

Regresión Avanzada

Universidad Austral

PhD. Débora Chan
Junio-Julio 2023

Facultad de Ingeniería

Organización

1 Análisis Diagnóstico

- Los Residuos
- Validación de los Supuestos
- Detección de Valores atípicos e Influyentes

2 Cuadrados Mínimos Ponderados

3 Modelos Robustos

Facultad de Ingeniería

Objetivos

Luego de Ajustar el Modelo

El análisis diagnóstico tiene por objetivo la validación de los supuestos del modelo, en este caso el modelo de regresión lineal simple.

Se observa fundamentalmente el comportamiento de los residuos del modelo. Es importante verificar que el ajuste no dependa en forma central de un pequeño subconjunto de datos y si así fuera resultará de interés detectar esos datos a los que luego llamaremos observaciones influyentes.

Los residuos del modelo

El análisis de los residuales en los modelos de regresión permite:

- a) Cuantificar la bondad de ajuste del modelo al patrón de los datos.
- b) Verificar el cumplimiento de los supuestos del modelo.

Hemos definido los residuos en el caso de la regresión lineal simple como $e_i = Y_i - a - bX_i$ o bien matricialmente como $e = Y - X\hat{\beta}$.

Facultad de Ingeniería

Residuales: Propiedad ①

Los residuos tienen esperanza nula; es decir



$$E(e) = 0$$

En efecto:

$$e = Y - X\hat{\beta}$$

Entonces:

$$E(e) = E(Y - X\hat{\beta}) = E(Y) - E(Y) = 0$$

Residuales: Propiedad ②

Los residuos tienen promedio nulo:



$$\sum_{i=1}^n e_i = 0$$

Entonces:



$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$$

Lo que equivale a que:



$$\bar{Y} = \overline{\hat{Y}}$$

Residuales: Propiedad ③

Los residuos son ortogonales a las X 's.



$$E(eX) = 0$$

En efecto:

$$E(X^t e) = E(X^t (Y - X\hat{\beta})) = E(X^t (Y - X\hat{\beta})) = E(X^t Y - X^t X\hat{\beta})$$

$$E(X^t e) = E(X^t Y - \underbrace{X^t X (X^t X)^{-1} X^t Y}_{=I}) = E(X^t Y - X^t Y) = 0$$

En virtud de la segunda ecuación normal vale también que:

$$\sum_{i=1}^n (e_i X_i) = \sum_{i=1}^n (Y_i - a - bX_i) X_i = 0$$

Residuales: Propiedad ④

Los residuos son ortogonales a los valores ajustados:



$$E(e^t \hat{Y}) = 0$$

$$E(e^t \hat{Y}) = E[(Y - \hat{Y})^t \hat{Y}] = E[Y^t \hat{Y} - \hat{Y}^t \hat{Y}] = E[Y^t X \hat{\beta} - (X \hat{\beta})^t X \hat{\beta}]$$

$$E(e^t \hat{Y}) = E[Y^t X (X^t X)^{-1} X^t Y - ((X^t X)^{-1} X^t Y)^t (X^t X) (X^t X)^{-1} X^t Y]$$

$$E(e^t \hat{Y}) = E(Y^t H Y - Y^t H Y) = 0$$

Para el caso de la regresión lineal simple aplicando las propiedades 1 y 2 :

$$\sum_{i=1}^n e_i(a + bx_i) = \sum_{i=1}^n ae_i + \sum_{i=1}^n be_i X_i = 0 + 0 = 0$$

Varianza de los residuales



$$V(e) = (I - H)\sigma^2$$

$$V(e) = V(Y - \hat{Y}) = V(Y - X\hat{\beta}) = V(Y - X(X^tX)^{-1}X^tY) = V(Y(I - H))$$

$$V(e) = (I - H)V(Y)(I - H)^t = (I - H)^2\sigma^2 = (I - H)\sigma^2$$

Siendo la matriz $(I - H)$ es simétrica e idempotente.

Varianza de los residuales

La expresión de varianza de un residuo particular es:



$$V(e_i) = \sigma^2(1 - h_{ii})$$

siendo h_{ii} el i-ésimo elemento diagonal de la matriz Hat o $H = X(X^tX)^{-1}X^t$

Luego su valor estimado es:



$$\widehat{V(e_i)} = \hat{\sigma}^2(1 - h_{ii})$$

Covarianza entre Residuales

La expresión de la covarianza entre dos residuos es:




$$\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2$$

Notemos que de lo expuesto se desprenden las siguiente propiedades:

- a) Tanto los errores ε_i como los residuales e_i tienen media 0.
- b) La varianza de los errores es constante, pero la de los residuales no lo es.
- c) Los errores no están correlacionados, pero los residuales si.

Matriz Hat para la Regresión Lineal Simple

La expresión de los elementos diagonales de la matriz Hat para el caso de la Regresión Lineal Simple son:


$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Facultad de Ingeniería

Residuos estandarizados

Residuales Estandarizados

- 🧐 Los residuos estandarizados son los residuos corregidos con la estimación
- 🧐 Esta corrección se realiza para que tengan media cero y desviación estándar 1, de modo tal que los valores de la distribución normal resulten comparables para detectar residuos altos en valor absoluto de su desvío estándar.



$$r_{st} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

Validación de la normalidad

Objetivo del Análisis Diagnóstico

Realizar inferencia basada en el modelo ajustado.

Para verificar el cumplimiento del supuesto de normalidad se dispone de estrategias analíticas y gráficas.

Facultad de Ingeniería

Validación de la normalidad: Estrategias analíticas

Se dispone de varios test: Shapiro-Wilks, Lilliefors, Anderson-Darling entre otros, todos ellos disponibles en R.

Retomando el ejemplo de los frutos de damasco:

```
library(nortest) # REALIZA 10 PRUEBAS DE NORMALIDAD #  
residuos=residuals(mod_hojas)
```

```
shapiro.test(residuos)
```

```
ad.test(residuos)
```

```
lillie.test(residuos)
```

Validación de la normalidad: Estrategias analíticas

`shapiro.test(residuos)`

Shapiro-Wilk normality test

data: residuos

$W = 0.97355$, p-value = 0.8611

`ad.test(residuos)`

Anderson-Darling normality test

data: residuos

$A = 0.21563$, p-value = 0.8184

`lillie.test(residuos)`

Lilliefors (Kolmogorov-Smirnov) normality test

data: residuos

$D = 0.10504$, p-value = 0.8644

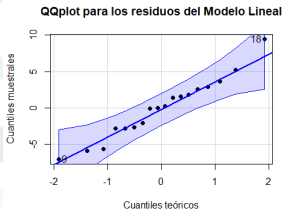
Conclusión

Todos los test coinciden en que puede sostenerse el supuesto de normalidad.

Validación de la Normalidad: Estrategias gráficas

Los gráficos de cuantil-cuantil o qq-plots visibilizan los apartamientos respecto de la normalidad de los residuos del modelo ajustado.

```
qqPlot(residuos, pch=19,  
main='QQplot para los residuos del Modelo Lineal',  
xlab='Cuantiles teóricos',  
ylab='Cuantiles muestrales')
```



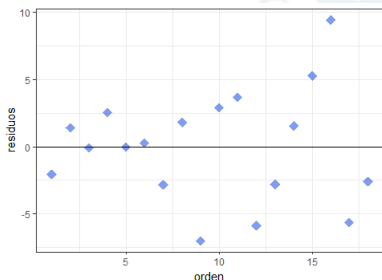
Validación de la independencia

La independencia de las observaciones puede ser especialmente importante en algunos casos, por ejemplo donde la variable temporal está presente. Para la inspección gráfica se representan los a fin de detectar (si existiera) la presencia de un patrón.

```
ggplot(dwdata,aes(x=orden,y=residuos))+  
geom_point(color = '#013ADF' ,  
fill = ' #013ADF' , size = 4, shape = 18, alpha = 0.5)+xlab(' orden'  
) +  
geom_abline(slope = 0)
```

Facultad de Ingeniería

Validación de la Independencia



Para realizar esta verificación analíticamente, se puede aplicar el test de Durbin-Watson.

```
library(lmtest)
```

```
dwtest(mod_hojas,alternative  
="two.sided",iterations = 1000)
```

Durbin-Watson test

data: mod_hojas

DW = 2.1074, p-value = 0.9

alternative hypothesis: true

autocorrelation is not 0

En este gráfico se observa que no hay estructura visible en los datos y se corrobora esta visualización con el resultado del test.

