019
## 7  -0.3579095761
## 8   0.1883585688
## 9  -0.1527158541
## 10  0.1368814673
## 11  0.3157112672
## 12 -0.2124856922
## 13  0.0434220528
## 14  0.0027127311
## 15  0.0869650668
## 16 -0.0540499956
## 17 -0.2501056940
## 18  0.0078291126
## 19 -0.4493036850
## 20  0.1046496628
## 21 -0.0006475322
## 22  0.1671626549
## 23  0.7235908530
## 24 -0.3118762352
```

```
## 25  0.0093271430
## 26  0.2423495048
## 27 -0.0435913864
## 28  0.1443212996
## 29  0.0491344828
## 30 -0.0300376666
## 31 -0.1866297584
## 32 -0.3107003230
## 33  0.3385505468
## 34  0.2776732954
## 35  0.0303075964
## 36  0.0882439108
## 37  0.2484046659
## 38  0.0468248185
## 39 -0.3756797198
## 40  0.1908180252
## 41 -0.1430153897
## 42 -0.2585822239
## 43 -0.2623211700
## 44 -0.1378227655
## 45  0.1718106094
## 46  0.6943786527
## 47 -0.4358181453
## 48 -0.0008099649
## 49 -1.9421717467
## 50 -0.2360274962
```

IV.III. Casos de Influencia Global: Distancia de Cook

La distancia de Cook o la D de Cook es una estimación de uso común sobre la influencia de una observación cuando se realiza un análisis de regresión de mínimos cuadrados. En un análisis aplicado de mínimos cuadrados ordinarios, la distancia de Cook se puede utilizar de varias formas: para indicar observaciones influyentes cuyo impacto en el modelo merece la pena estudiar; o para indicar regiones del espacio de diseño (construido mediante las covariables) donde sería deseable poder obtener más observaciones. Puede entenderse como el resumen de qué tanto un modelo de regresión cambiaría cuando la i-ésima observación es removida o, de manera formal, mide el efecto de eliminar una observación en el vector de parámetros combinados.

En (Glen, 2016)se señala que existen diversas interpretaciones de la distancia de Cook:

1) Una regla general es que las observaciones con una D de Cook mayor que 3 veces la media del conjunto de observaciones, μ, son posibles valores atípicos. Es la regla que se usará aquí y suele ser utilizada en el contexto del aprendizaje automático (Thieme, 2021).

2) Una interpretación alternativa es investigar cualquier punto por encima de $4/n$, donde n es el número de observaciones.

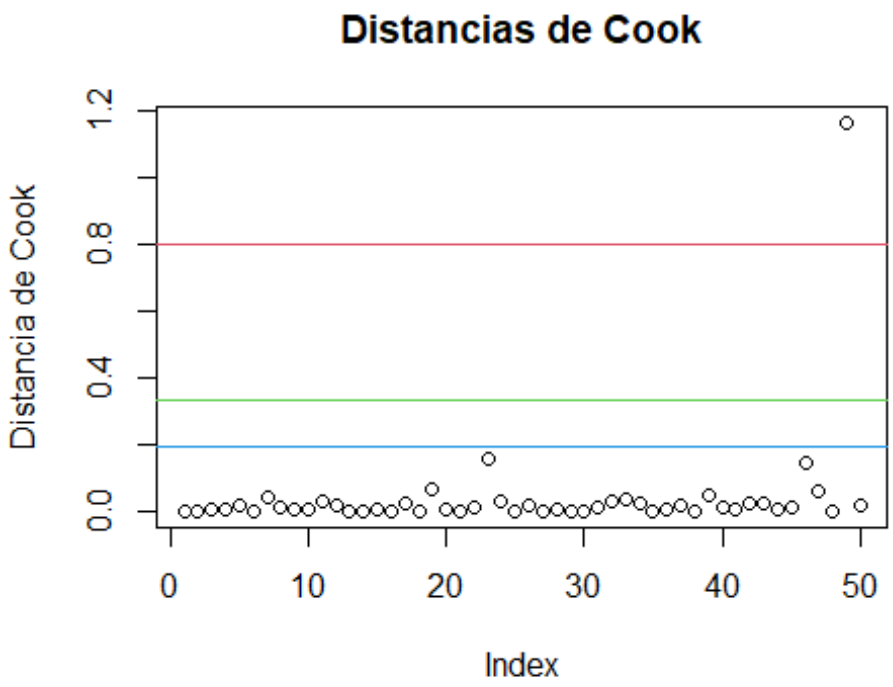
3) Otros autores sugieren que se debe investigar cualquier D "grande". ¿Qué tan grande es "demasiado grande"? El consenso parece ser que un valor D mayor que 1 indica un valor influyente, pero es posible que se deseen observar valores por encima de 0.5. También se debe investigar cualquier valor que sobresalga del otro.

4) Una forma alternativa (pero un poco más técnica) de interpretar D es encontrar el valor del percentil del valor atípico potencial utilizando la distribución F. Un percentil de más de 50 indica un punto muy influyente.

Complementariamente, otros textos como (Fox, 1991, pág. 34) establecen un umbral que puede ser no solo de $4/N$ sino también de $4/(N-k-1)$. En su caso, la última fórmula debería producir un umbral de alrededor de 0,1. Sin embargo, John Fox es bastante cauteloso cuando se trata de dar umbrales numéricos. En el lugar citado, aconseja el uso de gráficos y examinar con más detalle los puntos con "valores de D que son sustancialmente mayores que el resto". Según Fox, los umbrales solo deberían usarse para mejorar las pantallas gráficas.

En los gráficos generados anteriormente, aquellas observaciones de alta *influencia* en el modelo (a esto en inglés se conoce como alto *leverage*) son aquellas que se encuentran fuera de la región comprendida entre las dos líneas rojas punteadas en la parte superior e inferior del gráfico titulado "Residuals vs Leverage". En este caso, no parecen haber puntos fuera de tales líneas, sin embargo, el análisis formal revela (aunque desaconsejado por Fox) que las observaciones 2, 3, 6, 16, 19 y 45 son de alta influencia.

