Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Especialización en Estadística Aplicada

Análisis de Regresión

Examen 3

Jesus Alberto Urrutia Camacho (urcajeal@gmail.com)

Ciudad de México

9 de junio de 2021

1. Se realiza un estudio del el efecto que tiene la renta per capita (R), la zona costera (ZC) o archipiélago (ZA) sobre las estadías hoteleras por habitante en diferentes provincias (P). Y se obtienen los siguientes resultados

A. Indique qué variables son estadísticamente significativas a nivel individual, usando como nivel de significancia el 5 %. ¿Y si usamos el nivel de significancia del 10 %?

Saber si una variable es estadísticamente significativa implica realizar pruebas de hipótesis. En regresión lineal múltiple es posible hacer dos tipo:

1. Relación lineal entre la Variable de respuesta "Y" y *alguna* de las variables regresoras.
2. Relación lineal entre las Variable de respuesta "Y" y de forma *individual* una coeficiente de regresión.
3. La primer prueba se expresa, de la forma:

$$H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0 vs H_1 : \beta_j \neq 0, p.a.j \in 1, \ldots, p$$

.

4. La segunda prueba se expresa, de la forma:

$$H_o : \beta_j vs H_1 : \beta_j \neq 0$$

. Además, se rechaza la hipótesis nula si

$$\mid t_o \mid > t_{\alpha/2, n-p-1}$$

. Cabe rememorar que la hipótesis nula afirma que la variables independinte X_j no contribuye a la respuesta "por lo que puede ser eliminada del modelo" (Juarez, 2021).

Ahora, cabe recordar que la estadística de prueba es una t, y esta es su expresión:

$$t_o = \frac{\hat{\beta}_j}{\sqrt[2]{\hat{\sigma}^2 * A_j}}$$

El tamaño de muestra es $n = 51$. Donde, $A_j = (X^t * X)^{-1}$, la cuál es la matriz de diseño. Esta se calcula, como se muestra a continuanción, y se extrae la diagonal.

$$\begin{bmatrix} 0.884 & -0.850 & -0.056 & -0.001 \\ -0.850 & 0.009 & 0.002 & 0.0001 \\ -0.056 & 0.002 & 0.086 & -0.05 \\ -0.001 & 0.001 & -0.05 & 0.3830 \end{bmatrix}$$

De donde la matriz diagonal tiene un valor de:

$$MI = 0.884, 0.009, 0.086, 0.383$$

Además, se conoce la varianza estimada: $\sigma^2 = 0.58$

Posteriormente, se realiza un producto de escalares, entre los elementos de la diagonal de , $A_j = (X^t X)^{-1}$ y $\hat{\sigma}^2$. Este producto tiene por valor:

$$A_j \sigma^2 = 0.5127, 0.0052, 0.0498, 0.2221$$

Finalmente, se aplica raiz cuadrada a los elementos. Lo que genera el resultado de:

$$\sqrt{A_j\sigma^2} = 0.7160, 0.07224, 0.2233, 0.47131$$

La anterior operación es el dividendo para calcular el estadístico t. Además, se requiere del estimador de β_j. Entonces, como β_j son conocidas, se procede a calcular una división para calcuar el estadístico t_o.

$$t_0 = \frac{\beta_j}{\sqrt{A_j\sigma^2}} = -0.33517, 3.3218, 2.6865, 3.3947$$

Cabe destacar que el anterior resultado se compara con un t teórica. Dado que el cuantil que se busca es $t_{\alpha/2,n-p-1} = t_{0.025,51-3-1} = 2.011741$. Esta t se busca en tablas.

Ahora, se contrastan la hipótesis nula con la siguiente expresión: $\mid t_o \mid > t_{\alpha/2,n-p-1}$.

Dado lo anterior, es posible rechazar la hipótesis nula para β_{R_i} , β_{ZC_i} y para β_{ZA_i} al 95 % de confianza. Es decir, las variables R, ZC y ZA sí tienen una relación lineal con la respuesta.

Finalmente, si se realiza la pregunta "¿Y si se usa el nivel de significancia del 10 %?". Se contruye la $t_{\alpha/2,n-p-1} = t_{0.05,51-3-1} = 1.6779$. Esta t se busca en tablas.

Luego de realizar el cálculo, es posible sostener que las mismas variables que son significativas ,continúan siendolo al 90 % de confianza.

B. Construya un intervalo del 95 % de confianza para β_2 (ZC) y otro para β_3(ZA).

Tal como afirma Juarez (2021), "al buscar acotar a alguno de los coeficientes de la regresión en particular, se hablará de intervalor de confianza, sin embargo, al buscar acotar al vector de parámetros β se hablará de regresiones de confianza". Por tal razón, se procede a contruir Intervalos *de confianza* para β_2 (ZC) y otro para β_3(ZA) de forma individual.

Un intervalo de $(1-\alpha)100$ de confianza para β_j tiene la forma de

$$\hat{\beta}_j \pm t_{a/2,n-p-1}\sqrt{\hat{\sigma}^2 A_j}$$

.

Entonces, como se conocen los valores para las $\hat{\beta}_j$, y también de $t_{a/2,n-p-1}$ y de $\sqrt{\hat{\sigma}^2 A_j}$. Es posible hacer el cálculo para conocer los intervalor para las β.

$$\hat{\beta}_{ZC} \pm t_{a/2,n-p-1}\sqrt{\hat{\sigma}^2 A_j} =$$
$$0.6 \pm 2.011741 * 0.22333 = (0.2695, -0.2695)$$

$$\hat{\beta}_{ZA} \pm t_{a/2,n-p-1}\sqrt{\hat{\sigma}^2 A_j} =$$
$$1.6 \pm 2.011741 * 0.4713 = (1.517069, -1.517069)$$

Para concluir, es posible sostener que ambos intervalos de confianza contienen a los parámetros de tanto β_{ZC} como β_{ZA}, con una confianza del 95 %.

2. Para un conjunto de empresas pertenecientes a cierto sector económico se ha ajustado la siguiente función de producción:

$$Q_i = \beta_o L_i^{\beta_i K_i^{\beta_2} e^{e_i}}$$

Para que el modelo sea lineal, se toma el logaritmo obteniendo el siguiente modelo transformado

$$ln(Q_i) = ln(\beta_o) + \beta_1 ln(L_i) + \beta_2 ln(K_i) + e_i$$

Y se obtiene que $\hat{\beta}_0 = e^{0.5}$, $\hat{\beta}_1 = 0.76$, $\hat{\beta}_2 = 0.19$, $se(\hat{\beta}_1) = 0.71$, $se(\hat{\beta}_2) = 0.14$, y $R^2 = 0.969$.

A. Realice las pruebas de hipótesis individuales para determinar la significancia de β_1 y β_2 la prueba de significancia conjunta.

Tal como se afirmó en el primer inciso, las pruebas de hipótesis permiten conocer si una variable tiene dependencia lineal con la repuesta. En regresión lineal múltiple es posible hacer dos tipo:

1. Relación lineal entre la Variable de respuesta "Y" y las variables regresoras. También conocida como *conjunta*. Cuya hipótesis es: $H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0 vs H_1 : \beta_j \neq 0, p.a.j \in 1, \ldots, p$
2. Relación lineal entre las Variable de respuesta "Y" y de forma *individual* una coeficiente de regresión. Cuya hipótesis es: $H_o : \beta_j vs H_1 : \beta_j \neq 0$

En esta sección no se cuenta con $(X^t X)^{-1}$. Sin embargo, se tiene directamente la expresión del Error estándar (SE), a fin de construir el intervalo y calcular las pruebas, pues $SE(\hat{\beta}) = \sqrt{\hat{\sigma}^2 A_i}$.

Entonces, el estadístico t tiene la siguiente expresión:

$$t_o = \frac{\hat{\beta}_j}{\sqrt[2]{\hat{\sigma}^2 * A_j}}$$

Y al realizar las operaciones para $\hat{\beta}_1$,

$$t_o = \frac{0.76}{0.71}$$
$$t_o = 1.070423$$

El anterior valor debe de ser contrastado mediante la siguinete expresión. Donde se rechazaría la hipótesis nula si:

$$\mid t_o \mid > t_{\alpha/2, n-p-1}$$

Para lo cuál se busca el valor de $t_{\alpha/2,n-p-1} = t_{5/2,23-2-1} = t_{\alpha/2=5/2,20} = 2.0859$ en tablas de valores.

Entonces, es posible sostenar la siguiente expresión: $1.070423 < 2.0859$. Por lo que no se rechaza la hipótesis nula de $H_0 : \beta_1 = 0$ al 95 % de confianza. Lo que implica que β_1 no tiene una relación lineal con la función de producción.

En contraste, al realizar el análisis para $\hat{\beta}_2$. Se realizan las siguiente operaciones:

$$t_o = \frac{0.19}{0.14} = 1.3571$$

El anterior valor debe de ser contrastado mediante la siguinete expresión:

$$\mid t_o \mid > t_{\alpha/2, n-p-1}$$

Para lo cuál se buscó el valor de $t_{\alpha/2,n-p-1} = t_{5/2,23-2-1} = 2.0859$ en tablas de valores. Y es el mismo que el estadístico anterio.

Entonces, dado que $1.3571 < 20856$, no se rechaza la hipótesis nula de $H_0 : \beta_2 = 0$ al 95 % de confianza. Lo que implica que β_2 no tiene una relación lineal con la función de producción.

Cabe destacar que β_j "depende de todas las variables regresoras, no solamente de x_j, de manera que en realidad esta prueba es para determinar si la contribución de x_j es significativa dadas todas las demás variables regresoras". Y se evidenció que de forma individual, ninguna de las dos variables son significativas para el modelo (Juarez, 2021).

A continuación, se realiza la prueba de **significancia conjunta**. Cuya hipótesis es: $H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs $H_1 : \beta_j \neq 0$, p.a. $j \in 1, \ldots, p$. Donde rechazar la hipótesis nula implica que al menos una variables es significativa para el modelo.

Esta hipótesis se prueba con la estadística F, que tiene la siguiente expresión:

$$F_o = \frac{SCE/p}{SSE/(n-p-1)} \sim F_{p,n-p-1}$$

Dado que el estadístico F también se puede escribir de esta forma:

$$F = \frac{(R^2)/(p-1)}{1-R^2/(n-p)}$$

.

Dada la anteriorr expresión, se posible sostener que

$$F = \frac{(0.969)/(3-1)}{1-0.969/(23-3)} = 3.125806$$

.

Entonces, el anterior resultado se debe de compara contra una $F_{p,n-p-1}$. Donde se rechazará la hipótesis nula si:

$$F_o > F_{p,n-p-1}$$

El valor del estadístico teórico se busca en tablas. Donde el valor de $F_{2,20} = 3.4928$. Por lo tanto, no se rechaza la hipotesis nula, y es posible arüir que ningún parámetro es estadísticamente significativo al 95 % de confianza.

B. La estimación de los modelos de regresión lineal simple de ln(Q) en función de ln(L) y de ln(Q) en función de ln(K) produjo los siguientes resultados: $ln(Q_i) = -5.5 + 1.7ln(L_i)$, con $se(\hat{\beta}_1) = 0.09$ y $R^2 = 0.964$

$ln(Q_i) = 5.3 + 0.34ln(K_i)$, con $se(\hat{\beta}_1) = 0.02$ y $R^2 = 0.966$

Realice las pruebas de hipótesis para evaluar la significancia de en cada uno de los modelos. Explique la aparente contradicción entre los resultados obtenidos del inciso A) y con los obtenidos en este inciso.

Para ambas ecuaciones, se requiere partir de la siguiente hipótesis nula:

$H_o : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$

La cuál está "relacionada con la significancia de la regresión, ya que no rechazar la hipótesis nula implica que no existe relación lineal entre x y y" (Juarez, 2021). Además, la estadística que se emplea es una f, que tiene la siguiente expresión:

$$f = \frac{\hat{\beta}^2 S_{xx}}{S^2} \sim F_{(1,n-2)}$$

Donde la anterior expresión es análoga a $F_{(1,n)} = t_n^2$ Dado que solamente se cuenta con las β_i , $se(\hat{\beta}_1) = 0.02$ y $R^2 = 0.966$, se debe proceder como sigue.

