

Regresión Avanzada

Problemas de la Regresión

Adecuación del modelo

Supuesto que se hacen al ajustar un modelo lineal:

- ▶ la relación entre la variable regresora y las de respuesta es (aproximadamente) lineal.
- ▶ el error tiene media cero.
- ▶ el error tiene varianza constante.
- ▶ los errores son no correlacionados.
- ▶ los errores están normalmente distribuidos.

Adecuación del modelo

Si los supuestos no se cumplen:

- ▶ los resultados no son significativos.
- ▶ el ajuste se vuelve inestable (muestras distintas ajustes muy diferentes).

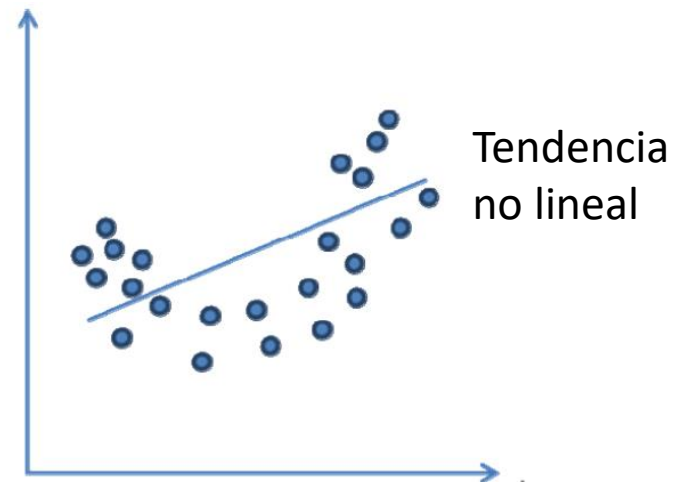
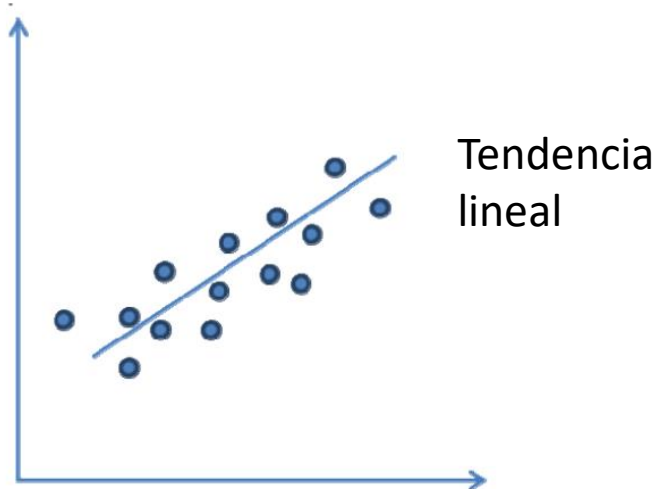
Importante: esta información no está ni en el R^2 , ni en el test F , ni en la RSS , etc. A continuación daremos técnicas de diagnóstico para detectar estos problemas.

Adecuación del modelo

Verificando linealidad entre la variable de respuesta y las variables regresoras.

Regresión simple:

En este caso basta hacer una scatter plot entre la variable de respuesta y y la variable regresora X .

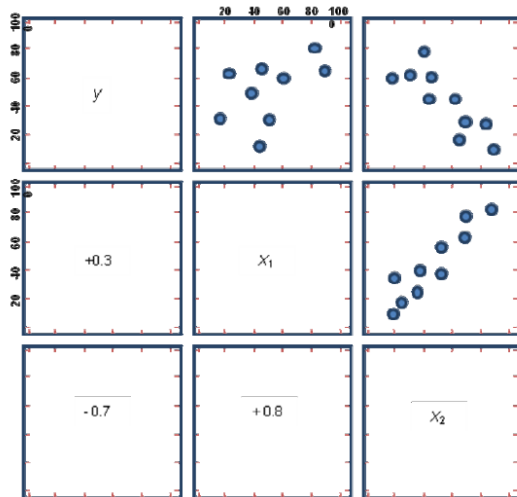


Adecuación del modelo

Regresión múltiple:

Hacer los scatter plots entre la variable de respuesta y y cada una de las variables regresoras X_j .

Mirar los coeficientes de correlación entre la variable de respuesta y y cada una de las variables regresoras X_j .



El coeficiente de correlación mide únicamente relación lineal.