Validación de la homocedasticidad

Supuesto de Homocedasticidad

El modelo lineal OLS planteado supone que la varianza de los errores es constante.



$$V(\varepsilon_i) = \sigma^2$$

El correlato muestral es una varianza constante de los residuos para los distintos valores de la variable predictora o regresora.

Entre las consecuencias de la falta de homocedasticidad podemos mencionar la pérdida de eficiencia del estimador del modelo y el error de cálculo del estimador de la matriz de covarianzas de los estimadores.

Validación de la homocedasticidad

Existen varias pruebas como el test de Breusch-Pagan o el de Goldfeld-Quandt disponibles en R.

```
bptest(mod_hojas)
gqtest(mod_hojas, order.by = LONGF, data=hojas)
```

studentized Breusch-Pagan test

data: mod_hojas

BP = 3.7166, df = 1, p-value = 0.05387

Goldfeld-Quandt test

data: mod_hojas

GQ = 3.1849, df1 = 7, df2 = 7, p-value = 0.07468

alternative hypothesis: variance increases from segment 1 to 2

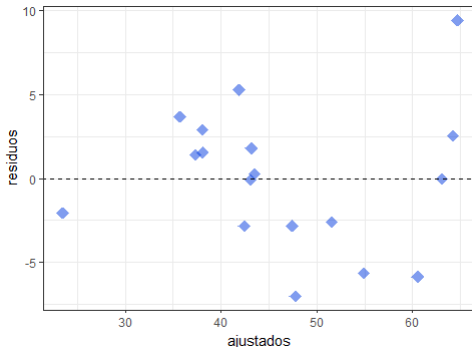
Validación Gráfica de la homocedasticidad

Se utiliza el gráfico de residuales versus la variable predictora que permite observar si el error del modelo aumenta o disminuye al aumentar la magnitud de los valores de la variable predictora. Una estructura de 'embudo' alerta respecto de heterocedasticidad.

Veamos el caso del modelo de los frutos de damasco:

```
ajustados=mod_hojas$fitted.values
databp=data.frame(ajustados,residuos)
ggplot(databp,aes(x=ajustados,y=residuos))+
  geom_point(color = "#013ADF", fill = '#013ADF', size = 4,
  shape = 18, alpha = 0.5)+xlab(' ajustados' )+
  geom_abline(slope = 0,linetype=' dashed' )
```

Validación Gráfica de la homocedasticidad



Si bien el test no rechaza la homocedasticidad de los errores el p valor es cercano al nivel de significación y se aprecia una leve estructura en la gráfica.

Detección de Outliers

En ocasiones una observación o un subgrupo de observaciones puede presentarse muy alejado del conjunto mayoritario de observaciones o no seguir exactamente el patrón de relación del conjunto general. Esta observación o conjunto de observaciones pueden ser outliers, pero también pueden ser valores influyentes.


Decimos que un dato alejado es un outlier cuando está alejado del conjunto general pero sigue el mismo patrón de distribución. La detección de estos valores es vital dada la sensibilidad del modelo de cuadrados mínimos ordinarios OLS a la presencia de observaciones atípicas o alejadas. Dicho de otra manera estas observaciones pueden afectar sensiblemente a la estimación de los coeficientes del modelo de OLS.

Detección de Ouliers

Si sospechamos que una observación es un outlier podemos aplicar un método clásico dentro de la regresión que es estimar el modelo con sus coeficientes pero excluyendo en la estimación a esta observación en particular.

Si se trata de la i -sima observación, podemos estimar su variable respuesta con el modelo original \hat{Y}_i y también con el modelo que la excluye $\hat{Y}_i(i)$.

Se propone el siguiente estadístico de contraste:


$$t_{obs} = \frac{Y_i - \hat{Y}_{i(i)}}{\sqrt{\hat{\sigma}^2(Y_i - \hat{Y}_{i(i)})}} = \frac{Y_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

Detección de Outliers

Si se puede sostener normalidad de la distribución de los errores, el estadístico propuesto tiene distribución t-Student con $n - (p + 1)$ grados de libertad, siendo p la cantidad de coeficientes estimados en el modelo.

Para controlar el nivel de significación global del contraste se puede aplicar la corrección del nivel por Bonferroni, estableciendo como hipótesis de nulidad: Ninguna de las observaciones es un outlier versus alguna lo es.

Veamos la aplicación en R para el ejemplo de los frutos de damasco:

```
outlierTest(mod_hojas)
```

No Studentized residuals with Bonferroni $p < 0.05$

Largest|rstudent| :

rstudent unadjusted p-value Bonferroni p

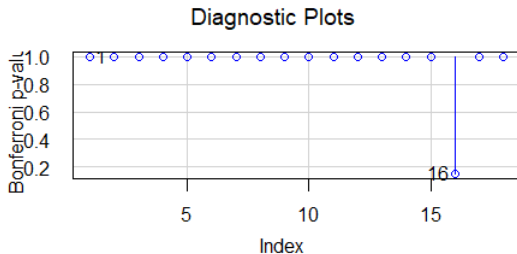
16	3.036495	0.0083297	0.14994
----	----------	-----------	---------

La observación 16 quedó señalada como valor extremo u outlier.

Es posible dibujar los resultados de la prueba de Bonferroni para cada observación. En la siguiente figura se observa que sólo la observación 16 es identificada como un posible outlier ya que su valor-p es muy pequeño.

```
library(car)
```

```
influenceIndexPlot(mod_hojas, vars="Bonf", las=1,col="blue")
```



Los residuos estandarizados nos indican qué tan extremos son los valores



Residuos que exceden más de 3 desviaciones estándar indican caso

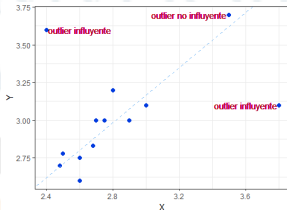
Detección de Valores influyentes

Valor Influyente

Una observación se considera un valor influyente cuando su presencia afecta sensiblemente al comportamiento del modelo.

En ML simple remover un valor influyente podría cambiar drásticamente el valor de la pendiente estimada.

Los outliers no necesariamente son valores influyentes.



Qué aportan?

Valor Influyente

Estas observaciones nos dicen cosas importantes respecto del modelo y de nuestro conjunto de observaciones. Si el fin del ajuste es predictivo, un modelo sin estas observaciones puede ser más útil para predecir con precisión la mayoría de casos. Sin embargo, prestar atención a estos valores es fundamental puesto que si no corresponden a errores de medición o carga, pueden corresponder a los casos más interesantes. Una alternativa adecuada cuando se sospecha de algún posible valor atípico o influyente es calcular el modelo de regresión incluyendo y excluyendo dicho valor.