Debido a que $VAR(\hat{\beta}_i) = \frac{\sigma^2}{S_{xx}}$. Lo cuál puede argumentarse que $SE(\hat{\beta}_i) = \frac{\sigma^2}{S_{xx}}$. Sn embargo, dado que no se conoce σ^2 es necesario estimarlo mediante S^2. Lo cuál generaría la función de $SE(\hat{\beta}_i) = \frac{S^2}{S_{xx}}$.

A partir de lo anterior, es posible expresar la F como

$$f = \hat{\beta}_i^2 / SE(\hat{\beta}_i) \sim F_{(1,n-2)}$$

Particulamente para el caso de β_L.

$$F_o = 1.71/0.09 = 19$$

. Lo cuál se contrasta contra una $F_{(1,n-2)} = 4.3247$. Entonces, se **puede concluir que** $F_{o,L} >> F_{(1,n-2)}$. Por lo que el efecto de β_L sí es significativo para este modelo.

Ahora respecto a β_K.

$$F_o = 0.34/0.02 = 17$$

. Lo cuál se contrasta contra una $F_{(1,n-2)} = 4.3247$. Entonces, se **puede concluir que** que $F_{o,k} >> F_{(1,n-2)}$. Por lo que el efecto de β_K sí es significativo para este modelo.

Curiosamente, el hecho de que los estimadores salgan significativos en un modelo simple se podría deber a que hay multicolinealidad en el modelo múltiple. Además, las manifestaciones de la colinealidad es que el R^2 es muy cercano a 1, los coeficientes son menores (o hasta cambian de signo) respecto al modelo más simple o sin multicolinealidad. Además de que se puede deber a que al ser una función de producción de empresas del sector, es posible que la recolección de información se haya levantado para agrupaciones similares de empresas, y que la muestra es relativamente pequeña.

3. Utilizando el estadístico de Mallows como referencia, seleccione que modelo(s) es (son) el (los) mejor para predecir el puntaje de un estudiante en un examen. Suponga que un profesor quisiera utilizar las horas de estudio, exámenes de prueba tomados y el actual promedio como variables para predecir el puntaje que obtendrá un estudiante en un examen. El profesor realiza el ajuste con siete distintos modelos y obtiene el coeficiente de Mallows para cada uno.

El estadístico de Mallows depende del coeficiente de rergesión parcial. Donde la esperanza del estadístico es igual al número de parámetros. Es decir :

$$C_p = \frac{SSE}{\sigma^2} - (n - 2p)$$

Con esperanza $E(C_p) \approx p$. Por lo que se esperaría que el modelo que mejor ajuste tenga un C_p con valor similar al número de parámetros. Cebe aclarar que la p debe incluir al intercepto. De manera que en la siguiente table, el valor de p es el número de parámetros.

A partir del análisis de la anterior tabla, es posible afirmar que el mejor subconjunto de variables que ajustan el modelo es cuando están **todas** variables ("Las tres juntas"), ya que tienen un $C_p = 4$, y el valor de $P = 4$.

Cuadro 1: Comparación modelos con Cp de Mallows

Variables	P	Cp_Mallows	Mejor
Horas		45.5	7
Exámenes de prueba	2	31.4	6
PRomedio		29.3	5
Horas y examanes de prueba		3.4	4
Horas y promedio	3	2.9	2
Exámenes de prueba y promedio		2.7	3
Las tres juntas	4	4.0	1

En segundo lugar, las varibales "Horas y promedio" ($P = 3$) son el segundo mejor subconjunto, con $C_p = 2.9$. En tercer lugar, las variables "Examenes de prueba y promedio" juntas se llevan el tercer lugar. Debido a que $P = 3$ y un valor de $C_p = 2.7$.

En contraste, el modelo exclusivamente que tiene la variables "Horas" es el que más variación tendría, es decir el de peor desempeño. Cabe destacar que la columna "Mejor" indica qué modelo es el que mejor ajusta en función del estadístico de Mallows, con un órden descendente.

Finalmente, cabe destacar que se busca "predecir el puntaje de un estudiante", para lo cuál el estadístico de Mallows no es la mejor herramienta. Ya que el C_p es una medida de sesgos y variación de las las betas del modelo. Si bien esto permitirá evaluar el modelo, se sugiere complementar con pruebas de hipótesis, para poder predicir mejor, y también con la suma de cuadrados de errores predichos (PRESS).

4. Considere el siguinete modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

$ConVAR(e_i) = \frac{\sigma^2}{z_i}$

A. ¿Qué hipótesis del modelo de regresión lineal múltiple viola este modelo?

Este modelo viola en principio de homocedasticidad, el cuál afirma que los errores tienen varianza constante, en los modelos de regresión. Esta violación se puede deber a que hay problemas de especificación, de muestreo o propios del fenómeno a tratar. Cabe recordar que los coeficientes de regresión continúen siendo insesgados, por el método de mínimos cuadrados ordinarios (OLS, por sus siglas en Inglés). Pero no son los mejores, pues su varianza no es la mínima.

B. ¿Qué expresión tendría el modelo ponderado que corrige este modelo?

Una forma de corregir el anterior problema es al aplicar el método de mínimos cuadrados *generalizados* (GLS). El cuál aplica una "transformación al modelo original para que se cumpla el supuesto de homocedasticidad y encontrar los mejores estimadores" (Juarez, 2021). Entonces, se parte de la ecuación inicial:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

Dado que se conocen las varianza heterocedástica, la cuál es: $VAR(e_i) = \frac{\sigma^2}{z_i}$. Es posible, multiplicar el modelo por $\frac{z_i}{\sigma^2}$. Lo que generaría la siguinete expresión:

$$Y_i \frac{z_i}{\sigma^2} = \beta_0 \frac{z_i}{\sigma^2} + \beta_1 X_{1i} \frac{z_i}{\sigma^2} + \beta_2 X_{2i} \frac{z_i}{\sigma^2} + e_i \frac{z_i}{\sigma^2}$$

C. Compruebe que el modelo ponderado del inciso anterior es heterocedástico.

De esta forma,

$$VAR(e_i) = \frac{\sigma^2}{z_i}$$

Y al aplicar la trasnformación:

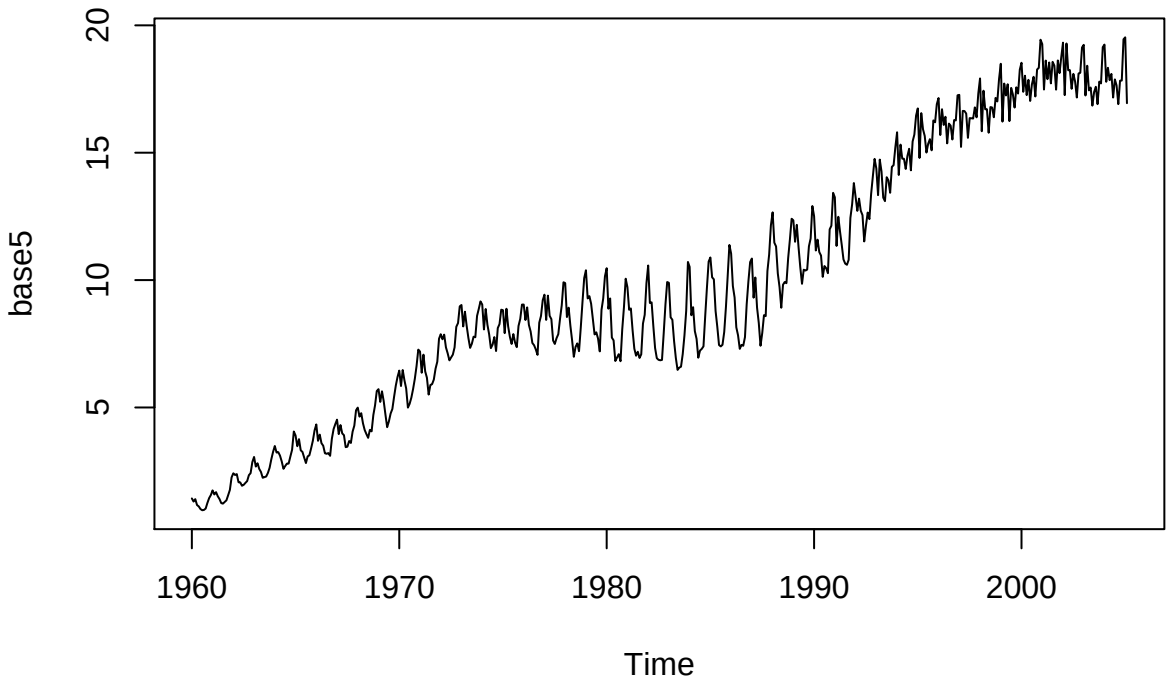
$$VAR(e_i^*) = \frac{\sigma^2}{z_i} \frac{z_i}{\sigma^2} = 1$$

. Lo cuál cumple con los supuestos para los mejores estimadores, según el Teorema de Gauss-Markov, pues los errores tendrían varianza constante.

5. Utilizando el conjunto de datos cangas del paquete expsmooth responda: ¿por qué la transformación de Box-Cox no es de verdadera ayuda?

La base de datos describe una serie de tiempo desde 1960 hasta el año 2005. Se puede percibir que conforme pasan los años, el valor de la variable también aumenta, lo cuál se visualiza en la siguiente gráfica. Además, a la base de datos se le agregó la variable (columna) de años.

```
base5new <- cbind(anios5, base5) #revisar que la ongitud de anios == a
#con anio5 <- rep(1:46, each=12)
anios5 <- anios[1:542]
```

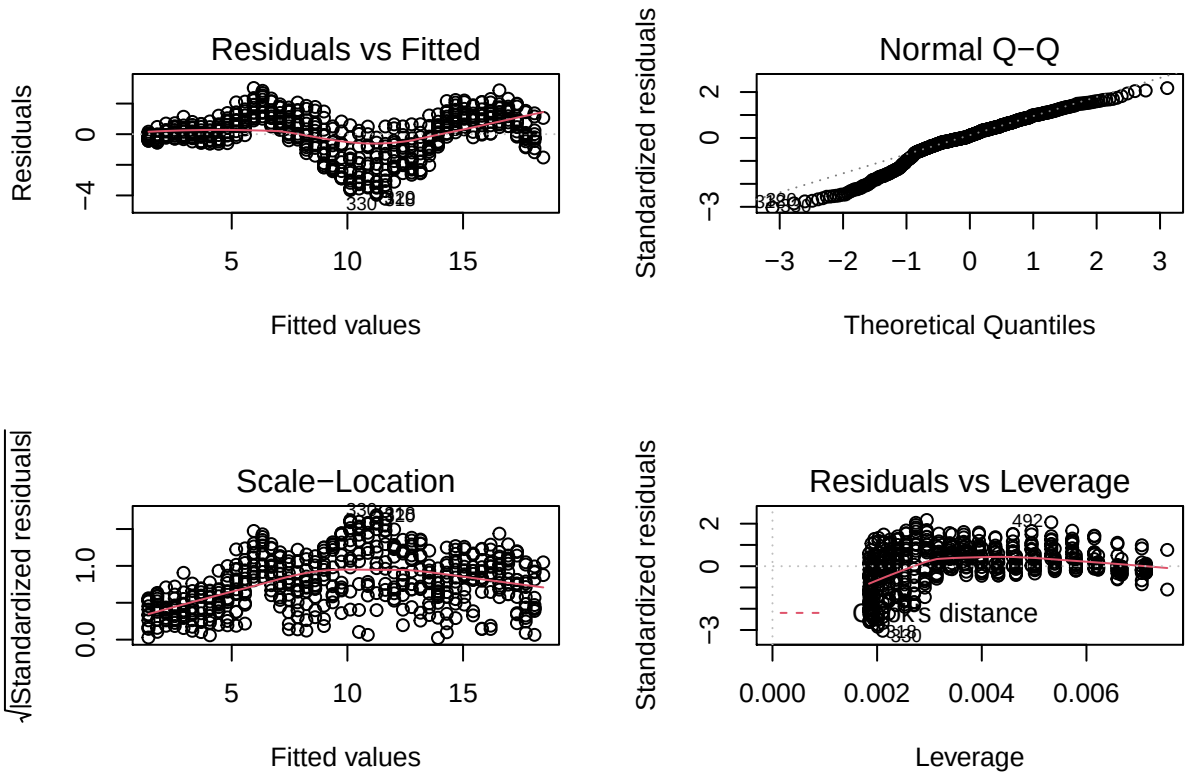


Respecto a la transformación BoxCox, ésta es de gran utilidad cuando se busca transformar los datos hacia una distribución normal y generar una varianza constante. Además, esta transformación busca "encontrar una relación lineal cuando esta es una curva". Desafortunadamente, la gráfica permite visualizar que la hay ciclos en la tendencia, y no solo una curva (Juarez, 2021).

```
##
## Call:
## lm(formula = base5 ~ anios5, data = base5new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2186 -0.6096  0.1336  0.9477  3.0167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.021764   0.121486   8.411 3.66e-16 ***
## anios5      0.379242   0.004582  82.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.391 on 540 degrees of freedom
## Multiple R-squared:  0.9269, Adjusted R-squared:  0.9268
## F-statistic:  6850 on 1 and 540 DF,  p-value: < 2.2e-16
```

Al desarrollar una regresión lineal simple, entre los datos y los años, es posible visualizar que tantos los parámetros son estadísticamente significativos hasta al 99 % de confianza, incluso que la $R^2_{adj} = 0.9268$.
Sin embargo, al revisar gráficamente los supuestos es evidente que el modelo no ajusta. Específicamente, no hay independencia entre residuales y los valores ajustados, tampoco hay homoscedasticidad en el modelo.