El coeficiente de correlación no es robusto.

Si se observa relación lineal se afianza el modelo, pero no implica que el mismo este mal. Ya que la relación lineal no es de a pares sino entre la variable de respuesta y el vector de variables regresoras.

Adecuación del modelo

Consideramos el conjunto de datos Auto de la librería ISLR. El conjunto de datos está conformado por 9 variables que muestran diferentes características de autos. Hay 392 observaciones.

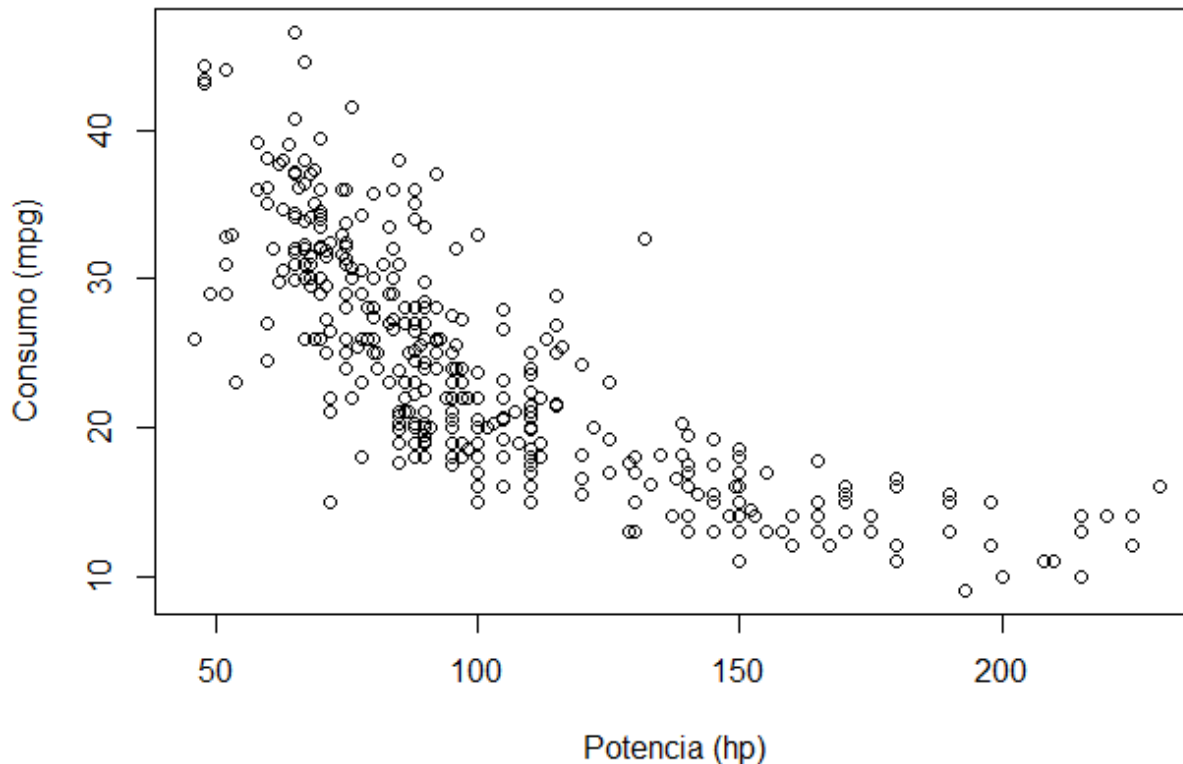
Se quiere explicar el consumo de combustible en función de estas características.


Comenzamos analizando el consumo de combustible (mpg) en función de la potencia (hp) del auto.

Adecuación del modelo

```
library(ISLR)
```

```
plot(Auto$horsepower,Auto$mpg,xlab="Potencia (hp)", ylab="Consumo (mpg)")
```



El gráfico sugiere
 incorporar un
término cuadrático

Adecuación del modelo

```
auto.lm1<-lm(Auto$mpg ~ Auto$horsepower)
summary(auto.lm1)
```

Call:
lm(formula = Auto\$mpg ~ Auto\$horsepower)

La regresión es significativa,
El R^2 es moderado.