Leverage (Apalancamiento)

El 'leverage', cuantifica el grado de apalancamiento de una observación, midiendo la contribución de la i -ésima observación a la suma total de cuadrados de la variable dependiente (Y). Para un modelo de regresión lineal simple su expresión es:



$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Observaciones con leverage superior a $h_{ii} > 2p/n$ se consideran influyentes. Algunos autores dicen $3p/n$

Propiedades del Leverage (Apalancamiento)

- El leverage h_{ii} mide la distancia entre el valor de x para la i -ésima observación y la media de los valores de x para las n observaciones.
- El leverage es un número entre 0 y 1; es decir: $0 < h_{ii} \leq 1$. Esto se debe a que la matriz H es simétrica e idempotente.
- La suma de leverages sobre todas las observaciones es $\sum_{i=1}^n h_{ii} = p$ siendo p el número de coeficientes del modelo.
- El puntaje de apalancamiento también se conoce como la auto-sensibilidad o auto-influencia de la observación, dado que:



$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$$

lo cual se deduce de la expresión: $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

Leverage (Apalancamiento)

El leverage de la i -ésima observación es igual a la derivada parcial del i -ésimo valor dependiente ajustado \hat{y}_i con respecto al i -ésimo valor dependiente observado y_i . Esta derivada parcial describe el grado en que el i -ésimo valor observado influye en el i -ésimo valor ajustado.

Veamoslo en el ejemplo de los frutos de damasco:

```
library(stats)
cota=3 * mean(hatvalues(mod_hojas))
leverage <- hatvalues(mod_hojas) > cota
cbind(hatvalues(mod_hojas),cota,leverage)
```

Ningún valor es señalado como outlier en este caso!

Distancias de Cook

Como se relacionan valore influyentes y outliers'

Cuando un valor es influyente, generalmente será atípico o tendrá un alto leverage; sin embargo no vale la recíproca; es decir que hay valores atípicos que no son influyentes o con alto leverage que no lo son tampoco.

La pregunta es entonces cuándo un valor es influyente y cómo evaluarlo analítica y gráficamente.



La distancia de Cook es una medida muy utilizada que combina único valor, la magnitud del residuo y el grado de leverage.

Distancias de Cook

La Distancia de Cook es una medida de cómo influye la observación i -ésima en la estimación del coeficiente β si es excluida del conjunto de datos para la estimación.



$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}^2}$$

Una distancia de Cook grande significa que una observación tiene una influencia importante en la estimación de β .



No existe un punto de corte aceptado universalmente.

Distancia de Cook (punto de corte)

Existen diferentes propuestas para el punto de corte:



Algunos autores sugieren investigar cualquier punto sobre $4/n$, siendo n el total de observaciones.



Otros autores proponen investigar cualquier distancia mayor a 1, o bien superior a 0.5.

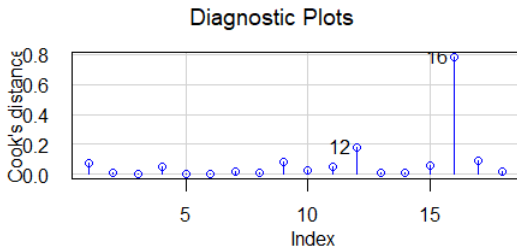


Otros indican que cualquier valor que sobresalga notablemente de los restantes valores de distancias de Cook debería inspeccionarse.

Distancias de Cook: aplicación

Veamos las distancias de Cook en el caso de los frutos de damasco:

```
library(car)  
influenceIndexPlot(mod_hojas, vars='Cook', las=1,col='blue')
```




Se destacan dos observaciones con valores altos de distancias de Cook,

DFFITS

Cuantifican la influencia

de la i -ésima observación sobre su propio valor predicho o ajustado, la expresión para calcularla es la siguiente:


$$DFFITS_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

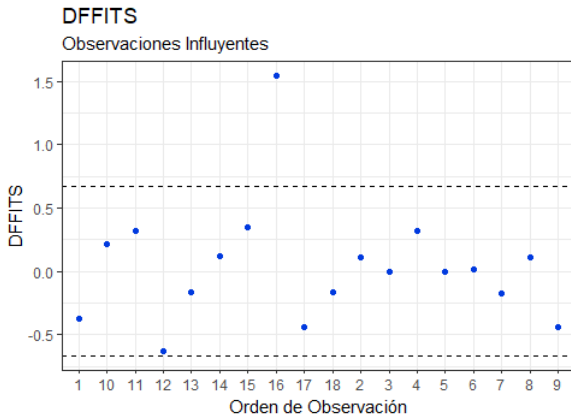
Las observaciones con DFFITS superiores a $2\sqrt{p/n}$ en valor absoluto se consideran influyentes.

DFFITS

Veamos los DFFITS para el caso de los frutos de damasco:

```
df <- mod_hojas$df.residual
p <- length(mod_hojas$coefficients)
n <- nrow(mod_hojas$mod_hojas)
dffits_crit = 2 * sqrt(p / n)
dffits <- dffits(mod_hojas)
df <- data.frame(obs = names(dffits), dffits = dffits)
ggplot(df, aes(y = dffits, x = obs)) + geom_point(color='
#013ADF' ) +
geom_hline(yintercept = c(dffits_crit, -dffits_crit), linetype=' dashed'
) + labs(title = "DFFITS", subtitle = ' Observaciones Influyentes' ,
x = ' Orden de Observación' , y = ' DFFITS' )+theme_bw()
```

DFFITS



Nuevamente es la observación 16 la que se destaca con un valor alto de DFFITS.

DFBeta

Objetivo: Evaluar el cambio en los coeficientes de regresión tras excluir la observación

: se trata de un proceso iterativo en el que cada vez se excluye una observación distinta y se reajusta el modelo. En cada iteración se registra la diferencia en los coeficientes de regresión con y sin la observación, dividida entre el error estándar del predictor en el modelo sin la observación.



$$DFbeta_i = \frac{\widehat{\beta} - \widehat{\beta}_{(i)}}{ES_{\widehat{\beta}_{(i)}}}$$

DFBETA

Interpretación

El valor de DFbeta indica el impacto de una observación sobre la estimación del coeficiente del modelo medido sobre dicho coeficiente, mientras que la distancia de Cook lo mide sobre los valores estimados. El valor de corte para el DFBeta es 1.

Veamos cómo buscar en R observaciones con DFbeta alto, usamos nuevamente el ejemplo de los frutos de damasco:

```
dfbetas(mod_hojas)[,2] > 1 # esta función entrega los valores estandarizados
```

En este caso tampoco se aprecian observaciones influyentes.

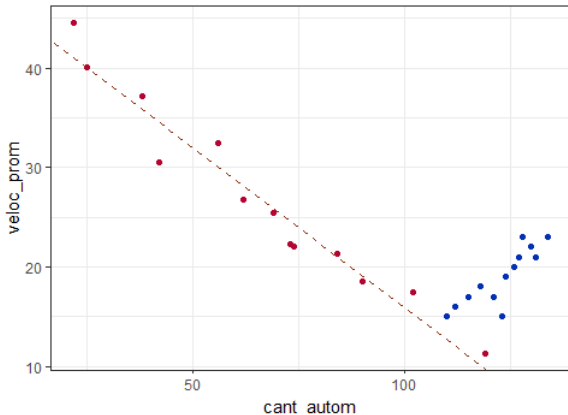
Alcances del Modelo

Después de estimar los coeficientes y testear su validez si:

- 😊 El modelo resulta estadísticamente significativo.
- 😊 No se rechaza la independencia de las observaciones.
- 😊 No se rechaza la normalidad de los residuos.
- 😊 No se aprecia presencia de estructura en los residuos.
- 😊 No se rechaza la homocedasticidad de los residuos.

Entonces el modelo puede utilizarse para hacer estimaciones puntuales o por intervalos teniendo en consideración la limitación el dominio sobre el que fue construido.