Cabe agregar que la normalidad se ve alterada en los cuantiles más extremos.




```
##
##  studentized Breusch-Pagan test
##
## data:  mod5
## BP = 14.888, df = 1, p-value = 0.0001141
```

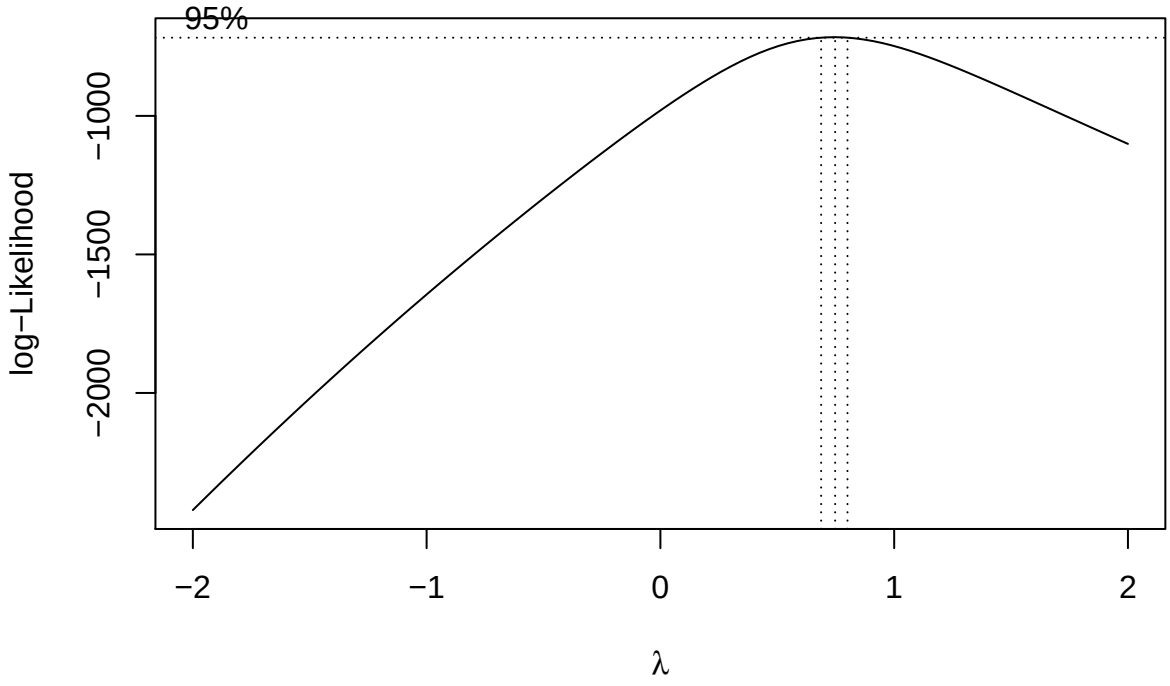
```
##
##  Durbin-Watson test
##
## data:  mod5
## DW = 0.26171, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod5$residuals
## W = 0.95903, p-value = 3.896e-11
```

Además, al realizar las pruebas para los supuestos, es evidente que con un P valor de 0.0001 se rechaza la hipótesis nula de moscedasticidad. Y que con un P valor de práticamente 0, también se rechaza la hipótesis de no autocorrelación de los residuales. Además, de que también se rechaza la hipótesis nula de normalidad con la prueba Shapiro-will.

Entonces, se procede a realizar la prueba Box-Cox, con valor de $\lambda = 0.7474$

```r
bc5 <- boxcox(mod5)
```




```r
bc5$x[ which(bc5$y==max(bc5$y)) ] # 0.7474747
```

```
## [1] 0.7474747
```

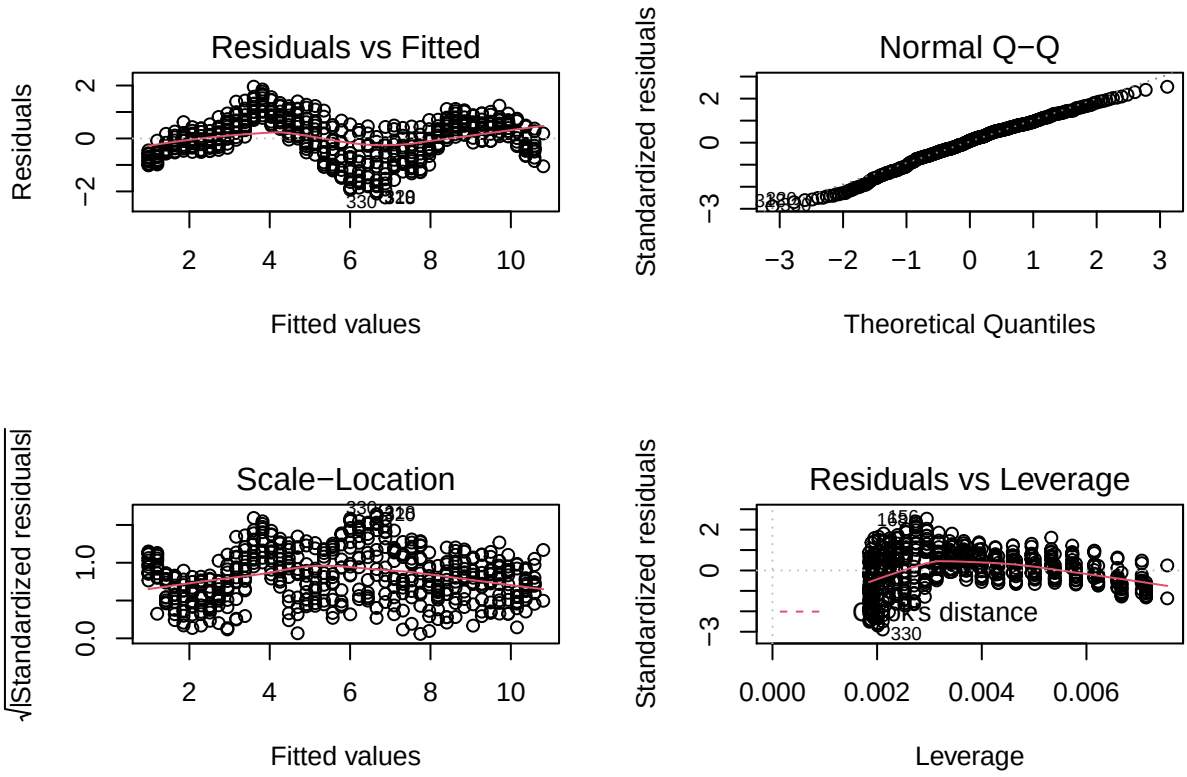
```r
mod5nu <- lm(((base5^0.7474747)-1)/0.7474747 ~ anios5, base5new)

summary(mod5nu)
```

```
##
## Call:
## lm(formula = ((base5^0.7474747) - 1)/0.7474747 ~ anios5, data = base5new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22952 -0.46449  0.06418  0.54583  1.95529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.763577   0.067459   11.32   <2e-16 ***
## anios5      0.218338   0.002544   85.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7723 on 540 degrees of freedom
## Multiple R-squared:  0.9317, Adjusted R-squared:  0.9315
```

```
## F-statistic:  7363 on 1 and 540 DF,  p-value: < 2.2e-16
```

Se observa que el modelo mejoró su coeficiente de determinación. Y también los coeficiente continúan siendo significativos. Además, en la parte inferior se muestran las gráficas de los supuestos que sin duda mejoraron.




```
##
##  studentized Breusch-Pagan test
##
## data:  mod5nu
## BP = 0.19046, df = 1, p-value = 0.6625
```

```
##
##  Durbin-Watson test
##
## data:  mod5nu
## DW = 0.24974, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod5nu$residuals
## W = 0.98897, p-value = 0.0004196
```

Los anteriores resultados muestran que ahora hay homoscedasticidad, con un P valor de 0.6625. Además, obviamente, se arregló el supuesto de normalidad. Sin embargo, continúa habiendo auto-correlación entre los residuales.

Entonces, las causas por las que puede haber auto-correlación entre errores, pueden ser las siguientes:

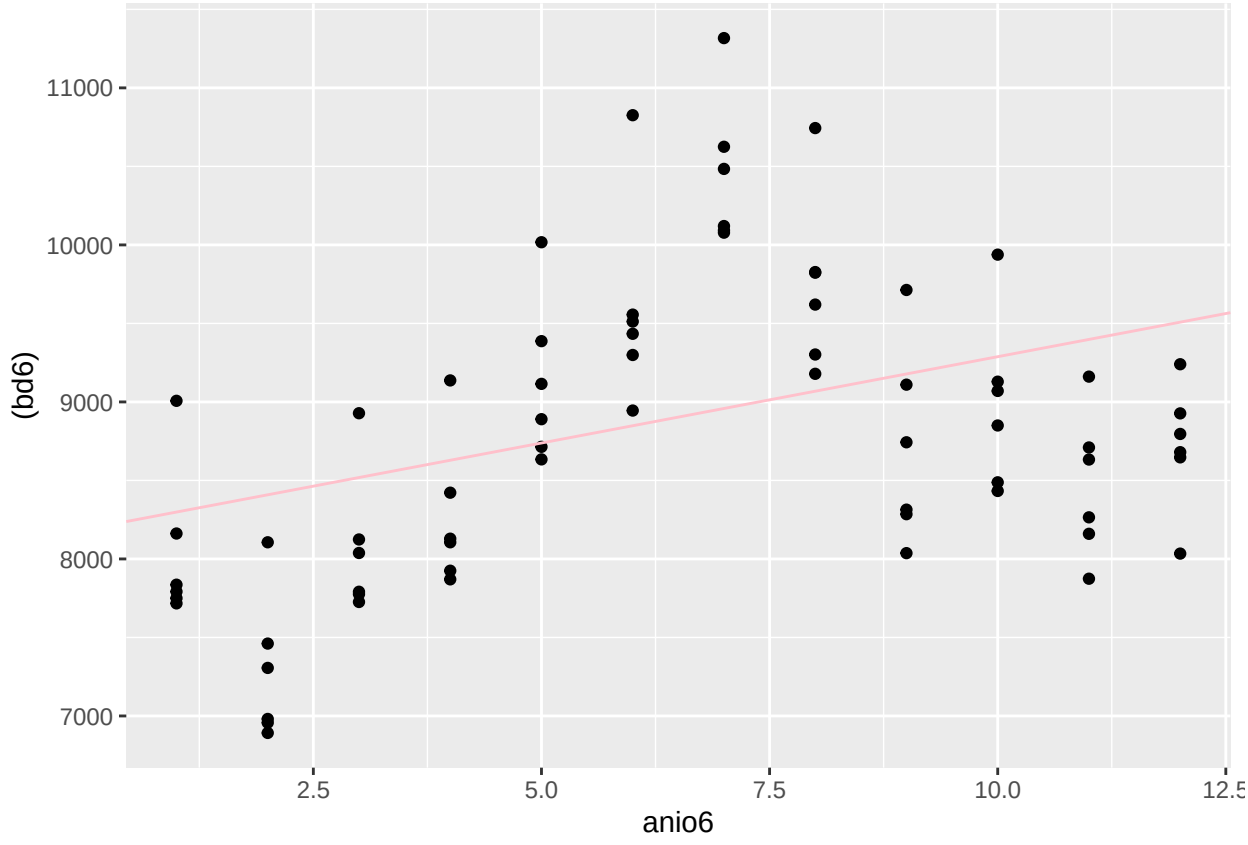
1. La naturaleza del fenómeno a tratar, siendo que haya relación entre errores, como eventos temporales. Y esta es la causa del presente problema.
2. Errores de especificación.