Residuals:

Min	1Q	Median	3Q	Max
-13.5710	-3.2592	-0.3435	2.7630	16.9240

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	39.935861	0.717499	55.66	<2e-16 ***
Auto\$horsepower	-0.157845	0.006446	-24.49	<2e-16 ***

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

Residual standard error: 4.906 on 390 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

Adecuación del modelo

En este caso el modelo quedaría,

$$\text{mpg} \approx 39.94 - 0.16 \text{ hp}$$

$$R^2=0.606$$

$$\text{adj}R^2=0.605$$

```
hp2=Auto$horsepower^2
auto.lm2<-lm(Auto$mpg ~ Auto$horsepower + hp2)
summary(auto.lm2)
```

Call:

```
lm(formula = Auto$mpg ~ Auto$horsepower + hp2)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.7135	-2.5943	-0.0859	2.2868	15.8961

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.9000997	1.8004268	31.60	<2e-16 ***
Auto\$horsepower	-0.4661896	0.0311246	-14.98	<2e-16 ***
hp2	0.0012305	0.0001221	10.08	<2e-16 ***

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

Residual standard error: 4.374 on 389 degrees of freedom

Multiple R-squared: 0.6876, Adjusted R-squared: 0.686

F-statistic: 428 on 2 and 389 DF, p-value: < 2.2e-16

Adecuación del modelo

En este caso el modelo quedaría,

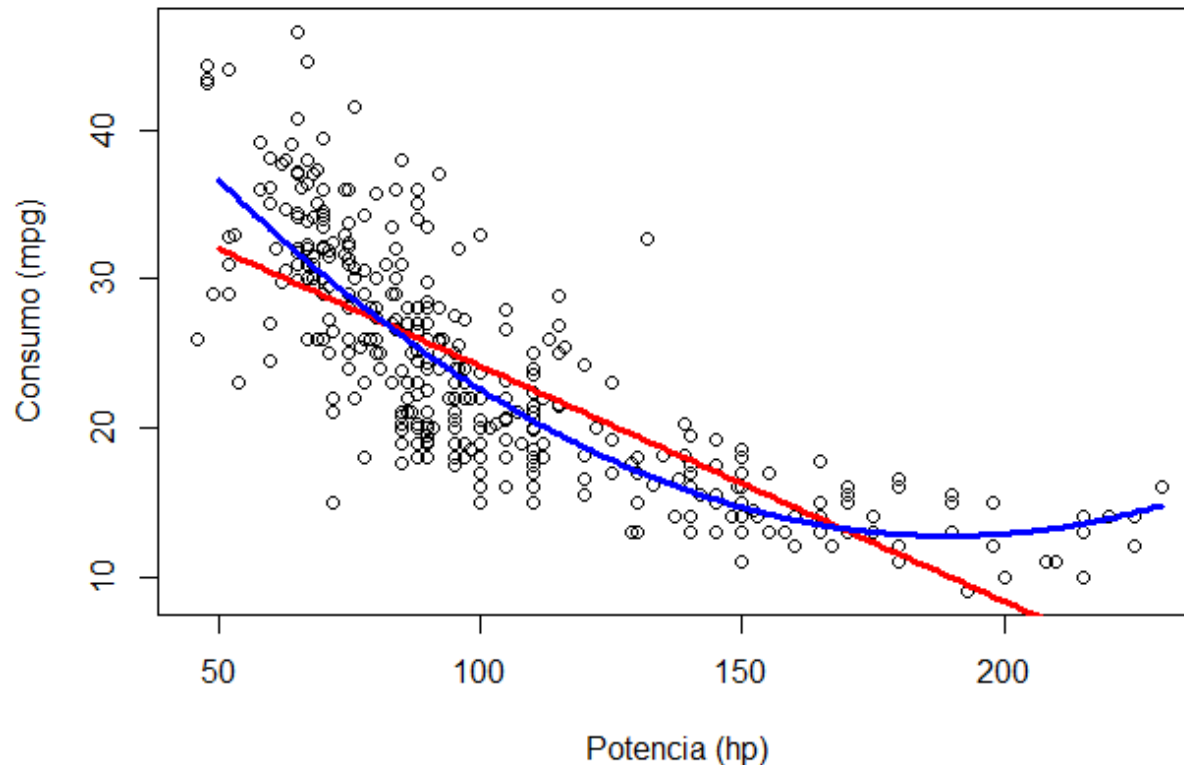
$$\text{mpg} \approx 56.9 - 0.47 \text{ hp} + 0.001 \text{ hp}^2$$

$$R^2=0.688$$

$$\text{adj}R^2=0.686$$

Adecuación del modelo