Es decir, que si la variable predictora $X \in (0; 100)$ las inferencias basadas en este modelo deben circunscribirse a este intervalo dado que fuera del mismo la función podría dejar de ser lineal. En la Figura se aprecia que el modelo lineal puede ser adecuado en ciertos intervalos pero no globalmente.



Cuando no se cumplen los supuestos

En algunas ocasiones no se satisface alguno o más de uno de los supuestos, esto puede deberse por ejemplo a la presencia de curvatura, a la falta de normalidad o bien a la falta de homogeneidad de los residuos.

Según corresponda tenemos alternativas para solucionar el problema y construir el modelo más adecuado.

Facultad de Ingeniería

Los residuos presentan estructura

Qué hacer?

Cuando los residuos muestran estructura es posible que la función elegida (hasta ahora la lineal) no sea la más adecuada. Si se desea mantener el entorno del modelo lineal simple puede cambiarse la función del modelo y en caso de admitir el paso a regresión múltiple podría incorporarse una nueva variable al modelo de modo tal de explicar la variabilidad no captada por el modelo inicial.

Debe destacarse que si se realiza tal cambio, deberá a continuación validarse la bondad de ajuste del modelo y el análisis diagnóstico para este nuevo modelo.

Ejemplo 5: Pulpa de papel

Se pretende analizar la relación existente entre la concentración de madera de la pulpa de papel, y la resistencia del papel elaborado con ésta. El objetivo del análisis es describir y cuantificar la tendencia observada. Se utiliza la base de datos **madera.xlsx** disponible en shorturl.at/BHU01

```
mod_mad=lm(resist ~madera,data=madera)
summary(mod_mad)
coef=mod_mad$coefficients
ggplot(madera,aes(madera,resist)) + geom_point(color=' #8A2908'
) + labs(x = ' Concentración de madera' , y = ' Resistencia del
papel' ) +
theme_bw() +
geom_abline(intercept=coef[1],slope=coef[2],linetype=' dashed' ,
color=' #2E2E2E' )
```

El Modelo

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.0156	3.7312	5.90	0.0000
madera	1.4144	0.4201	3.37	0.0034

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

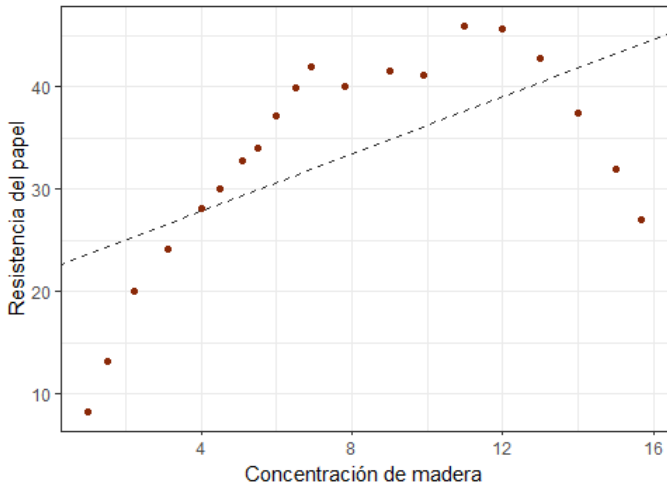
Residual standard error: 8.56 on 18 degrees of freedom

Multiple R-squared: 0.3712, Adjusted R-squared: 0.3363

F-statistic: 10.63 on 1 and 18 DF, p-value: 0.004351

Si bien el modelo es adecuado se observa que el valor del coeficiente de determinación no parece importante.

La salida



Análisis de normalidad de los residuos

Analizamos normalidad

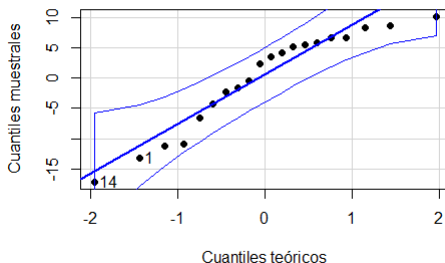
```
shapiro.test(residuals(mod_mad))  
qqPlot(residuals(mod_mad), pch=19,  
main='QQplot para los residuos del Modelo Lineal',  
xlab='Cuantiles teóricos',  
ylab='Cuantiles muestrales')
```

Shapiro-Wilk normality test

```
data: residuals(mod_mad)  
W = 0.91069, p-value = 0.0657
```

Q-q plot de los residuos

QQplot para los residuos del Modelo Lineal



No se aprecian apartamientos del supuesto de normalidad, ni en el test de Shapiro, ni en el qqplot, sin embargo se señalan un par de puntos problemáticos: el primero y el catorceavo.

Independencia de los Residuos

Analizamos independencia

```
dwtest(mod_mad, alternative = 'two.sided', iterations = 1000)
```

Durbin-Watson test

data: mod_mad

DW = 1.3873, p-value = 0.1114

alternative hypothesis: true autocorrelation is not 0

No se evidencian violaciones del supuesto de independencia.

Buscamos Outliers

```
# Detectamos presencia de outliers
```

```
outlierTest(mod_mad)
```

```
influenceIndexPlot(mod_mad, vars="Bonf", las=1,col='blue')
```

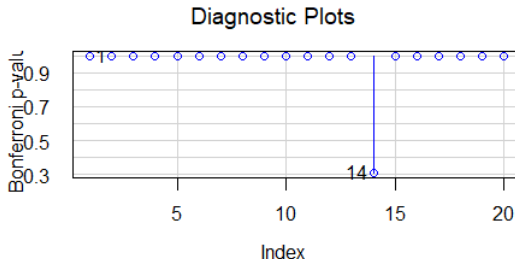
No Studentized residuals with Bonferroni $p < 0.05$ Largest —rstu—
dent—:

rstudent unadjusted p-value Bonferroni p

14	-2.692756	0.015406	0.30812
----	-----------	----------	---------

Facultad de Ingeniería

Bonferroni



Nuevamente se destacan las mismas observaciones que en el gráfico cuantil-cuantil de los residuos.

Buscamos puntos influyentes

#Estudiamos el leverage

```
cota=3 * mean(hatvalues(mod_mad)) leverage < - hatval-  
ues(mod_mad) > cota sum(leverage)
```

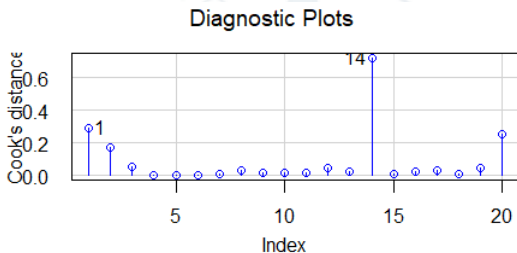
0

Primero definimos la cota para este modelo, luego sumamos la cantidad de observaciones que superan la cota y en este caso ninguna lo hace. Es decir que no hay valores influyentes.

Buscamos puntos influyentes

```
# Buscamos influencia por distancias de Cook
```

```
influenceIndexPlot(mod_mad, vars='Cook', las=1,col='blue')
```



La distancia de Cook correspondiente a las observaciones 1 y 14 quedan señaladas.