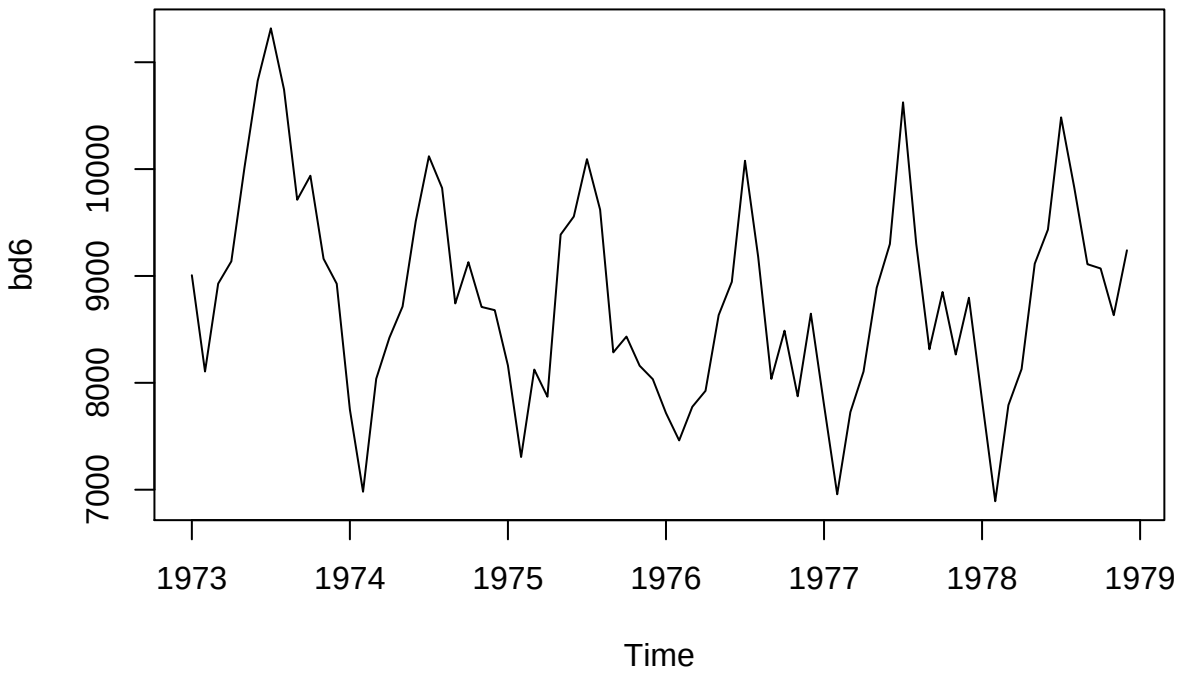
Entonces, a pesar de que los estimadores de regresión son los estimadores insesgados, **no son los estimadores de varianza mínima**, y por tanto no son los mejores. Además, se subestima la σ^2, se sobre estima R^2, tal como se ha presentado. Y las pruebas con las estadísticas t y F "dejan de ser válidas", por lo que la hipótesis de relación lineal entre estimadores deja de ser válida, también (Juarez, 2021).

Finalmente, es posible sostener que dado que la autocorrelación proviene del hecho de que los datos son temporales, la transformación Box-Cox deja de ser útil para mejorar el modelo. Esto se debe a que esta transformación "mueve" los datos hacia la normalidad, mas no arregla el supuesto de correlación entre errores. Pues la correlación se debe a que el fenómeno es temporal. Es más, se requiere de modelar los errores respecto al tiempo, del método de mínimo cuadrados generalizados (GLS) o de otras transformaciones donde la covarianza de los errores se corriga.

6. Para el conjuntos de datos usdeaths del paquete fma

A. GRafique los datos





Es evidente que se trata de una base de datos de información temporal. Si se traza una recta de regresión, es evidente que hay gran distancia entre las observaciones y la recta, lo cuál podría llevar a pensar que el modelo no ajusta bien.

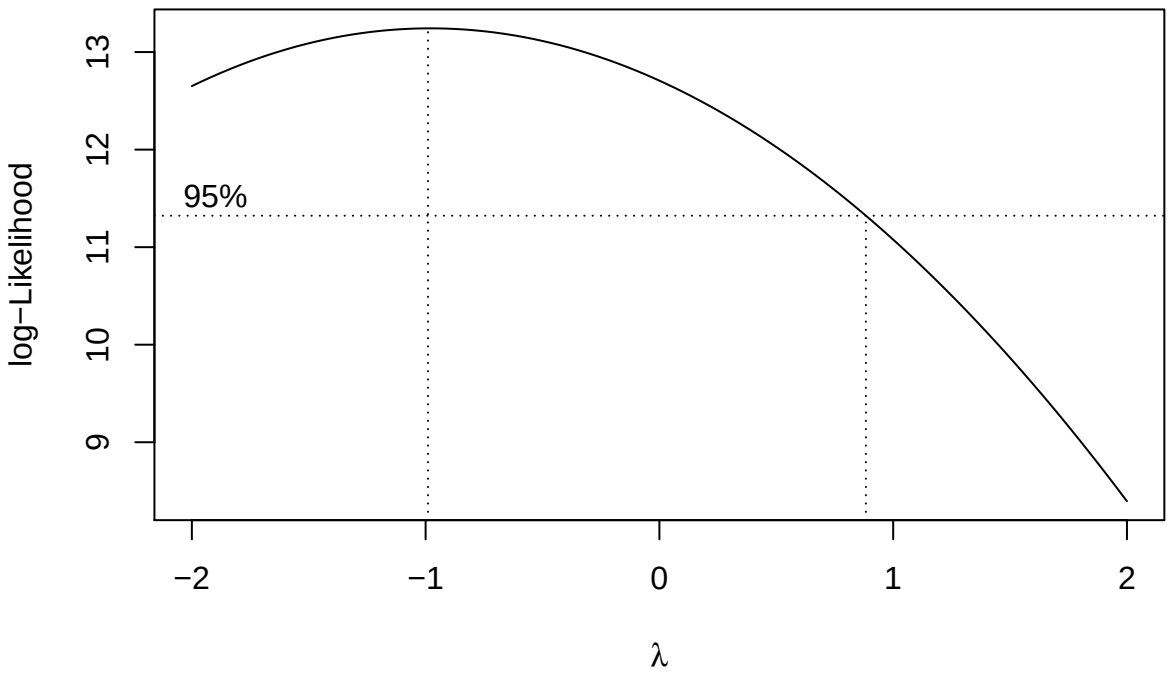
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6892    8089    8728    8788    9323   11317
```

```
##  Time-Series [1:72] from 1973 to 1979: 9007 8106 8928 9137 10017 ...
```

Lo anterior se confirmar con la función *str* , la cual describe que los datos son recogidos entre 1973 a 1979. Y el tamaño es de 72 unidades, donde cada año observado es integrado por 12 meses, por lo que hay 6 años enla base.

B. Si cree que es apropiado realizar una transformación para ajustar un modelo lineal, aplíquela y describa los efectos de su aplicación.

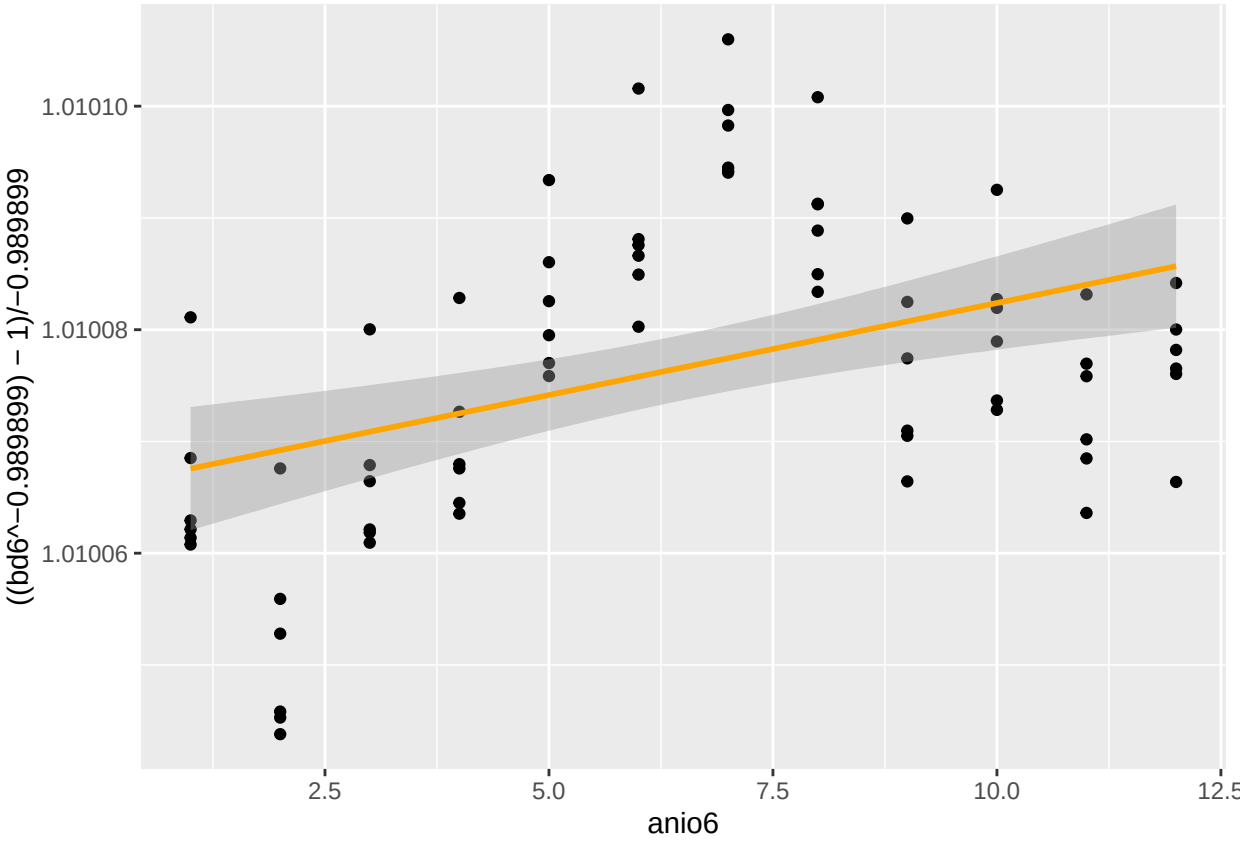
Se aplica un modelo de regresión lineal simple. A pesar de que los estimadores son estadísticamente significativos, el coeficiente de determinación es *muy mal*, pues es de 0.13. Respecto a los supuestos básciso, es posible identificar que los datos son normales, pero los extremos se alejan; parece que no hay datos *outliers* influyentes; parece que no hay heteroscedasticidad, pero que los datos con identificación 7, 6, y 8 podrían incidir en la curva central roja; finalmente, parece haber tendencia entre los residuales y los datos ajustados, lo que se muestra como una *colina*, y los mismos datos (6,7,8) ,son los más incluyentes. Entonces, se procede a realizar una transformación para mejorar el ajuste del modelo, y comprobar la relación entre residuales y datos.



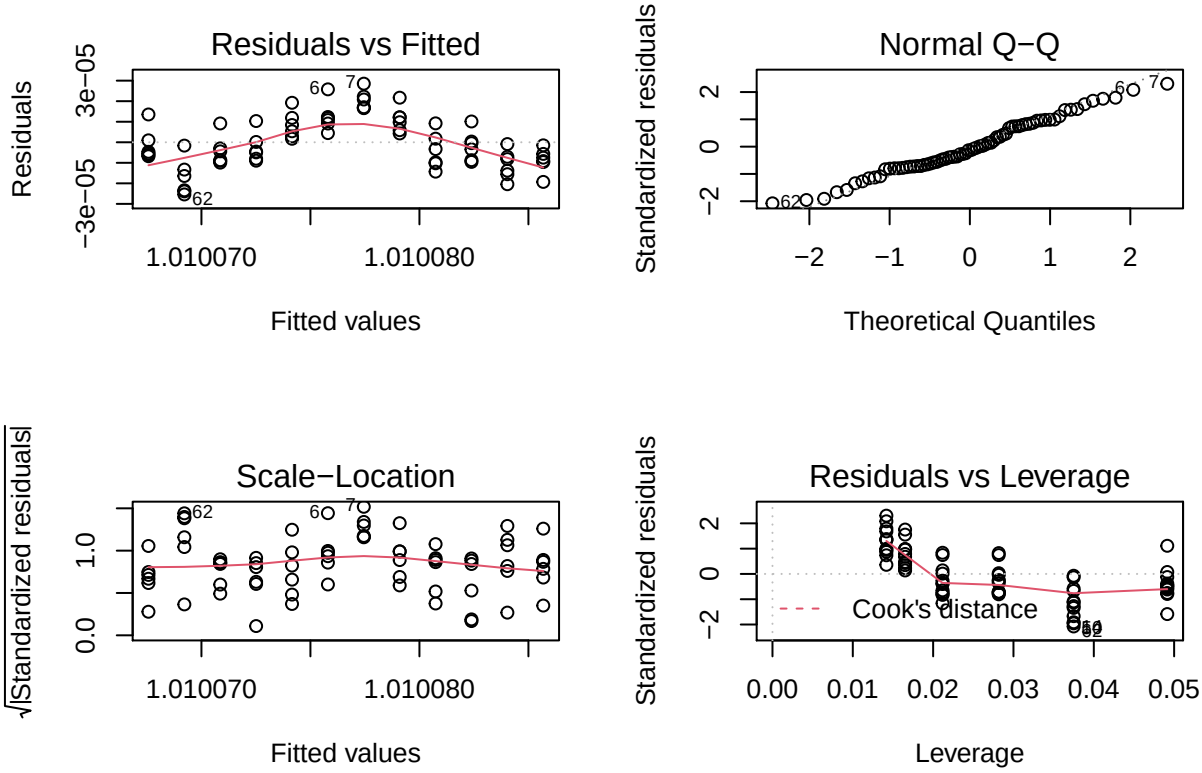