```
plot(Auto$horsepower,Auto$mpg,xlab="Potencia (hp)",ylab="Consumo (mpg)")  
lines(50:230, auto.lm1$coefficients[1]+auto.lm1$coefficients[2]*50:230,col = "red", lwd = 3)  
lines(50:230, auto.lm2$coefficients[1]+auto.lm2$coefficients[2]*50:230+  
      auto.lm2$coefficients[3]*(50:230)^2,col = "blue", lwd = 3)
```



Análisis de los residuos Recordemos que el i -ésimo residuo es la diferencia entre el i -ésimo valor observado y su predicción,

$$r_i = y_i - \hat{y}_i.$$

Los residuos miden la variabilidad en la variable de respuesta que no es explicada por la regresión.

Algunas propiedades de los residuos:

- ▶ $E(r_i) = 0$.
- ▶ Un estimador de la varianza de los residuos es

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (r_i - \bar{r})^2}{n - p}.$$

- ▶ Los residuos no son independientes ya que sus grados de libertad son $n - p$ en n residuos. En general

$$p \ll n - p$$

entonces esa dependencia no suele afectar el análisis de los residuos.

Adecuación del modelo

La unidad de medida de los residuos es la de la variable de respuesta, por lo tanto, para realizar su análisis es necesario escalarlos.

Estandarización de los residuos

Ya que $E(r_i) = 0$ proponemos,

$$d_i = \frac{r_i}{\hat{\sigma}}.$$

Luego, $E(d_i) = 0$ y $var(d_i) = 1$.

En general, si $|d_i| > 3$ indica un potencial outlier.

Adecuación del modelo

Estudentización de los residuos

En lugar de usar un estimador de la varianza de r_i , $\hat{\sigma}$, usaremos la varianza exacta de r_i .

En primer lugar tenemos que $r = (I - H)\epsilon$.

Luego,

$$\text{var}((I - H)\epsilon) = (I - H)\text{var}(\epsilon)(I - H) = \sigma^2(I - H).$$

Como la matriz $(I - H)$ es simétrica e idempotente, pero generalmente no es diagonal, los residuos tienen diferente varianza y suelen no ser independientes.

Adecuación del modelo

Luego sea h_{ij} al elemento que está en la posición (i,j) -ésima de la matriz H .

Entonces,

$$\begin{aligned} \text{var}(r_i) &= \sigma^2(1 - h_{ii}), \\ \text{cov}(r_i, r_j) &= -\sigma^2 h_{ij}. \end{aligned}$$

Como $0 < h_{ii} < 1$ entonces $\hat{\sigma}^2$ sobreestima la varianza de r_i .

Finalmente $\widehat{\text{var}}(r_i) = \hat{\sigma}^2(1 - h_{ii})$.

Observación: h_{ii} es el **leverage** del i -ésimo dato, en cierto modo mide el aporte de la i -ésima observación en la varianza muestral, no depende del valor observado y_i .

Adecuación del modelo