Buscamos puntos influyentes

#Estudiamos los DFFITS

```
dffits_crit = 2 * sqrt(p / n)
```

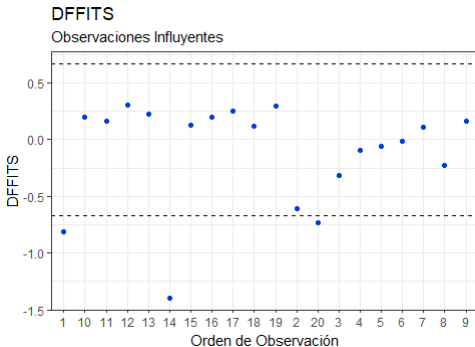
```
dffits <- dffits(mod_mad)
```

```
df <- data.frame(obs = names(dffits), dffits = dffits)
```

```
ggplot(df, aes(y = dffits, x = obs)) + geom_point(color='  
#013ADF' ) +
```

```
geom_hline(yintercept = c(dffits_crit, -dffits_crit), linetype='dashed')  
+ labs(title = 'DFFITS', subtitle = 'Observaciones Influyentes',  
x = 'Orden de Observación', y = 'DFFITS')+theme_bw()
```

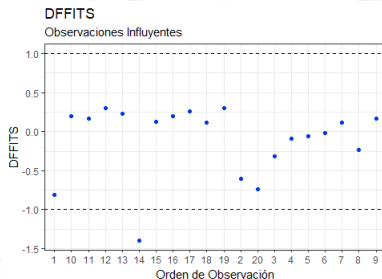
Buscamos observaciones influyentes



Las observaciones 1, 14 y 20 superan la cota establecida para los DFFITS.

Con la cota $3\sqrt{\frac{p}{n}}$, sólo la observación 14 caería fuera.

Buscamos observaciones influyentes



Estudiamos los DFBetas

```
sum(dfbetas(mod_mad)[, 2] > 1)
```

```
0
```

No hay observaciones con DFBeta superior a la cota.

Homocedasticidad de los Residuos

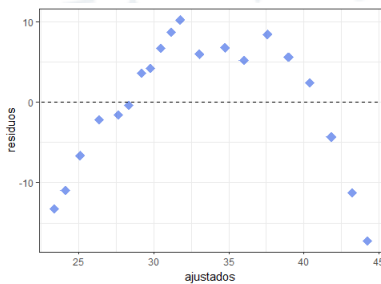
```
bptest(mod_mad) # test de Breusch Pagan
ajustados=mod_mad$fitted.values; residuos=residuals(mod_mad)
databp=data.frame(ajustados,residuos)
ggplot(databp,aes(x=ajustados,y=residuos))+
  geom_point(color = '#013ADF', fill = '#013ADF', size = 4,
  shape = 18, alpha = 0.5)+xlab('ajustados')+
  geom_abline(slope = 0,linetype='dashed')+theme_bw()
```

studentized Breusch-Pagan test

data: mod_mad BP = 1.037, df = 1, p-value = 0.3085

No se rechaza la hipótesis de homocedasticidad de los residuos, pero se aprecia estructura en los mismos.

Se aprecia estructura en los residuos!



¿Cómo podemos resolver este problema?

Vamos a incorporar una nueva variable al modelo sugerida por la estructura de los residuos. Nos estamos yendo a un modelo de regresión lineal múltiple dado que estamos proponiendo dos variables predictoras o explicativas.

```
mod_mad2=lm(resist ~ madera+l(madera**2),data=madera)
summary(mod_mad2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.9119	1.5744	0.58	0.5701
madera	8.5430	0.4425	19.31	0.0000
(<i>madera</i> ²)	-0.4270	0.0258	-16.58	0.0000

Residual standard error: 2.077 on 17 degrees of freedom

Multiple R-squared: 0.9643, Adjusted R-squared: 0.9601

F-statistic: 229.3 on 2 and 17 DF, p-value: 5.035e-13

Interpretación de la Salida

Cómo seguimos?

- 😊 Ambos coeficientes resultan significativos.
- 😊 La regresión es significativa.
- 😊 El poder explicativo del modelo aumentó considerablemente.

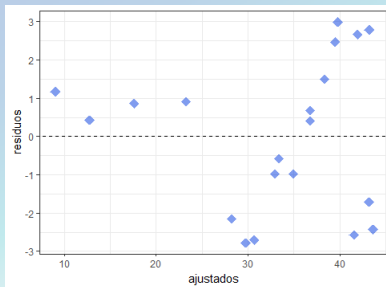


Sin embargo...

- 😬 Resta realizar el análisis diagnóstico del nuevo modelo.

Consideraciones Finales

Todos los análisis realizados se repiten y no aparecen outliers ni observaciones muy influyentes, luego volvemos a graficar los residuos versus los ajustados y visualizamos que ya no hay estructura.




Transformaciones de Box & Cox

Box y Cox desarrollaron una familia de transformaciones con el propósito de lograr normalidad en la distribución de los errores. Estas transformaciones a menudo también reduce la no linealidad, y la heteroscedasticidad.

El procedimiento básico de Box-Cox trata de encontrar el mejor exponente para transformar los datos en una forma normal. Posteriormente, se aplica la transformación a todos los datos del conjunto.

- 😊 La transformación de Box-Cox está definida solamente para valores de respuesta positivos. Si la respuesta contiene ceros o valores negativos se agrega automáticamente un desplazamiento.
- 😊 Cuando las transformaciones de Box & Cox no logran normalizar los residuos, es conveniente considerar una regresión robusta o una regresión beta u otra metodología.

La forma de estas transformaciones es:


$$L(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y) & \text{si } \lambda = 0 \end{cases}$$

siendo y la variable a ser transformada y λ el parámetro de transformación. Si bien $\lambda \in \mathbb{R}$, se suele seleccionar valores de λ múltiplos de 0.25 o 0.1.

La función boxcox (MASS), devuelve una gráfica que analiza diferentes valores de potenciales λ para la transformación.

Si se realiza una transformación de variables (respuesta o predictora/s), también se modifica la interpretación de todo. Generalmente podemos aplicar transformaciones inversas para dar una interpretación en términos de nuestras variables originales.

Ejemplo 7: cars

Utilizamos la base de datos cars de la biblioteca datasets, Intentaremos ajustar un modelo para estimar la distancia en función de la velocidad.



Facultad de Ingeniería

Ejemplo 7: cars

```
data(' cars' , package = ' datasets' )
```

```
mod_cars <- lm(dist ~ speed, data = cars)
```

```
summary(mod_cars) # Visualizamos el ajuste del modelo propuesto
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.5791	6.7584	-2.60	0.0123
speed	3.9324	0.4155	9.46	0.0000

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Ejemplo 7: cars

El modelo ajusta razonablemente bien, sin embargo falta analizar los supuestos del modelo. Veamos la normalidad de los residuos.

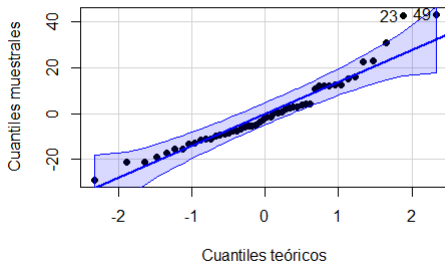
```
# Analizamos la normalidad de los residuos  
shapiro.test(mod_cars$residuals)
```

Shapiro-Wilk normality test

```
data: mod_cars$residuals  
W = 0.94509, p-value = 0.02152
```

Gráfico cuantil-cuantil modelo resistencia de papel

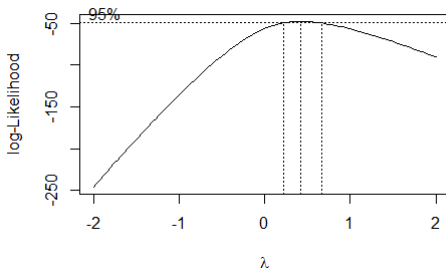
QQplot para los residuos del Modelo Lineal



Vemos que los residuos no satisfacen el supuesto de normalidad.
distribuciona

Transformación de Box & Cox

Buscamos la mejor transf de Box & Cox para normalizar los residuos
`boxcox(object = mod_cars, plotit = TRUE)`