```
## [1] -0.989899
```

```
##
## Call:
## lm(formula = ((bd6^-0.989899) - 1)/-0.989899 ~ anio6)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.541e-05 -8.797e-06 -1.567e-06  9.861e-06  2.853e-05
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.010e+00  3.134e-06 3.223e+05  < 2e-16 ***
## anio6       1.646e-06  4.259e-07 3.866e+00 0.000245 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.247e-05 on 70 degrees of freedom
## Multiple R-squared:  0.1759, Adjusted R-squared:  0.1642
## F-statistic: 14.95 on 1 and 70 DF,  p-value: 0.0002448
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Primero se realiza una transformación box-cox, con $y_i^{\lambda} = y_i^{-0.989899}$. Aunque los estimadores continúan siendo significativos al 95 % de confianza, con P valores de 0.0000 y 0.0002, es evidente que el modelo de mejora su ajuste, pues la $R^2 = 0.17$. Además, se grafica la recta de regresión y es visible la gran distancia entre los datos y la recta.




```
##
##  studentized Breusch-Pagan test
##
## data:  mod6new
## BP = 0.21444, df = 1, p-value = 0.6433
```

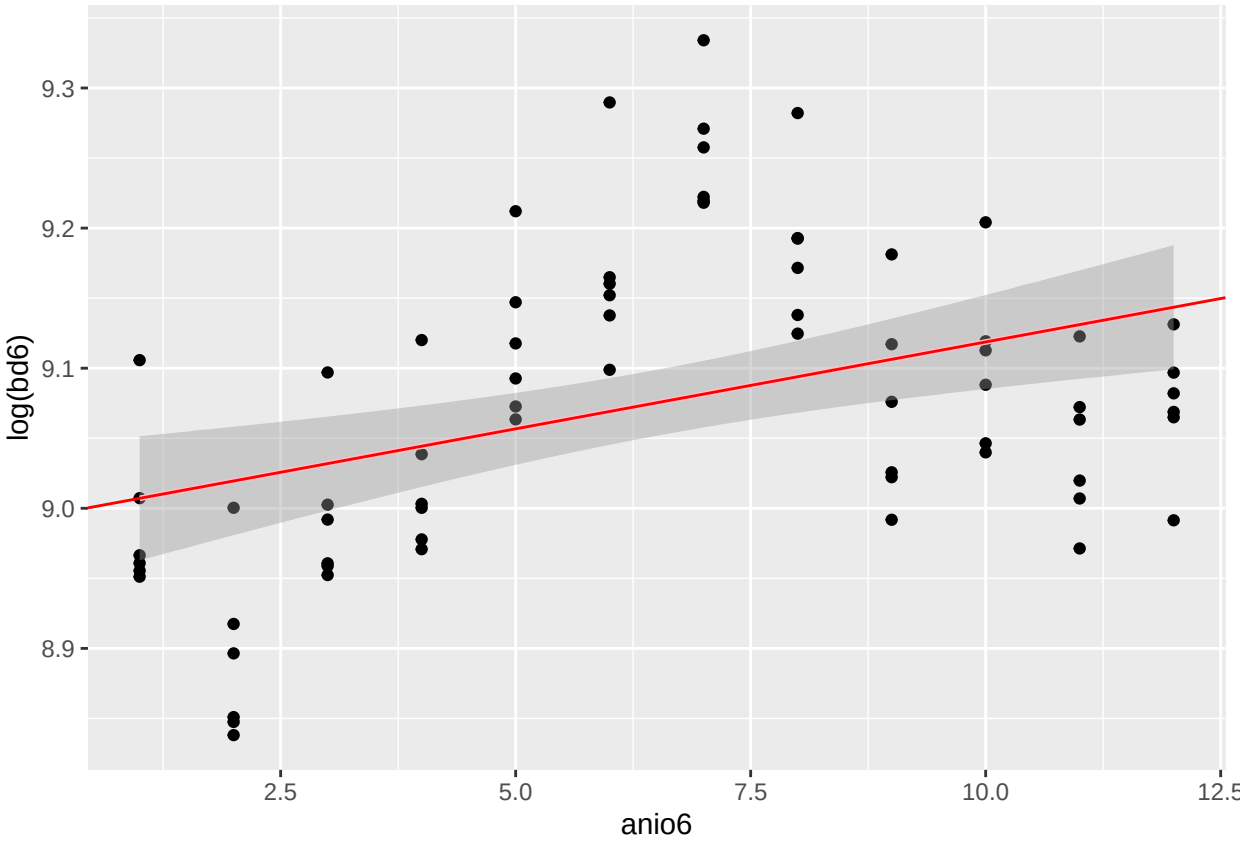
```
##
##  Durbin-Watson test
##
## data:  mod6new
## DW = 0.7024, p-value = 1.1e-10
## alternative hypothesis: true autocorrelation is greater than 0
```

Respecto a los supuestos, parece que los supuestos de normalidad, outliers influyentes y homocedasticidad permanecen constantes, pero la correlación de residuales continúa siendo violada. Por lo que se procede a realizar pruebas de hipótesis. Se realiza la prueba Breush Pagan y no se rechaza la hipótesis nula de homocedasticidad con un P valor de 0.6433. Pero sí se rechaza la hipótesis nula de no correlación de errores con la prueba Durbin-Watson al 95 % de confianza, y un P valor de 0.0000.

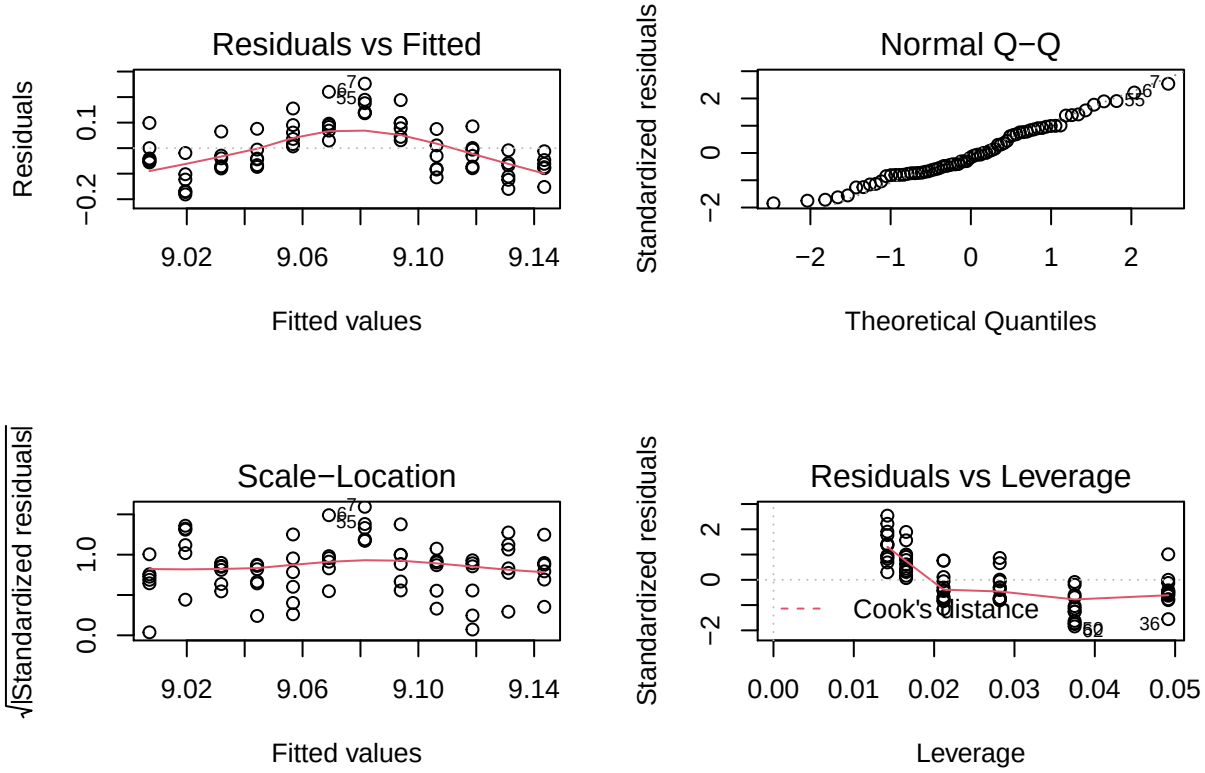
```
##
## Call:
## lm(formula = log(bd6) ~ anio6)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.18139 -0.07250 -0.01567  0.07626  0.25256
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.99471    0.02517 357.324  < 2e-16 ***
## anio6        0.01240    0.00342   3.625 0.000544 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1002 on 70 degrees of freedom
## Multiple R-squared:  0.1581, Adjusted R-squared:  0.146
## F-statistic: 13.14 on 1 and 70 DF,  p-value: 0.0005438
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Segundo, se realiza la transformación del logaritmo a la variable respuesta, y el escenario es similar a los anteriores: el $R^2 = 0.15$, pero los estadísticos siguen siendo significativos. También, la gráfica con la recta de regresión parece haber gran distancia entre los datos y la recta.