```
#ANÁLISIS GRÁFICO
cook = cooks.distance(mod.b)
plot(cook,ylab="Distancia de Cook",main="Distancias de Cook")
q1=qf(0.5,p,n-p) #calculando F para una probabilidad de 0.5
q2=qf(0.2,p,n-p) #calculando F para una probabilidad de 0.2
q3=qf(0.1,p,n-p) #calculando F para una probabilidad de 0.1
abline(h=c(q1,q2,q3),col=c(2,3,4))
identify(1:n,cook,pais)
```




```
## integer(0)

#ANÁLISIS FORMAL
library(mvinfluence)

## Loading required package: heplots

CooksD <- cooks.distance(mod.b)
influential <- CooksD[(CooksD > (3 * mean(CooksD, na.rm = TRUE)))]
influential

##       23       46       49
## 0.1553363 0.1418105 1.1620992
```

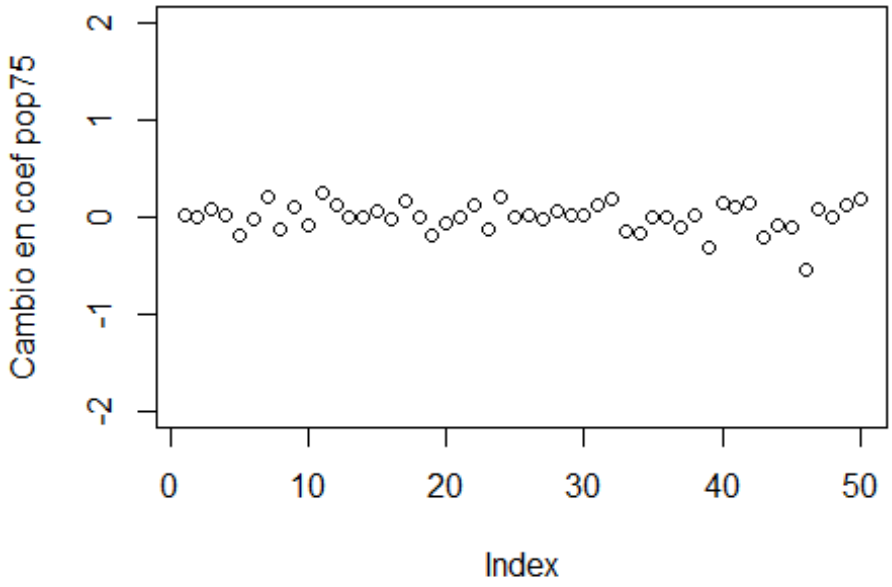
IV.IV. Casos de Influencia sobre los Coeficientes de Regresión: DFBETAS.

Como se señala en (Rockefeller College, 2007, págs. 1-2), el DFBETA de una observación en particular es la diferencia entre el coeficiente de regresión para un predictor incluido calculado para todos los datos y el coeficiente de regresión

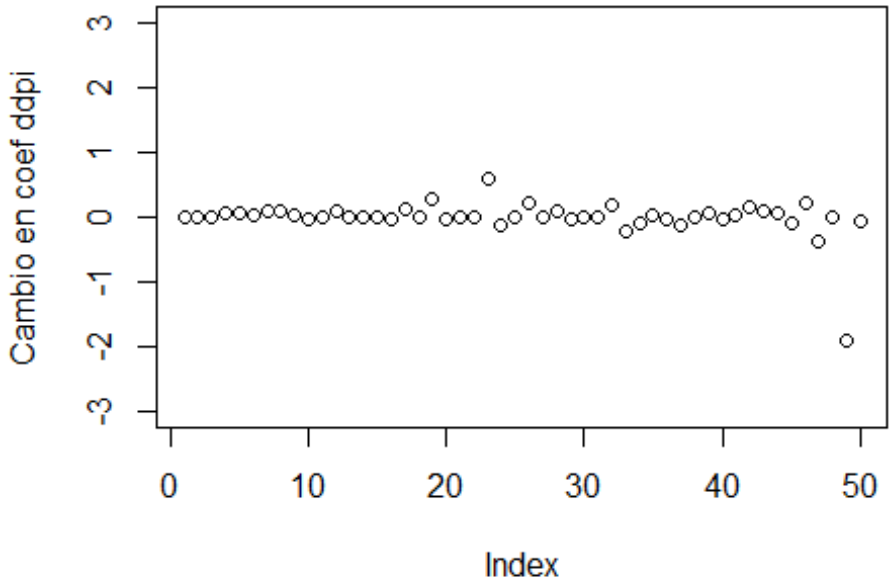
calculado con la observación eliminada, escalado por el error estándar calculado con la observación eliminada. El valor de corte para las DFBETA es $2\sqrt{n}$, donde n es el número de observaciones. Sin embargo, otro