- ▶ h_{ij} es una medida de locación de la observación x_i en la nube de puntos de las x .
- ▶ Los puntos con h_{ij} alto estarán en la zona perisférica de la nube de puntos de las $x = (x_1, \dots, x_{p-1})$, que son los puntos donde la densidad de las x es baja.
- ▶ Los puntos con h_{ij} bajo, tienen varianza alta, esto indica que el ajuste es malo en la zona central de la nube de puntos.

Las violaciones al modelo suelen ocurrir debido a puntos perisféricos de la nube de puntos de x , que son difíciles de identificar con los residuos o con los residuos estandarizados, porque en general son menores.

Adecuación del modelo

Residuos estudentizados:

$$sr_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

tenemos que $E(sr_i) = 0$ y $var(sr_i) = 1$ independientemente de la posición de x_i cuando se cumplen los supuestos del modelo.

Una observación tiene mucha influencia en la regresión lineal cuando:

- ▶ tiene sr_i alto.
- ▶ tiene h_{ii} alto.

Adecuación del modelo

Residuos PRESS

Se calcula el residuo considerando la el valor ajustado para la observación considerando la regresión hecha con todos los datos menos el i -ésimo. $r_{(i)} = y_i - \hat{y}_{(i)}$. Si (x_i, y_i) fuese un valor atípico se lo detectaría.

- ▶ Sacar la i -ésima observación.
- ▶ Ajustar la regresión con las $n - 1$ observaciones restantes.
- ▶ Calcular el valor ajustado en base a la regresión hallada en el item anterior, $\hat{y}_{(i)}$.
- ▶ Calcular el *residuo* correspondiente $r_{(i)} = y_i - \hat{y}_{(i)}$.
- ▶ Realizar este procedimiento para todas las observaciones.

Adecuación del modelo

Este procedimiento parece ser muy engorroso, sin embargo, se puede ver que,

$$r_{(i)} = \frac{r_i}{1 - h_{ii}},$$

donde $h_{ii} = x_i(X'X)^{-1}x_i'$, igual que en los residuos estudentizados. Luego, los residuos PRESS son los residuos ponderados, donde los ponderadores se obtienen de la diagonal de la matriz H .

- ▶ observaciones con alto h_{ii} , tienen alto residuo PRESS y son puntos muy influyentes en la regresión.
- ▶ una diferencia grande entre el residuo y el residuo PRESS indica un punto donde la regresión ajusta bien, sacar ese punto empobrece el modelo.

Adecuación del modelo

Dificultad,

$$\text{var}(r_{(i)}) = \frac{\sigma^2}{1 - h_{ii}},$$

cambia para cada observación, luego se propone estandarizarla,
residuo PRESS estandarizado,

$$\frac{r_{(i)}}{\sqrt{\text{var}(r_{(i)})}} = \frac{\frac{r_i}{1-h_{ii}}}{\sqrt{\frac{\sigma^2}{1-h_{ii}}}} = \frac{r_i}{\sqrt{\sigma^2(1-h_{ii})}}.$$