El gráfico señala que la mejor opción de λ es cercana a 0.5. También se puede pedir el valor exacto de la grilla que maximiza la logverosimilitud.

```
mod_cars2=lm( $dist^{0.5} \sim speed$ , data = cars)
summary(mod_cars2) # Visualizamos el ajuste del nuevo modelo
propuesto
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.2771	0.4844	2.64	0.0113
speed	0.3224	0.0298	10.83	0.0000

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

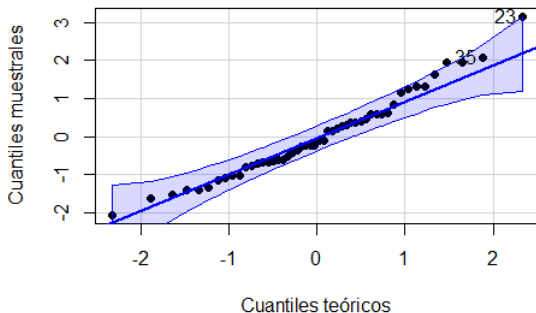
Residual standard error: 1.102 on 48 degrees of freedom

Multiple R-squared: 0.7094, Adjusted R-squared: 0.7034

F-statistic: 117.2 on 1 and 48 DF, p-value: 1.773e-14

Ejemplo 7: cars

QQplot para los residuos del Modelo Lineal



Ahora se satisface el supuesto de normalidad de los residuos, sin embargo debemos ser cuidadosos en la interpretación del modelo.

Organización

- 1 Análisis Diagnóstico
- 2 Cuadrados Mínimos Ponderados
- 3 Modelos Robustos

UNIVERSIDAD AUSTRAL

Facultad de Ingeniería

WLS como generalización de OLS

El modelo de mínimos cuadrados ponderados o pesados (WLS) es una regresión en la cual las observaciones tienen diferentes pesos o importancia en el modelo; es decir que cada observación realiza un aporte diferente a la estimación de los coeficientes del modelo.

Existe heterocedasticidad cuando los residuos están relacionados con los predictores. Por ejemplo, la heteroscedasticidad podría estar presente en un modelo que pretenda explicar los ingresos en función de la edad, dado que el ingreso es generalmente más variable para los adultos que para los jóvenes.

Puede ocurrir también cuando la variable respuesta es el promedio de varias repeticiones del experimento y la varianza resulta proporcional a la cantidad de repeticiones.

Por qué es un problema la heterocedasticidad?

Complica la inferencia (pruebas de hipótesis e IC) pues éstos asumen errores no correlacionados y homocedásticos. Luego las predicciones basadas en estos modelos podrían ser incorrectas en presencia de heterocedasticidad. Si la cantidad de observaciones, sin embargo es suficientemente grande, la varianza del estimador OLS aún puede ser pequeña y las estimaciones precisas.

Los mínimos cuadrados ponderados o Weighted Least Squares (WLS) corrigen la varianza no constante al ponderar cada observación por el recíproco de su varianza estimada. Las observaciones con pequeñas varianzas estimadas se ponderan más que las observaciones con grandes varianzas estimadas. El método OLS minimiza la suma de cuadrados residuales para hallar los estimadores de los coeficientes, mientras que el método WLS minimiza la suma de cuadrados ponderada.

OLS es WLS con pesos iguales a 1

OLS

Se minimiza la suma de cuadrados residuales:

$$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Cuya expresión matricial es:

$$SCR = (Y - X\beta)^t (Y - X\beta)$$

Derivando respecto de β y despejando el vector de coeficientes estimado tiene la siguiente expresión:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

OLS versus WLS

WLS

Se minimiza la suma de cuadrados de los residuos ponderados:

$$\sum_{i=1}^n w_{ii}(y_i - \hat{y}_i)^2$$

Cuya expresión matricial es:

$$SCR_W = (Y - X\beta)^t W (Y - X\beta)$$

Derivando respecto de β y despejando el vector de coeficientes estimado tiene la siguiente expresión:

$$\hat{\beta} = (X^t W X)^{-1} X^t W Y$$

WLS: ventajas y desventajas

Las ventajas son:

- 😊 Extraer la máxima información de pequeños conjuntos de datos.
- 😊 Es la mejor alternativa para residuos heterocedásticos.

Las desventajas incluyen:

- 😞 Requiere el conocimiento de los pesos o su estimación cuidadosa.
- 😞 Si los pesos están mal establecidos y las muestras son pequeñas los resultados pueden ser impredecibles.
- 😞 Es muy sensible a outliers. Un valor atípico mal ponderado puede sesgar drásticamente las estimaciones.

La dificultad de la implementación de WLS es la estimación de los pesos. Sin embargo es muy probable que se deban estimar a partir de los gráficos de los

Ejemplo 8: choferes

Se desea estimar el número de inspectores adecuado para una línea de transporte urbano basado en el número de choferes de la misma. Para ello se dispone de datos de 29 líneas de transporte de esa ciudad y su correspondiente número de choferes. Los datos están en el archivo **inspectores.xlsx** disponible en shorturl.at/afkN7.




```
# Ajustamos un modelo de cuadrados mínimos ordinarios
ols_inspec=lm(inspectores choferes,data=inspectores)
summary(ols_inspec)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.1832	27.9891	0.54	0.5919
choferes	0.2582	0.0375	6.89	0.0000

Residual standard error: 64.47 on 27 degrees of freedom
 Multiple R-squared: 0.6373, Adjusted R-squared: 0.6239
 F-statistic: 47.45 on 1 and 27 DF, p-value: 2.116e-07

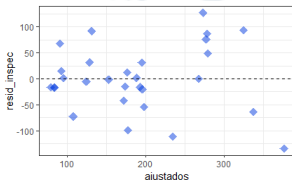
El modelo planteado resulta significativo, falta verificar el cumplimiento de los supuestos del modelo.

```
resid_inspec=ols_inspec$residuals # almacenamos los residuos  
ajust_insp=ols_inspec$fitted.values # idem valores ajustados  
shapiro.test(resid_inspec) # testeamos la normalidad de los residuos  
bptest(ols_inspec) # verificamos el supuesto de homocedasticidad
```

```
shapiro.test(resid_inspec)  
Shapiro-Wilk normality test  
data: resid_inspec  
W = 0.97498, p-value = 0.7002  
bptest(ols_inspec)  
studentized Breusch-Pagan test  
data: ols_inspec  
BP = 9.4436, df = 1, p-value = 0.002119
```

Se verifica normalidad, pero no se verifica homocedasticidad. Inspeccionamos la gráfica los residuos vs valores ajustados.

```
ggplot(dat_insp,aes(x=ajust_insp,y=resid_inspec))+  
geom_point(color = ' #013ADF' , fill = '#013ADF' , size = 4,  
shape = 18, alpha = 0.5)+xlab(' ajustados' )+  
geom_abline(slope = 0,linetype="dashed")+theme_bw()
```