criterio es buscar observaciones con un valor superior a 1. Aquí, el "punto de corte" significa que se interpreta como "esta observación podría influir demasiado en el coeficiente estimado".

Como se señala en (SAS, 2019), si se excluye una observación de los datos y reajusta el modelo, se obtendrán nuevas estimaciones de los coeficientes de regresión. ¿Cuánto cambian las estimaciones? Obsérvese que, a nivel formal, se obtiene una estadística para cada observación y también una para cada regresor (incluida la intersección). Por tanto, si se tienen n observaciones y k regresores, se obtendrán $n * k$ estadísticos.

```
#ANÁLISIS GRÁFICO
dfbetas = dfbetas(mod.b)
plot(dfbetas[,2],ylab="Cambio en coef pop75",ylim=c(-2,2)) #-2 y 2 son valores ajustables
identify(1:n,dfbetas[,2],pais)
```




```
## integer(0)

plot(dfbetas[,3],ylab="Cambio en coef ddpi",ylim=c(-3,3)) ##-3 y 3 son valores
ajustables
identify(1:n,dfbetas[,3],pais)
```




```
## integer(0)

mod.c=lm(sr ~ pop75 + ddpi, ahorro[-49,])
summary(mod.b)

##
## Call:
## lm(formula = sr ~ pop75 + ddpi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0223 -3.2949  0.0889  2.4570 10.1069
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4695     1.4101   3.879 0.000325 ***
```