Los residuo PRESS estandarizado coinciden con los residuos estudentizados.

R-student

Un problema que persiste es que al estimar σ^2 se utiliza toda la muestra. Se propone estimar la escala para cada observación removiéndola,

$$s_{(i)}^2 = \frac{1}{n - p} \left[(n - k - 1) \hat{\sigma}^2 - \frac{r_i^2}{1 - h_{ii}} \right].$$

Luego, el **residuo externamente estudentizado** o **R-student** está dado por

$$t_i = \frac{r_i}{\sqrt{s_{(i)}^2 (1 - h_{ii})}}.$$

- ▶ si la i -ésima observación es muy influyente la estimación de la varianza puede diferir en forma marcada.
- ▶ bajo el modelo t_i tiene distribución t de student con $n - p$ grados de libertad.

Adecuación del modelo

Gráficos de Residuos

- ▶ Se realizan varios gráficos basados en los residuos para verificar si se cumplen los supuestos del modelo
- ▶ es conveniente graficar los residuos, los residuos estandarizados y estudentizados, para observar los cambios en los gráficos.

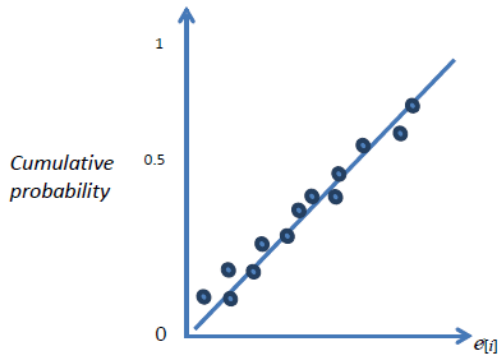
Adecuación del modelo

QQ-plot de los residuos

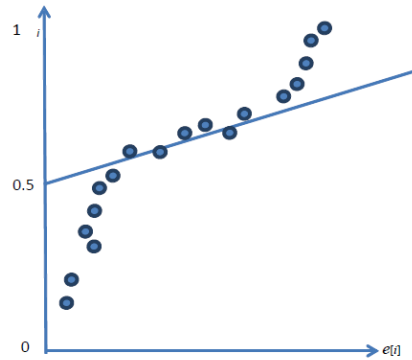
- ▶ se utiliza para verificar el supuesto de normalidad de los errores.
- ▶ si esto no se cumple se ve comprometida la validez de los test de hipótesis e intervalos de confianza y predicción.
- ▶ distribuciones con colas pesadas señalan que la regresión es sensible a un conjunto pequeño de observaciones.

Se grafican los residuos ordenados de menor a mayor r_i versus $\Phi^{-1} \left[\frac{i-0.5}{n} \right]$, donde Φ es la cdf de la distribución normal estándar. Si estos valores se parecen deberían alinearse.

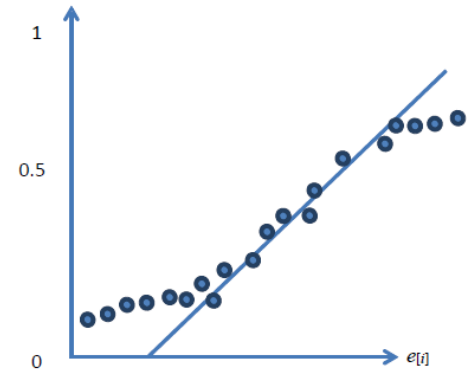
Adecuación del modelo



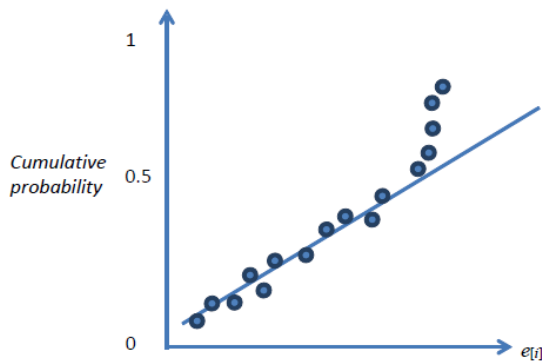
Distribución normal



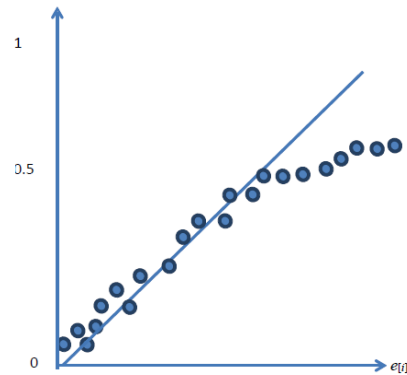
Colas pesadas



Colas livianas



Asimetría positiva



Asimetría negativa

Adecuación del modelo

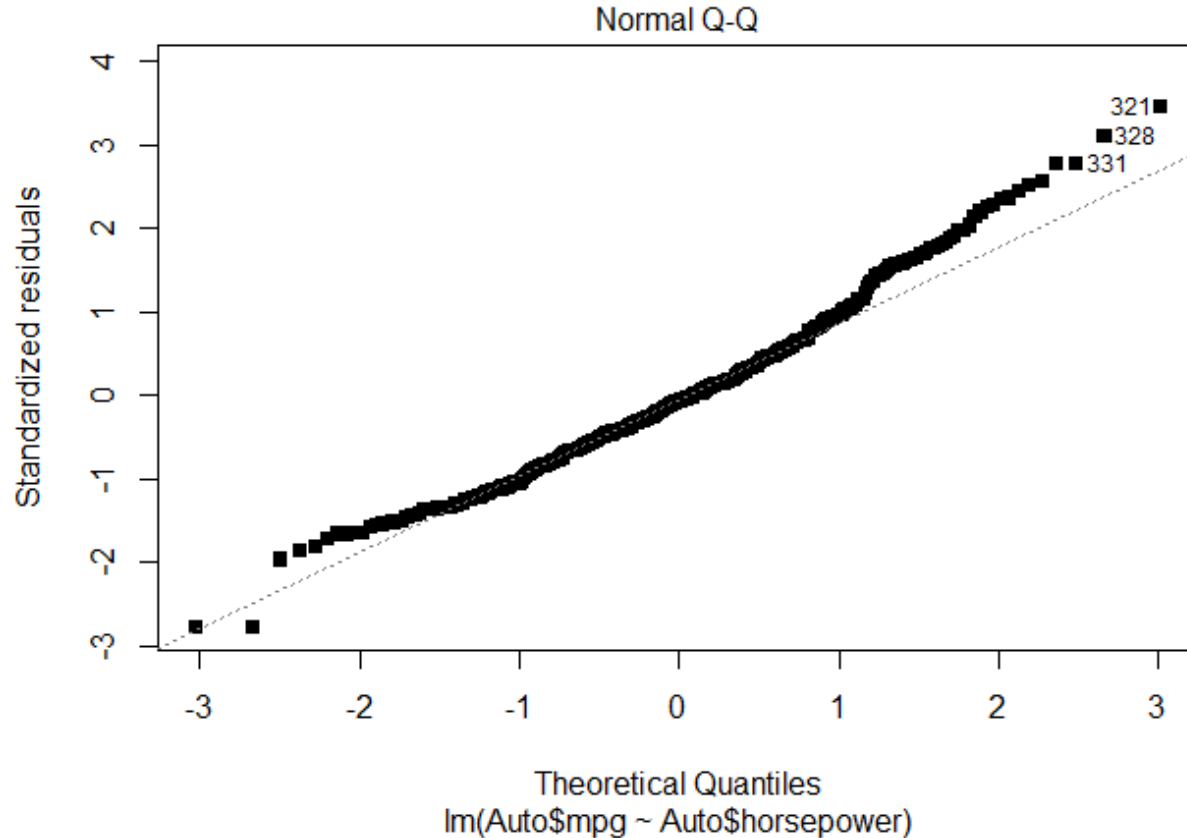
Observaciones:

- ▶ QQ-plot basados en muestras chicas ($n < 16$) son difíciles de analizar, aun generandolos normales no se ven de ese modo. Preferiblemente muestras más grandes ($n > 30$).
- ▶ Muestras no es aleatoria, en ocasiones no se aprecia en este gráfico.
- ▶ Muchos defectos de este gráfico provienen de la distorsión que provocan unas pocas observaciones atípicas.

Adecuación del modelo

Retomamos el ejemplo del Auto. QQ plot para el modelo lineal