```
##
##  studentized Breusch-Pagan test
##
## data:  mod6log
## BP = 0.0014583, df = 1, p-value = 0.9695
```

```
##
##  Durbin-Watson test
##
## data:  mod6log
## DW = 0.66132, p-value = 1.863e-11
## alternative hypothesis: true autocorrelation is greater than 0
```

Respecto a los supuestos, parece que los supuestos de normalidad, outliers influyentes y homocedasticidad siguen constantes, empero la correlación de residuales es incumplida todavía. Por lo que se procede a hacer pruebas de hipótesis. Se realiza la prueba Breusch Pagan y no se rechaza la hipótesis nula de homocedasticidad con un P valor de 0.9695. Pero sí se rechaza la hipótesis nula de no correlación de errores con la prueba Durbin-Watson al 95 % de confianza, y un P valor de 0.0000.

Finalmente, a pesar de que el modelo de regresión con los datos originales y en las dos transformaciones muestran que los estimadores son estadísticamente significativos, hay coeficientes de determinación muy bajos, entre el 0.1 al 0.2. Esto se debe a que los errores están correlados, pues los datos son temporales, y esto se compueba con las pruebas de Durbin-Watson. El efecto que tiene esto en el modelo es que sí es posible estimar los parametros, pero sus varianzas no son las mínimas, ya que éstas dependen de $Cov(e_i, e_j) \neq 0$, y por lo tanto no son los mejores estimadores por Gauss-Markov. Esto también se debe a que los datos al ser temporales tienen varianzas que dependen del tiempo. Además, esto repercute en que los Intervalo de confianza serán amplios y las pruebas no serán *confiables*. Para arreglar esto, sugiere aplicar mínimos cuadrados generalizados, aumentar el tamaño de la muestra y utilizar técnicas propias para datos temporales.

7. Utilice el conjunto de datos seatpos del paquete faraway para encontrar el mejor modelo lineal que describa a la variable hipcenter. Compare este modelo con el modelo que incluye a todas las variables independientes del conjunto.

La base de datos está compuesta por 9 variables, donde la variables **hipcenter** es la endógena y describe la "distancia horizontal del punto medio de la cadera desde una ubicación fija en el carro medida en milímetros" (Faraway, 2004). Todas son numéricas o de valores enteros.

A continuación se muestra un correlograma, donde se tacha a las correlaciones que no son significativas al 95 % de confianza. Cabe destacar, a excepción de *age*, todas las variables están fuertemente correlacionadas con **hipcenter**, y también entre sí. Esto podría ser un indicio para excluir a la variable *age* del modelo de regresión, para que ajuste bien.

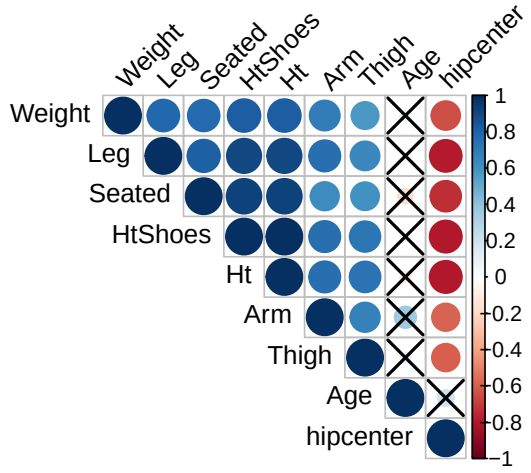
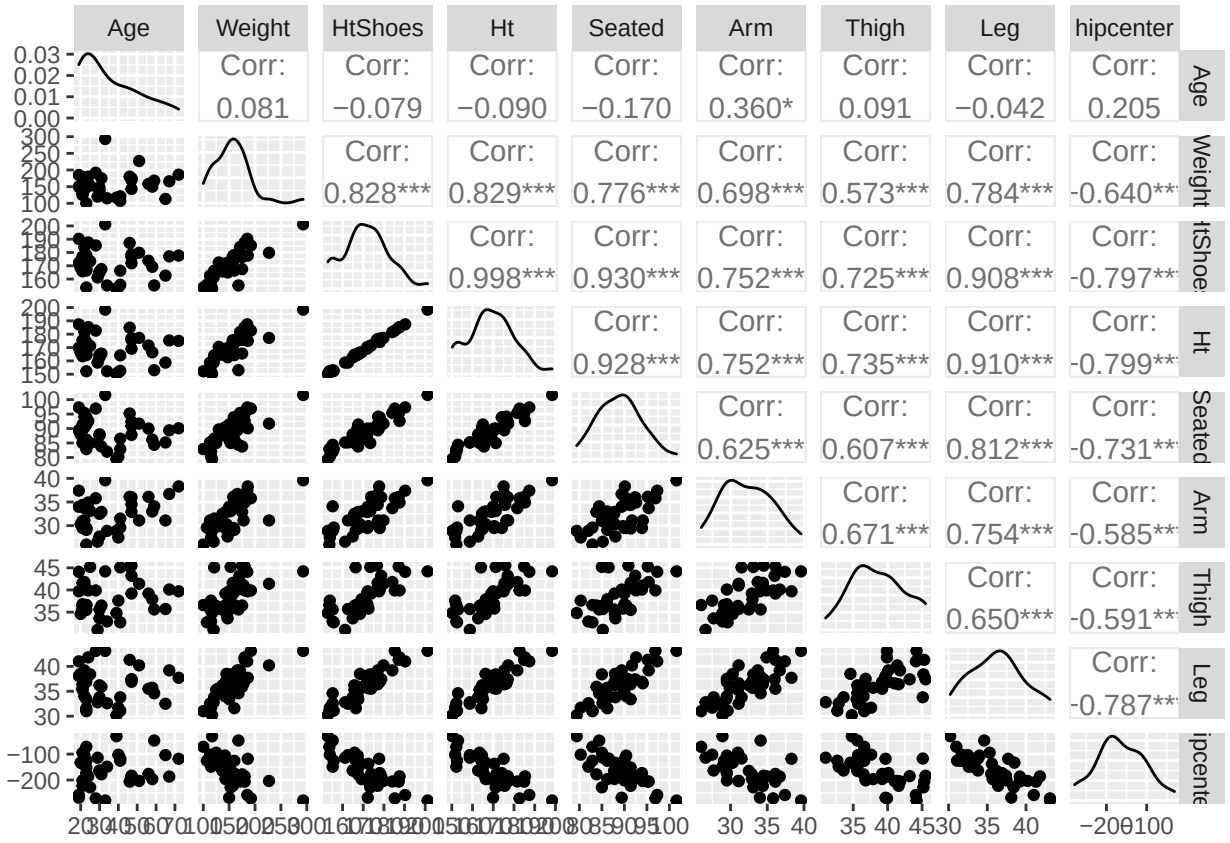
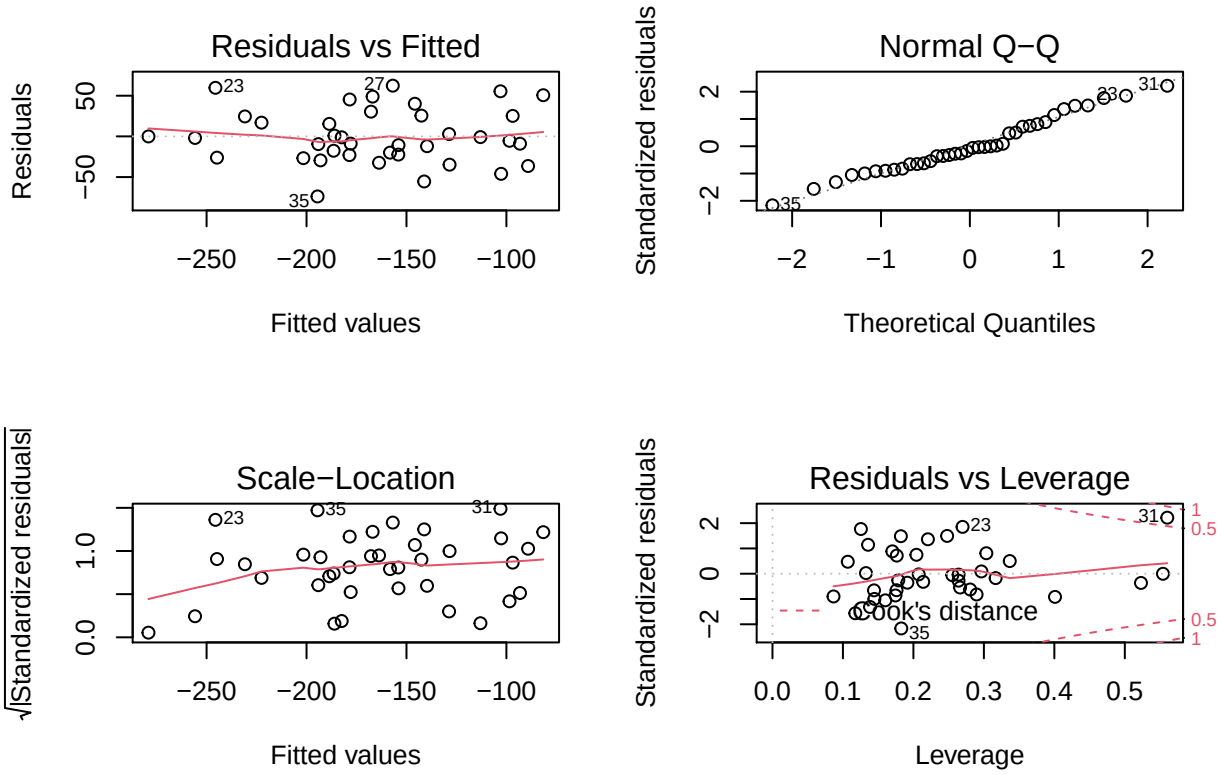


Figura 1: Correlograma con significancia




```
##
## Call:
## lm(formula = hipcenter ~ ., data = bd7, x = T, y = T)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```



Luego de aplicar una regresión múltiple con todas las variables independientes, el modelo tiene un coeficiente de regresión bajo, y a excepción del intercepto, ningun coeficiente de regresión parcial es significativo. También, el rango de los resoduales recorre desde -73 hasta 62, lo que se puede considerar un amplio margen. Esto se podría deber a errores de especificación de las variables que describen el modelo. Respecto a los supuestos, gráficamente parece haber independencia de residuales, y también cumplir laxamente la normalidad y homoscedasticidad. Sin embargo, sí hay outliers influyentes, el cuál es el punto 31, 35 y 23. Se procede a realizar pruebas para los anteriores supuestos.