```
## pop75        1.0726     0.4563   2.351 0.022992 *
## ddpi         0.4636     0.2052   2.259 0.028562 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.122 on 47 degrees of freedom
## Multiple R-squared:  0.1883, Adjusted R-squared:  0.1538
## F-statistic: 5.452 on 2 and 47 DF,  p-value: 0.007423
```

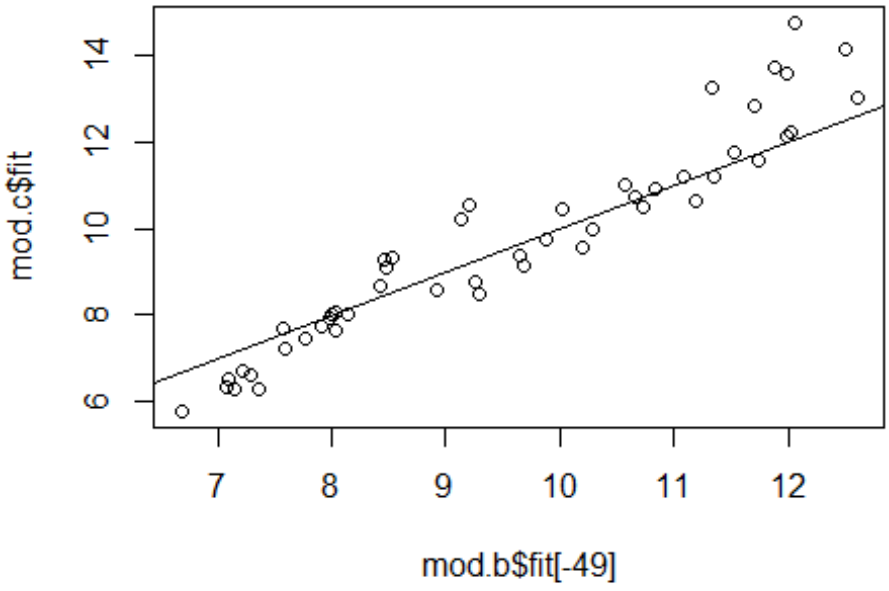
```
summary(mod.c)
```

```
##
## Call:
## lm(formula = sr ~ pop75 + ddpi, data = ahorro[-49, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4209 -2.8687  0.1448  2.3132  9.2655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4178     1.4372   3.074  0.00355 **
## pop75         1.0197     0.4393   2.321  0.02477 *
## ddpi          0.8377     0.2603   3.218  0.00237 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.962 on 46 degrees of freedom
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.2334
## F-statistic: 8.305 on 2 and 46 DF,  p-value: 0.0008328
```

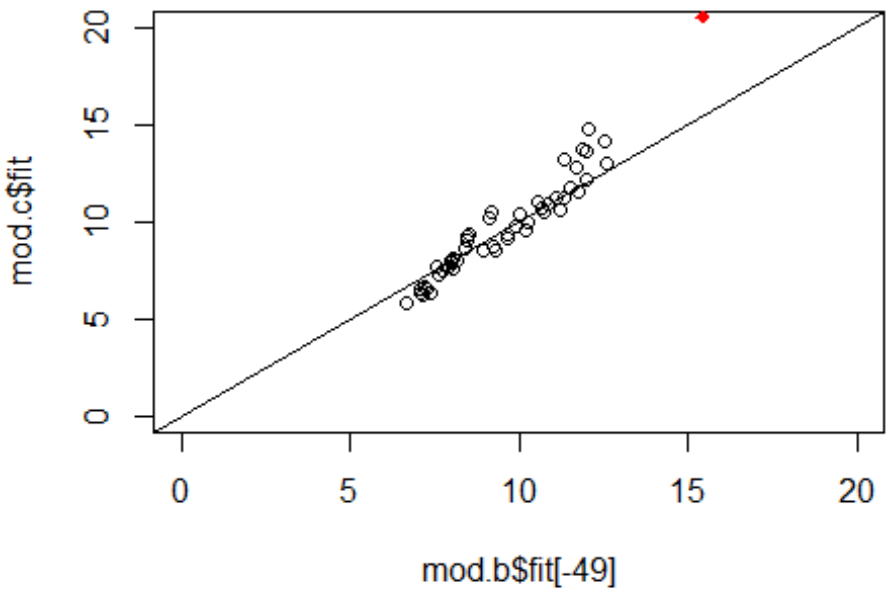
Con esto se ejemplifica el cambio que puede tener un coeficiente, en este caso el de 'ddpi', que es donde está influyendo la observación en cuestión. Adicionalmente, debe decirse que en todos los tipos de pruebas gráficas sobre valores extremos que se han realizado se ha identificado a Libia como tal, por lo que es recomendable investigar de manera específica a Libia, lo que se debe hacer no sólo desde la teoría estadística, sino también desde algún marco científico-técnico de referencia.

Complementariamente, puede construirse un modelo 'mod.c' que excluya a Libia, con la finalidad de observar gráficamente el comportamiento de las predicciones cuando no se considera a este país que representaba observaciones extremas respecto a los demás.