```
plot(auto.lm1,which=2,pch=15)
```

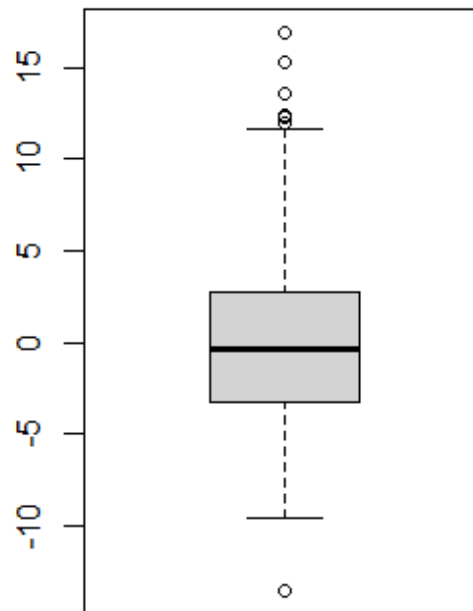
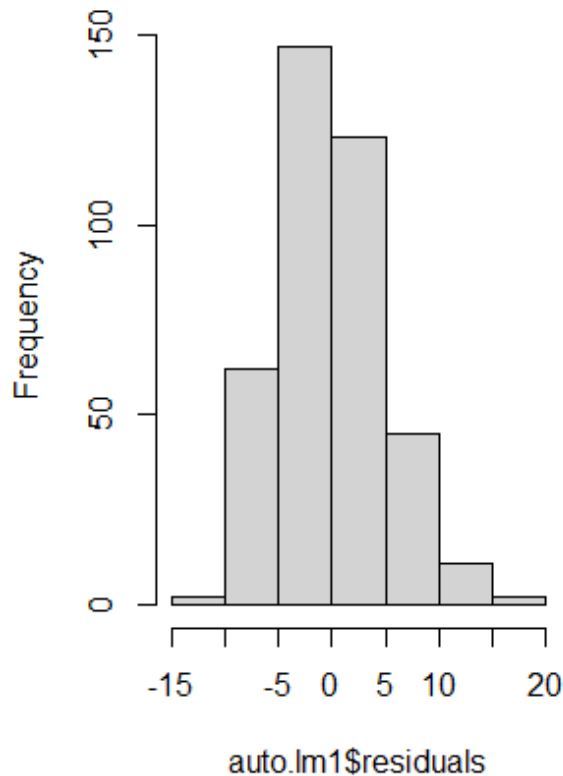


Asimetría positiva
leve

Adecuación del modelo