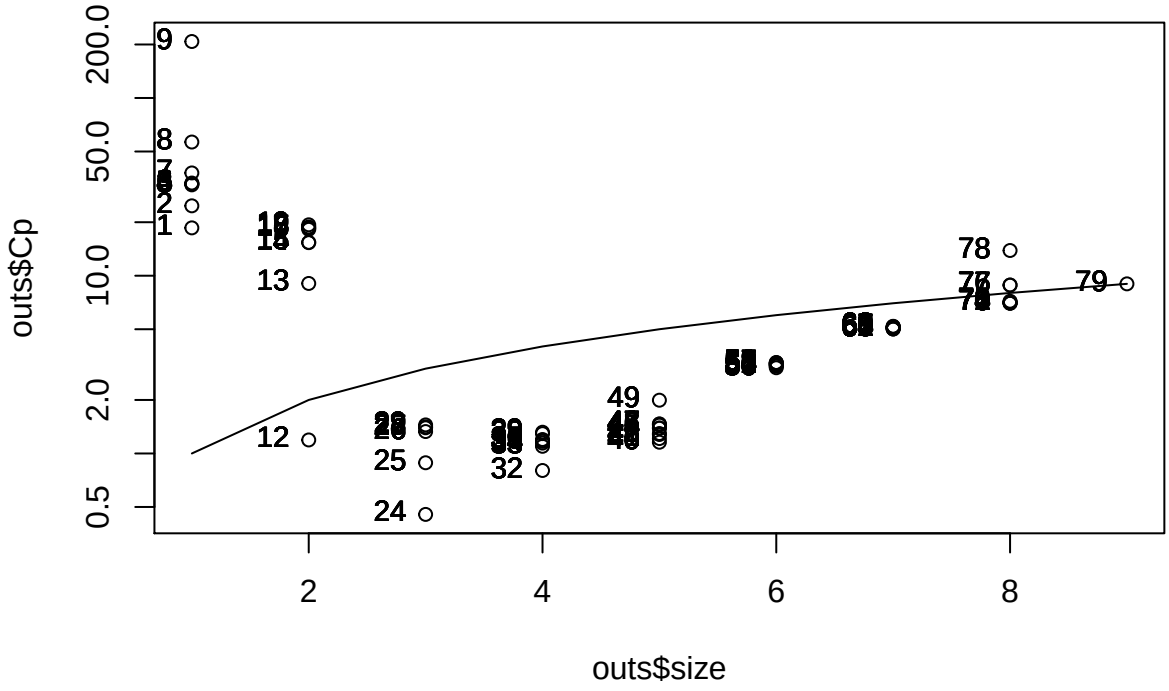
```
##
##  studentized Breusch-Pagan test
##
## data:  complet7
## BP = 14.037, df = 8, p-value = 0.0808
```

```
##
##  Durbin-Watson test
##
## data:  complet7
## DW = 1.7688, p-value = 0.2441
## alternative hypothesis: true autocorrelation is greater than 0
```

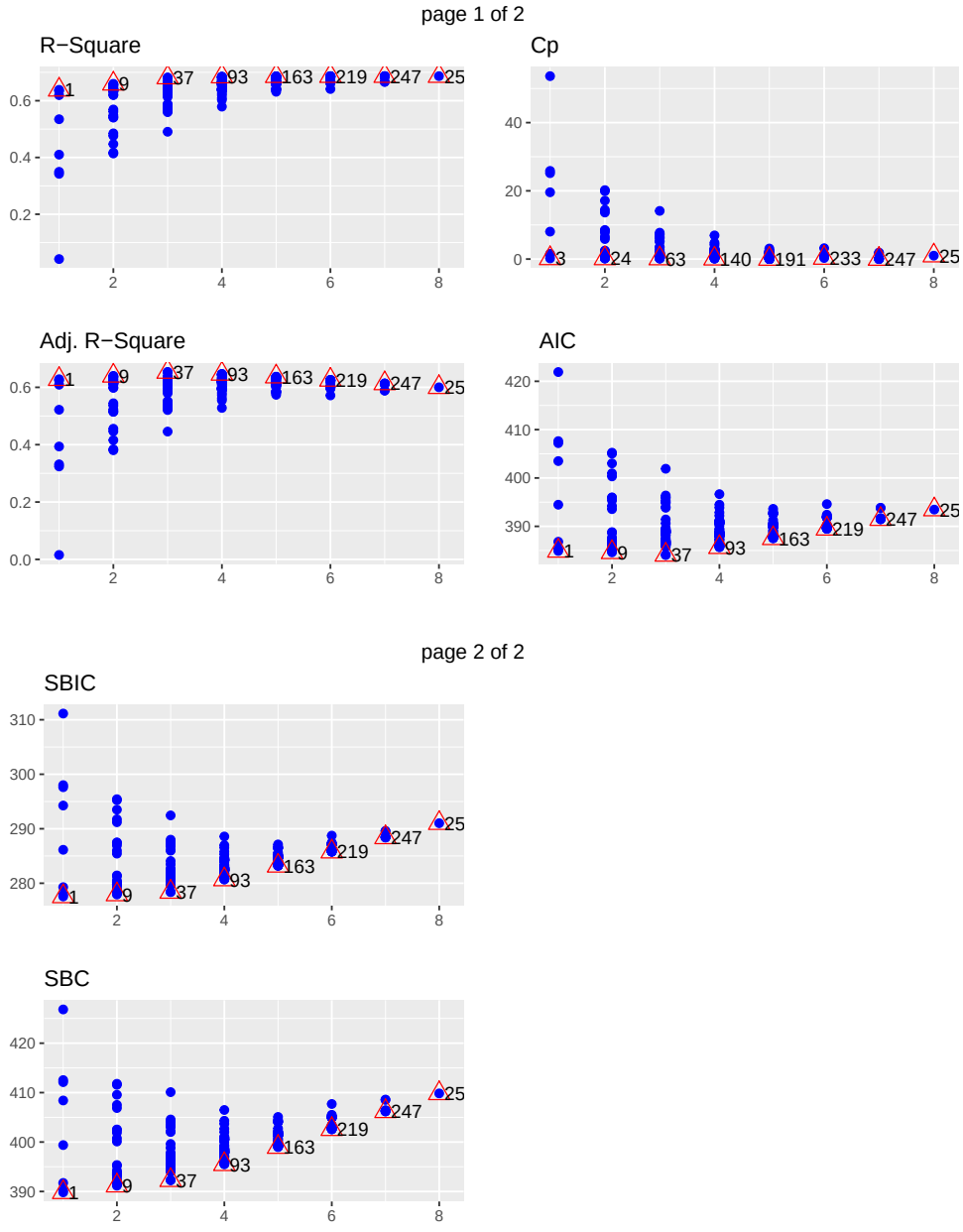
El test con hipótesis nula No correlación no se rechaza, con un P valor de 0.244. La prueba de hipótesis nula de homscedasticidad no se rechaza, pero está en la frontera con un P valor de 0.08. Cabe recordar que este último problema se puede deber a outliers influyentes, errores de especificación, entre otros.

Entonces, a continuación se procede a encontrar el mejor modelo. Cabe señalar que se debería de modela con respaldo de un experto de área, y este debería ser el principal criterio para crear modelos. Sin embargo, se procede a seleccionar el modelo a través de la Cp de Mallows.

```
outs<-leaps(complet7$x, bd7$hipcenter, int=FALSE) #output: que V.A. incluye modelo, $labels son los var
plot(outs$size,outs$Cp, log="y",cex=0.9)
lines(outs$size,outs$size)
text(outs$size, outs$Cp, labels=row(outs$which), cex=0.9, pos=2)
```



El análisis con la Cp de Mallows impica que se seleccionará aquel número cuyo valor sea similar a la cantidad de parámetros del modelos, ya que $E(C_p) \approx p$. En este sentido, y según la gráfica, los modelos 12, 75,79,72,78 son los mejores para este criterio.



Por otra parte, es necesario comparar más que sólamente un indicador como la Cp de Mallow. Afortunadamente la anterior gráfica muestra comparaciones para diferentes índicadores (como la R^2_{ajust}, la Cp, o el criterio AIC). Es posible visualizar lo siguiente:

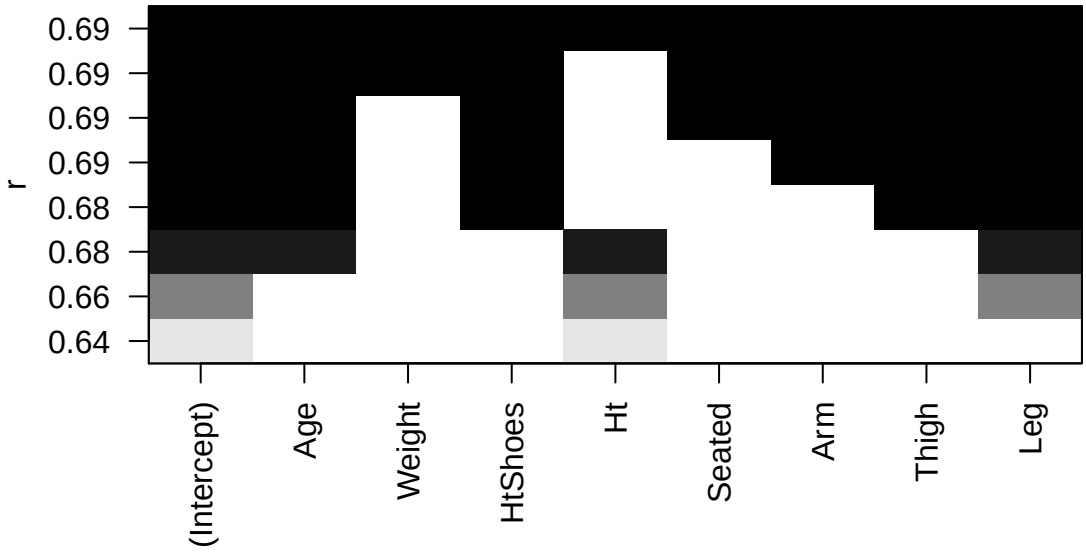
1. Cuando se tienen todos los parámetros (sin incluir β_0), el modelo "25" es el mejor. Ya que es el único.
2. Cuando se tienen 7 párametros (sin incluir β_0), el mejor modelo es el "247". Esto lo afirman los criterios de Cp, R cuadrada ajustada y el criterio AIC. Este modelo está integrado por las variables: Age, Weight, HtShoes, Arm, Thigh, Leg.
3. Al poseer 6 párametro (sin incluir β_0), el modelo "219", el mejor según los criterios R cuadrada ajustada y el criterio AIC. Este modelo está integrado por las variables: Age, HtShoes, Seated, Arm, Thigh, Leg.
4. Al poseer 5 párametro (sin incluir β_0), el modelo "163", el mejor según los criterios R cuadrada ajustada y el criterio AIC. Este modelo está integrado por las variables: Age, HtShoes, Arm, Thigh, Leg.

Curiosamente, las variables Age, HtShoes, Arm, Thigh, Leg se mantienen en los 4 modelos anteriores.

```
best7 <- regsubsets(hipcenter ~ . , data = bd7)
summary(best7)
```

```
## Subset selection object
## Call: regsubsets.formula(hipcenter ~ ., data = bd7)
## 8 Variables  (and intercept)
##         Forced in Forced out
## Age         FALSE      FALSE
## Weight      FALSE      FALSE
## HtShoes     FALSE      FALSE
## Ht          FALSE      FALSE
## Seated      FALSE      FALSE
## Arm         FALSE      FALSE
## Thigh       FALSE      FALSE
## Leg         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Age Weight HtShoes Ht  Seated Arm Thigh Leg
## 1  ( 1 ) " " " "    " "     "*" " "    " " " "   " "
## 2  ( 1 ) " " " "    " "     "*" " "    " " " "   "*"
## 3  ( 1 ) "*" " "    " "     "*" " "    " " " "   "*"
## 4  ( 1 ) "*" " "    "*"     " " " "    " " "*"   "*"
## 5  ( 1 ) "*" " "    "*"     " " " "    "*" "*"   "*"
## 6  ( 1 ) "*" " "    "*"     " " "*"    "*" "*"   "*"
## 7  ( 1 ) "*" "*"    "*"     " " "*"    "*" "*"   "*"
## 8  ( 1 ) "*" "*"    "*"     "*" "*"    "*" "*"   "*"
```

```
plot(best7, scale = "r")
```