Se aprecia en esta figura que la variabilidad de los residuos aumenta conforme aumentan los valores ajustados. Vamos a intentar con un modelo de cuadrados mínimos ponderados.

```
fitted=lm(abs(ols_inspec$residuals) ~ ols_inspec$fitted.values)$residuals
pesos=l(1/ols_inspec$fitted.values^2) # definimos el modelo con los pesos
wls_inspec<- lm(inspectores_choferes, weights = pesos, data=inspectores)
summary(wls_inspec)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	13.6846	18.9751	0.72	0.4770
choferes	0.2603	0.0387	6.73	0.0000

Residual standard error: 0.339 on 27 degrees of freedom Multiple R-squared: 0.6262, Adjusted R-squared: 0.6124 F-statistic: 45.24 on 1 and 27 DF, p-value: 3.201e-07

Este modelo también es adecuado, el valor del R_{adj}^2 es similar pero el valor residual en el primero es 64.47, mientras que en el segundo es 0.339, lo cual dice

Organización

- 1 Análisis Diagnóstico
- 2 Cuadrados Mínimos Ponderados
- 3 Modelos Robustos
 - Precisión de los métodos ajustados

Facultad de Ingeniería

Cuándo?

Cuando los errores no tienen distribución normal y las transformaciones de Box & Cox no corrigen esto, los resultados obtenidos por OLS pueden verse afectados.

Una alternativa podría ser eliminar los valores atípicos (outliers) que forman dichas colas, sin embargo, si no son errores de carga el no incluirlos puede afectar el estudio del fenómeno que trata de modelarse.




Otra alternativa más apropiada es reducir la influencia de estos valores atípicos mediante una regresión robusta. Los modelos robustos son técnicas de regresión lineal capaces de integrar datos atípicos y puntos de apalancamiento.

Estos modelos son también una regresión con ponderaciones que penalizan o subestiman los valores influyentes y/o atípicos. Por lo tanto, es un caso especial de regresión ponderada. Si no hubiera entre los datos puntos atípicos, el estimador se reduce al OLS.

Alternativas

Algunos estimadores disponibles


Existen varias técnicas de regresión robusta, podemos mencionar:

-  M- estimador de Huber
-  Mínimos cuadrados podados (LTS)
-  Desviaciones mínimas absolutas (LAD)

M - estimador de Huber

Un M-estimador se obtiene minimizando la suma de una función de los residuos; en este sentido el OLS utilizaría la función cuadrática.

Huber propuso una generalización para estimar los parámetros mediante la minimización de una función de la forma:


$$H(\theta) = \sum_{i=1}^n \rho \left(\frac{e_i(\theta)}{s} \right)$$

siendo s un estimador de escala de los errores y ρ una función que se elige de modo tal que tenga ciertas propiedades deseables, como eficiencia e insesgamiento, bajo el supuesto distribucional adecuado.

La función ρ

Esta función debe satisfacer las siguientes propiedades:



ρ es par; es decir $\rho(v) = \rho(-v) \quad \forall v \in \mathbb{R}$.



$\rho(0) = 0$.



ρ es monótona no decreciente, es decir: $0 \leq v \leq u \Rightarrow \rho(v) \leq \rho(u)$.



ρ es una función derivable y su derivada es denotada con ψ .



ρ es una función acotada.

Las soluciones que minimizan estas funciones se denominan M-estimadores.



$$\hat{\theta} = \arg \min_{\theta} \sum_{i=1}^n \rho \left(\frac{e_i(\theta)}{s} \right)$$

Por qué cambiar de norma?

Para disminuir el impacto de outliers/leverage en la estimación

La norma cuadrática se utiliza en la estimación de parámetros con mucha frecuencia y mientras los datos no contengan contaminaciones resulta apropiada. Sin embargo frente a datos atípicos e influyentes la estimación es ineficiente.

Los M-estimadores fueron propuestos para minimizar el efecto de los outliers, incluyéndolos sin embargo dentro de la estimación.

Para hallar los M - estimadores se diferencia la función a minimizar y se iguala a cero, obteniéndose la siguiente expresión:



$$\sum_{i=1}^n \psi \left(\frac{e_i(\theta)}{s} \right) x_i = 0$$

Cómo lograr mayor robustez?

Para alcanzar un mayor grado de robustez que el estimador OLS, es necesario que la función ρ tenga un crecimiento más lento que la cuadrática, de modo tal que las observaciones atípicas tengan menor peso.

La familia de funciones ρ propuesta por Huber es de la forma:

$$\rho_H(v) = \begin{cases} -cv - \frac{c^2}{2} & \text{si } v < -c \\ \frac{v^2}{2} & \text{si } |v| \leq c \\ cv + \frac{c^2}{2} & \text{si } v > c \end{cases}$$

Su derivada tiene la expresión:

$$\psi_H(v) = \rho'_H(v) = \text{signo}(v) \min \{c, |v|\}$$

Mínimos Cuadrados Recortados(Rousseeuw)

El estimador LTS es el que minimiza la suma de los cuadrados de los k residuos con valores absolutos más pequeños, donde el valor de k satisface $\frac{1}{2} \leq k \leq 1$. Es decir que en este caso la función que se minimiza es:




$$Q_k(\theta) = \sum_{i=1}^k e_i^2(\theta)$$

El estimador de mínimos cuadrados recortados es la solución a la minimización de la función $Q_k(\theta)$


Esta definición es muy similar al OLS, la diferencia está en que OLS realiza el mínimo sobre todas las observaciones y este estimador sólo considera k/n observaciones, o bien les asigna peso nulo a las restantes. Esto genera un estimador menos influenciado por los outliers.

Desviaciones mínimas absolutas

En este caso la función a minimizar es la suma de los valores absolutos de los residuos.


$$S(\theta) = \sum_{i=1}^k |e_i(\theta)|$$

Luego el estimador es la solución que minimiza esta expresión, simbólicamente:

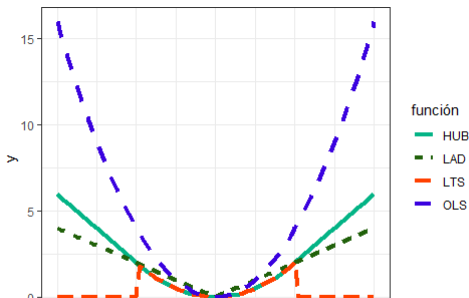

$$\hat{\theta} = \arg \min_{\theta} S(\theta)$$

No existe una solución analítica para la regresión de mínimos desvíos absolutos, por lo cual está en desventaja frente a OLS.

Una alternativa viable es aplicar métodos iterativos.

También es importante destacar que, para un determinado conjunto de datos, este método puede producir múltiples soluciones, mientras que el método OLS siempre produce una única solución.

En la figura se comparan las funciones que minimiza cada modelo.







Ejemplo8: Modelos Robustos en R

La base de datos Duncan disponible en la biblioteca 'carData' tiene 45 observaciones y 4 variables. Estos datos corresponden a ocupaciones de EEUU con sus correspondientes valoraciones de prestigio en el año 1950.



Ejemplo 8: Las variables

-  **type:** factor con tipo de ocupación. Niveles: docente, profesional, gerencial; de cuello blanco y cuello azul.
-  **income:** porcentaje de personas ocupadas con ese cargo en el censo de 1950 que ganaban más de 3500 dólares por año.
-  **education:** porcentaje de personas ocupadas con ese cargo que tenían terminados sus estudios secundarios.
-  **prestige:** porcentajes de opiniones en una encuesta que valoraron el cargo como prestigioso.

Facultad de Ingeniería

Inspeccionamos la asociación entre las variables



Se aprecia una relación lineal razonable entre el ingreso, la educación y el prestigio. Vamos a tratar de explicar el ingreso a partir de la educación.

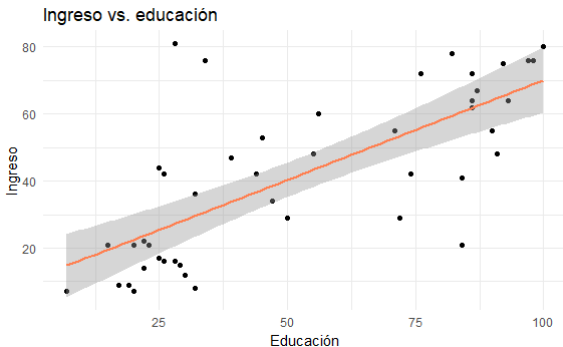
ajustamos un modelo lineal por OLS y visualizamos la calidad del ajuste

```
duncan_lm <- lm( income ~ education, data=Duncan)
summary(duncan_lm)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.6035	5.1983	2.04	0.0475
education	0.5949	0.0863	6.89	0.0000

Residual standard error: 17.04 on 43 degrees of freedom
 Multiple R-squared: 0.5249, Adjusted R-squared: 0.5139
 F-statistic: 47.51 on 1 and 43 DF, p-value: 1.84e-08

Visualizamos la calidad del ajuste



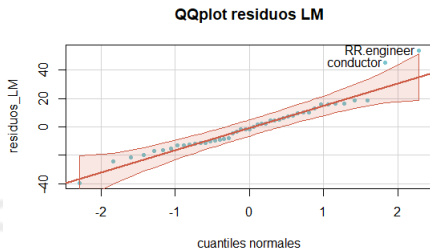
Se observa la presencia de datos alejados de la recta de ajuste, pueden ser outliers y también puntos influyentes. Vamos a realizar el análisis diagnóstico del modelo.