```
#ANÁLISIS GRÁFICO
(extr.0=mod.b$fit[49]) #la estimación, para el caso en que se considera Libia, es de 15.43

##      49
## 15.43674

ahorro[49,]

##       sr pop15 pop75    dpi  ddpi
## Libya 8.89 43.69  2.07 123.58 16.71

(extr.1=predict(mod.c,data.frame(pop75=2.07,ddpi=16.71))) #la estimación, para el caso en que no se considera a Libia, es de 20.

##       1
## 20.52621

#Así, el modelo con el extremo Libia estima 15 y sin este estima 20.


plot(mod.b$fit[-49],mod.c$fit)   #Este valor afecta la estimación de todos los demás valores
abline(0,1)
```




```
plot(mod.b$fit[-49],mod.c$fit,xlim=c(0,20),ylim=c(0,20))
points(extr.0,extr.1,pch=18,col="red")
abline(0,1) #Intercepto 0 y pendiente 1
```



#Es un valor que se encuentra alejado de ambos modelos, del que lo contiene y del que no, si de igual forma se intentase estimar.

#ANÁLISIS FORMAL

```
dfbetas <- as.data.frame(dfbetas(mod.b))
dfbetas
```

```
##      (Intercept)        pop75        ddpi
## 1   1.351530e-02  0.0248978041 -1.754920e-02
## 2  -1.466813e-03  0.0029421551  3.329427e-05
## 3  -3.536276e-02  0.0724383694 -8.831933e-04
## 4  -8.278417e-02  0.0264768306  7.157822e-02
## 5   1.748061e-01 -0.1834540770  4.958998e-02
## 6  -1.033715e-02 -0.0135094153  1.443254e-02
## 7  -3.191645e-01  0.2028037717  1.002652e-01
## 8   8.189367e-02 -0.1282869582  9.918319e-02
```

```
## 9  -1.349223e-01  0.1041832092  2.325856e-02
## 10  1.235777e-01 -0.0884842917 -3.110646e-02
## 11 -1.093207e-01  0.2484325307  9.587759e-03
## 12 -1.990796e-01  0.1268129735  7.913043e-02
## 13  1.126672e-02  0.0023505909  8.357472e-03
## 14 -1.392368e-03  0.0023754688  2.792672e-04
## 15 -8.673785e-03  0.0553495051 -8.851317e-03
## 16  1.896756e-02 -0.0224093757 -3.207783e-02
## 17 -2.446369e-01  0.1623474000  1.148532e-01
## 18  6.965961e-03 -0.0062232765 -7.727856e-04
## 19 -1.357132e-01 -0.1906989037  2.815375e-01
## 20  1.017890e-01 -0.0657221161 -4.842921e-02
## 21  1.872718e-04 -0.0005318949  1.097636e-04
## 22 -2.721582e-02  0.1137386885 -1.223234e-02
## 23 -7.513036e-02 -0.1300024256  6.031625e-01
## 24 -1.546840e-01  0.2076626321 -1.414742e-01
## 25  5.403938e-06  0.0062653826 -4.372722e-03
## 26 -6.585979e-02  0.0131626085  2.023820e-01
## 27  1.024575e-02 -0.0319574888  2.242952e-03
## 28 -6.716832e-02  0.0559387864  1.061046e-01
## 29  9.921269e-03  0.0245302550 -2.510034e-02
## 30 -2.762453e-02  0.0182183675  9.330400e-03
## 31 -1.508411e-01  0.1211328020  4.296599e-03
## 32 -3.013163e-01  0.1754802328  1.730726e-01
## 33  3.185004e-01 -0.1557952964 -2.171698e-01
## 34  2.638996e-01 -0.1690738201 -1.115650e-01
## 35 -1.065143e-02  0.0071958663  2.321366e-02
## 36  5.503111e-02  0.0002935281 -4.262501e-02
## 37  2.229958e-01 -0.1113200579 -1.139742e-01
```