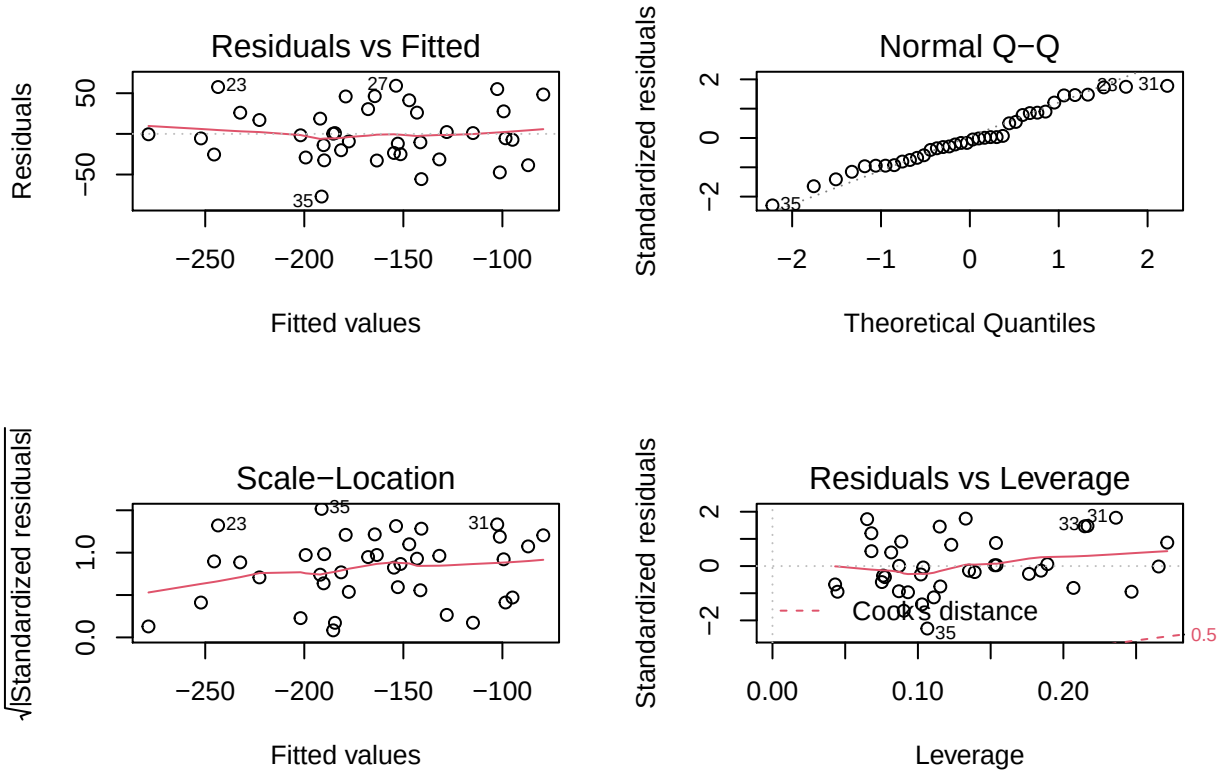
Tal como se indicó con el análisis de los indicadores, las variables: Age, HtShoes, Arm, Thigh, Leg también aparecen en el modelo de 5 coeficientes de regresión parcial (sin incluir β_0). Sin embargo, si se buscara solamente 4 coeficientes, se podría excluir el β_{arm}, esto en función de los resultados brindados por el análisis con la función *regsubsets*.

```
## [1] "Age"       "Weight"    "HtShoes"   "Ht"        "Seated"    "Arm"
## [7] "Thigh"     "Leg"       "hipcenter"
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + HtShoes + Thigh + Leg, data = bd7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.069 -24.643  -3.584  26.092  59.182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 445.7977   105.1452   4.240  0.00017 ***
## Age           0.6525     0.3910   1.669  0.10462
## HtShoes      -1.9171     1.4050  -1.365  0.18164
## Thigh        -1.3732     2.2392  -0.613  0.54391
## Leg          -6.9502     4.1118  -1.690  0.10040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.46 on 33 degrees of freedom
## Multiple R-squared:  0.6849, Adjusted R-squared:  0.6467
```

```
## F-statistic: 17.93 on 4 and 33 DF,  p-value: 6.535e-08
```

Al compara los modelos de cuatro y cincos variables sin "Arm", es evidente que el modelo de 4 variables, tiene un coeficiente de determinación ligeramente mayor, y en ambos casos sólamente β_0 es estadísticamente significativo.




```
## [1] "Age"       "Weight"    "HtShoes"   "Ht"        "Seated"    "Arm"
## [7] "Thigh"     "Leg"       "hipcenter"
```

Esto aún puede deberse a que había outliers influyentes y a que el supuesto de homoscedasticidad esté en la *frontera*. Pero al graficar los supuestos, ya nose observan outliers. Y al realizar la prueba, ya el supuesto de homoscedasticidad ha mejorado con un P valor de 0.1265.

```
##
##  studentized Breusch-Pagan test
##
## data:  mod74
## BP = 7.1835, df = 4, p-value = 0.1265
```

```
##
##  Durbin-Watson test
##
## data:  mod74
## DW = 1.8079, p-value = 0.2863
## alternative hypothesis: true autocorrelation is greater than 0
```

```
##
## Call:
## lm(formula = hipcenter ~ ., data = bd7, x = T, y = T)
##
## Coefficients:
```

```
## (Intercept)          Age       Weight      HtShoes           Ht       Seated
##   436.43213      0.77572      0.02631     -2.69241      0.60134      0.53375
##         Arm        Thigh          Leg
##    -1.32807     -1.14312     -6.43905
```

```
##
##                             Selection Summary
## --------------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Entered     R-Square    R-Square     C(p)        AIC         RMSE
## --------------------------------------------------------------------------------
##    1    Ht            0.6383      0.6282    -0.5342    384.9060    36.3684
##    2    Leg           0.6594      0.6399    -0.4889    384.6191    35.7909
##    3    Age           0.6814      0.6533    -0.5247    384.0811    35.1208
## --------------------------------------------------------------------------------
```

Se parte del modelo con una sóla variable explicativa: "Ht", con un$R^2 = 0.799$, y con un P valor de 0.051. Lo cuál es coherete con los modelos anteriores, donde Ht era seleccionado para modelos con menos de 3 variables independientes. Posteriomente, se incorpora la variables *Leg*, cabe destacar que el coeficiente de determinación ya converge al que se ha presentado previamente, pues $R^2_{adj} = 0.659$, desafortunadamente el significancia se va hasta 0.15. Luego, se agrega la variable Age. Pero en este punto las variablese ya no son significativas, y el R^2 no supera el 0.68. Y así termina la modelación *fordward*.

```
##
##
##                            Elimination Summary
## --------------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Removed     R-Square    R-Square     C(p)        AIC         RMSE
## --------------------------------------------------------------------------------
##    1    Ht            0.6865      0.6134     7.0035    391.4680    37.0885
##    2    Weight        0.6864      0.6257     5.0113    389.4782    36.4903
##    3    Seated        0.6862      0.6371     3.0360    387.5105    35.9309
##    4    Arm           0.6849      0.6467     1.1569    385.6684    35.4559
##    5    Thigh         0.6813      0.6531    -0.5108    384.0990    35.1291
## --------------------------------------------------------------------------------
```

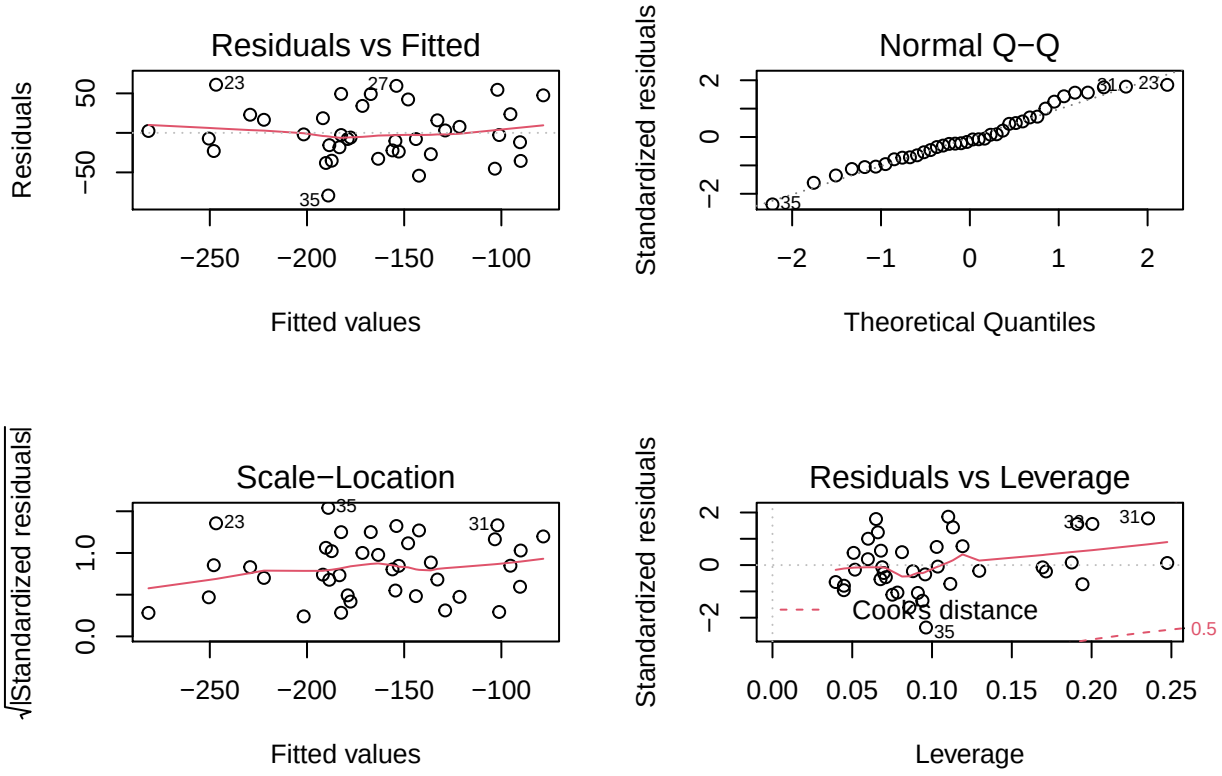
Respecto al método "hacia atrás", la primer variable en ser eliminada es *Ht*. Cabe señalar, que ninguna de las variables independientes son estadísticamente significativas. Posteriormente se quitan las variables *Weight* , *Seated* , *Arm*, y *Thingh* en ese órden. El valor de R^2 varía entre 0.61 a 0.65. Finalmente, el modelo seleccionado está integrado por Age, HtShoes, y Leg. Con un $R^2_{adj} = 0.65$

```
mod73 <- lm(hipcenter ~ Age + HtShoes + Leg, data = bd7)
summary(mod73)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + HtShoes + Leg, data = bd7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.269 -22.770  -4.342  21.853  60.907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 456.2137   102.8078   4.438 9.09e-05 ***
## Age           0.5998     0.3779   1.587   0.1217
```

```
## HtShoes      -2.3023     1.2452  -1.849   0.0732 .
## Leg          -6.8297     4.0693  -1.678   0.1024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.13 on 34 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.6531
## F-statistic: 24.22 on 3 and 34 DF,  p-value: 1.437e-08
```

Se muestra que el modelo de tres variables: Age, HtShies y Leg, tiene un coeficiente de determinación de 0.6531. Y que sólamente la variables HtShoes es significativamente al 90 %. Además, si se verifican los residuales, se puede evidenciar que el rango recorre desde -79 hasta 60, lo cuál se puede considerar un rango amplio.



Al comprobar gráficamente los supuestos, no aparecen outliers que sean influyentes, y parece que los tres supuestos restantes se cumplen someramente bien.

Cuadro 2: Comparación modelos con R ajustada y AIC

Indicador	ModeloCompleto	Modelo4variables	Modelo3variables	Modelo2varialbes
R Ajustada	0.6001	0.6467	0.6531	0.6399
AIC	393.4634	384.0990	385.6684	384.6191

Finalmente, es posible arguir que el modelo de tres variables es el que presenta mejores indicadores. Ya tiene el coeficiente de determinación más alto, y el un punto más alto que los modelos con 2 y 4 variables. Asimismo, los modelos reducidos no cuentan con los problema de outliers influyentes ni de homoscedasticidad en la frontera del P value. Sin embargo, se evidenció que en ningún modelo, las variables son estadísticamente significativas al 95 %. Esto se puede deber a falta de especificación, es decir que ninguna variables describe de forma correcta a *hipcenter*. Además, se mostró que los residuales tienen un amplio rango. Entonces, se sugiere que haya un especialista de área que oriente la formulación del modelo.

Referencias

[1] Juarez, Claudia, *Análisis de regresión*, IIMAS, UNAM. (2021)

[2] Pardoe, Iain, *Best subsets regression, Adjusted R-sq, Mallows Cp*, Course of the Department of Statistics, The Pennsylvania State University, (2018). Obtenido desde https://online.stat.psu.edu/stat462/node/197/ el 8 de junio de 2021.