Normalidad de los residuos

Shapiro-Wilk normality test

data: residuos

$W = 0.94462$, $p\text{-value} = 0.03189$

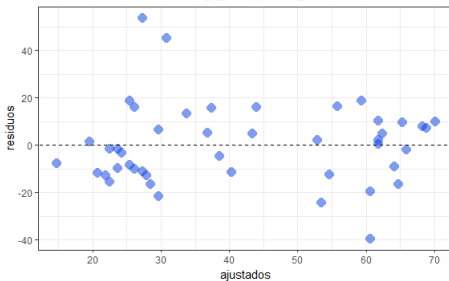


Homocedasticidad de los residuos

studentized Breusch-Pagan test

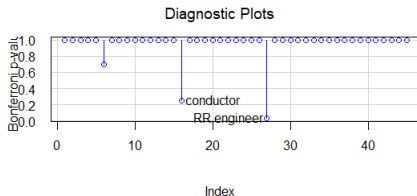
data: duncan_lm

BP = 0.28649, df = 1, p-value = 0.5925



Exploramos la presencia de outliers

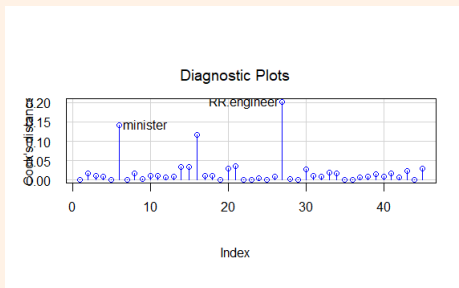
rrstudent unadjusted p-value Bonferroni p RR.engineer 3.646444
0.00072729 0.032728



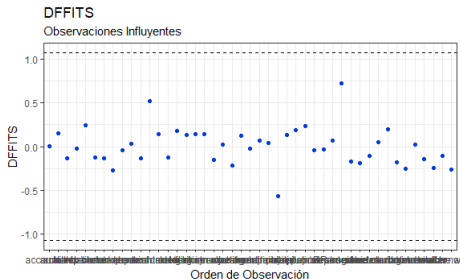
Se aprecian como atípicos los valores del ingeniero y del conductor.

Medidas de Influencia

No se detectan valores con alto leverage, si con distancia de Cook por encima de la cota.



DDFITS y DFBETA



Las observaciones de los DFFITS y de los DFBeta no denotan valores atípicos, mientras que en el caso de las distancias de Cook se señalan dos casos especiales el ministro y el ingeniero.

Estimación Robusta por Huber

```
library(robustbase); library(MASS)
library(olsrr); library(car); library(quantreg)
# Ajustamos un modelo robusto con la propuesta de Huber
duncan_Huber <- rlm(income ~ education, data = Duncan, k2 =
1.345)
duncan.lm$coefficients
```

```
rlm(formula = income ~ education, data = Duncan, k2 = 1.345)
Converged in 9 iterations
Coefficients:
(Intercept) education
6.3002197 0.6615263
Degrees of freedom: 45 total; 43 residual
```

Estimación Robusta por LTS

Ajustamos un modelo robusto con la propuesta de LTS

```
duncan_LTS <- lqs(income ~ education, data = Duncan, method  
= 'lts')  
duncan_LTS
```

```
lqs.formula(formula = income ~ education, data = Duncan, method  
= 'lts')
```

Coefficients:

(Intercept) education

-2.6929 0.8125

Scale estimates 10.43 11.19

Estimación Robusta por LAD

Ajustamos un modelo robusto con la propuesta de LAD

```
duncan_LAD <- rq(income ~ education, data = Duncan, tau =  
0.5)
```

```
duncan_LAD
```

```
rq(formula = income ~ education, tau = 0.5, data = Duncan)
```

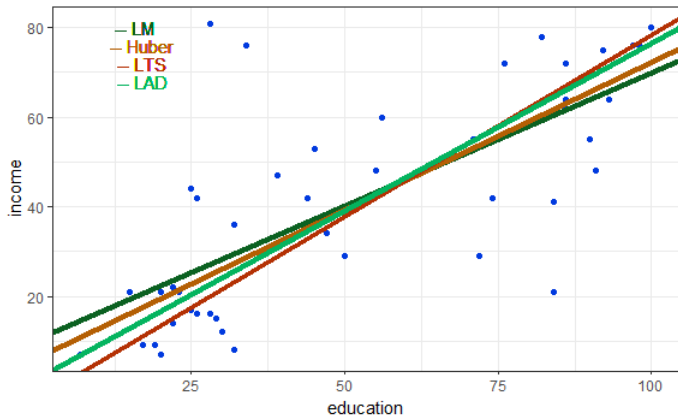
Coefficients:

(Intercept) education

1.75 0.75

Degrees of freedom: 45 total; 43 residual

Comparamos gráficamente las estimaciones



Comparación analítica de los ajustes

```
set.seed(2) # fijamos una semilla para reproducibilidad
train = Duncan %>% sample_frac(0.8) # separamos la base
test = Duncan %>% setdiff(train)
income=test$income # guardamos la v.resp de test
ols_Predicted <- predict(duncan_lm, test) # guardamos predichos
lts_Predicted <- predict(duncan_LTS, test)
lad_Predicted <- predict(duncan_LAD, test)
hub_Predicted <- predict(duncan_Huber, test)
ols_pred <- cbind(ols_Predicted, income) # armamos base con
observados y ajustados
lts_pred <- cbind(lts_Predicted, income)
lad_pred <- cbind(lad_Predicted, income)
hub_pred <- cbind(hub_Predicted, income)
```

```
EM_OLS=mean(apply(ols_pred, 1, min)/apply(ols_pred, 1, max))
EM_LTS=mean(apply(lts_pred, 1, min)/apply(lts_pred, 1, max))
EM_LAD=mean(apply(lad_pred, 1, min)/apply(lad_pred, 1, max))
EM_HUB=mean(apply(hub_pred, 1, min)/apply(hub_pred, 1, max))
EM=c(EM_OLS,EM_LTS,EM_HUB,EM_LAD)
sal=cbind(EM)
rownames(sal)<-c('OLS','LTS','HUB','LAD') sal%>%kbl()%>%
kable_classic()
```

Tabla: Errores Medios y Coeficiente de Determinación de los Modelos Ajustados

	EM
OLS	0.78
LTS	0.78
HUB	0.81
LAD	0.80

Los errores medios cometidos definidos como el promedio de las proporciones entre el mínimo y máximo entre ajustado y observado es similar, aunque el M-estimador de Huber logra mejorar esta proporción levemente. En cuanto al coeficiente de determinación el LAD logra el mejor ajuste. Hay que destacar que la separación de conjuntos de testeo y validación se hace con una semilla aleatoria, al variar esta semilla pueden variar estos valores.