```
## 38 -1.079745e-03  0.0187510380  8.280699e-03
## 39  1.337521e-01 -0.3241810643  5.650568e-02
## 40 -2.562410e-02  0.1389910220 -4.911630e-02
## 41 -1.277817e-01  0.0961045367  2.618463e-02
## 42 -2.467483e-01  0.1338885971  1.467690e-01
## 43  6.279840e-02 -0.2163381663  8.318391e-02
## 44 -8.103937e-04 -0.0873368007  4.677393e-02
## 45  1.679846e-01 -0.0986695169 -1.030457e-01
## 46  4.504554e-01 -0.5389579092  2.051481e-01
## 47  8.059105e-02  0.0856110155 -3.933330e-01
## 48 -3.366249e-04 -0.0002276316  4.342084e-04
## 49  7.757844e-01  0.1204620075 -1.895863e+00
## 50 -1.527264e-01  0.1794008045 -6.924540e-02
```

V. REFERENCIAS

Bhandari, A. (20 de Marzo de 2020). *What is Multicollinearity? Here's Everything You Need to Know*. Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Cross Validated. (22 de Mayo de 2014). *What's the difference between standardization and studentization?* Obtenido de Questions: https://stats.stackexchange.com/questions/99717/whats-the-difference-between-standardization-and-studentization

Cross Validated. (31 de Octubre de 2017). *Degrees of Freedom in Simple Linear Regression*. Obtenido de Questions: https://stats.stackexchange.com/questions/311091/degrees-of-freedom-in-simple-linear-regression

ETCH Zürich. (24 de Julio de 2022). *Intercountry Life-Cycle Savings Data*. Obtenido de R Documentation: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/LifeCycleSavings.html

Fox, J. (1991). *Regression Diagnostics: An Introduction.* Newbury Park: Sage.

Glen, S. (13 de Julio de 2016). *Cook's Distance/Cook's D: Definition, Interpretation*. Obtenido de StatisticsHowTo: https://www.statisticshowto.com/cooks-distance/

Gujarati, D., & Porter, D. (2010). *Econometría.* México, D.F.: Fondo de Cultura Económica.

Hayden Economics. (5 de Junio de 2022). *Detection Of Heteroscedasticity*. Obtenido de Regression Models: https://www.rhayden.us/regression-models/detection-of-heteroscedasticity.html

Penn State University. (24 de Julio de 2022). *Identifying Influential Data Points*. Obtenido de Applied Regression Analysis: https://online.stat.psu.edu/stat462/node/173/

Penn State University. (24 de Julio de 2022). *Studentized Residuals*. Obtenido de Applied Regression Analysis: https://online.stat.psu.edu/stat462/node/247/

ResearchGate. (24 de Septiembre de 2016). *Multicollinearity issues: is a value less than 10 acceptable for VIF?* Obtenido de Post: https://www.researchgate.net/post/Multicollinearity_issues_is_a_value_less_than_10_acceptable_for_VIF

Rockefeller College. (4 de Septiembre de 2007). *Outliers and DFBETA.* Obtenido de University at Albany:

https://www.albany.edu/faculty/kretheme/PAD705/SupportMat/DFBETA.pdf

SAS. (17 de Junio de 2019). *Influential observations in a linear regression model: The DFBETAS statistics*. Obtenido de Content: https://blogs.sas.com/content/iml/2019/06/17/influence-regression-dfbeta.html

The University of Auckland. (24 de Julio de 2022). *Savings*. Obtenido de Statistics: https://www.stat.auckland.ac.nz/~lee/330/datasets.dir/savings.txt

Thieme, C. (15 de Mayo de 2021). *Identifying Outliers in Linear Regression — Cook's Distance*. Obtenido de Towards Data Science: https://towardsdatascience.com/identifying-outliers-in-linear-regression-cooks-distance-9e212e9136a