```
par(mfrow=c(1,2))  
hist(auto.lm1$residuals)  
boxplot(auto.lm1$residuals)
```

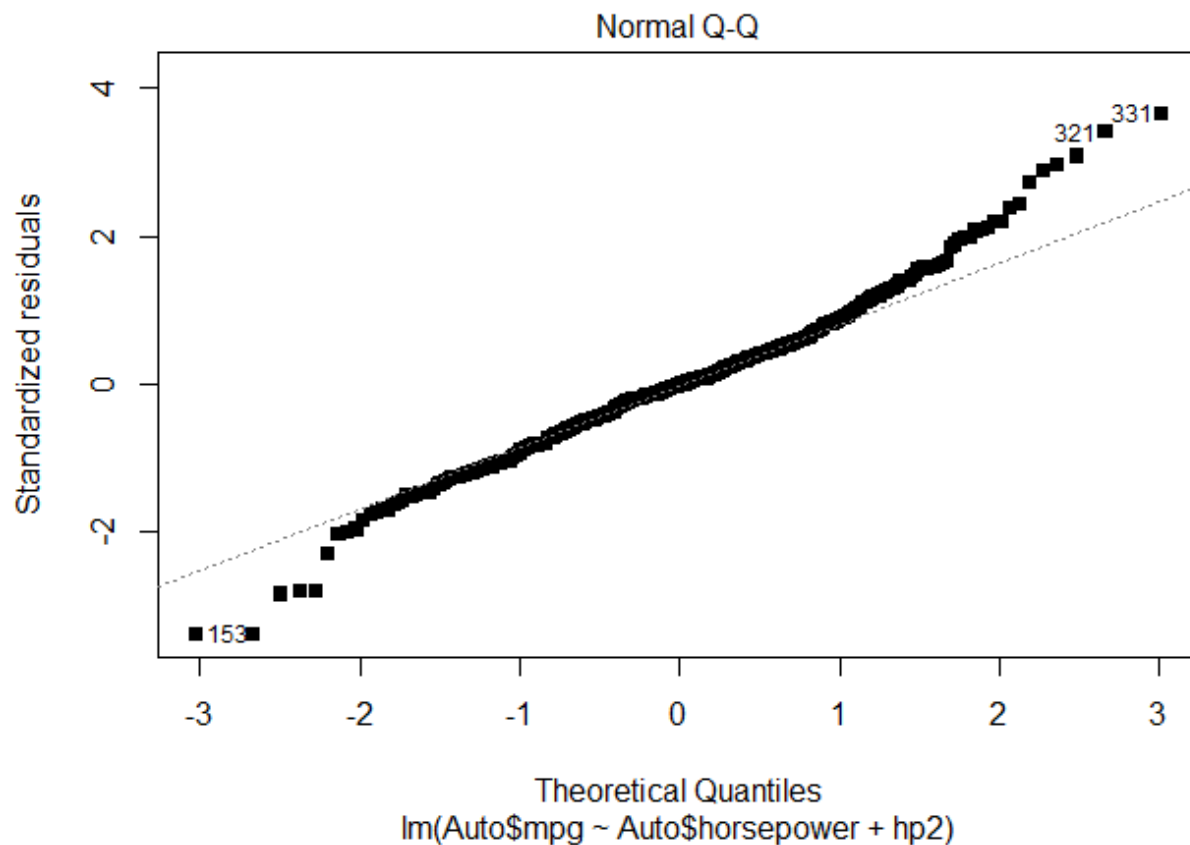
Histogram of auto.lm1\$residuals



Complementamos el análisis con el histograma y boxplot. Se confirma el análisis

Retomamos el ejemplo del Auto. QQ plot para el modelo cuadrático.

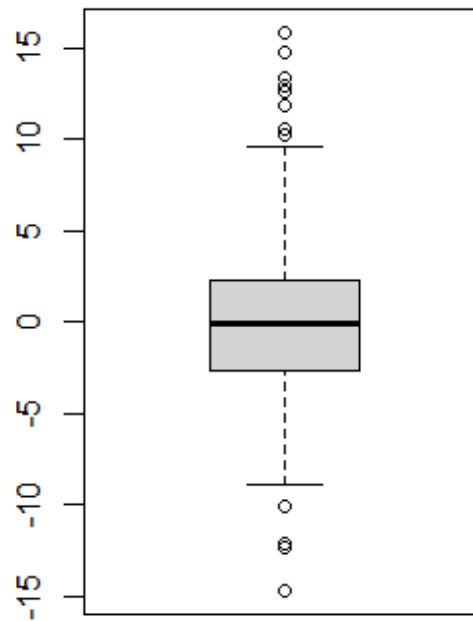
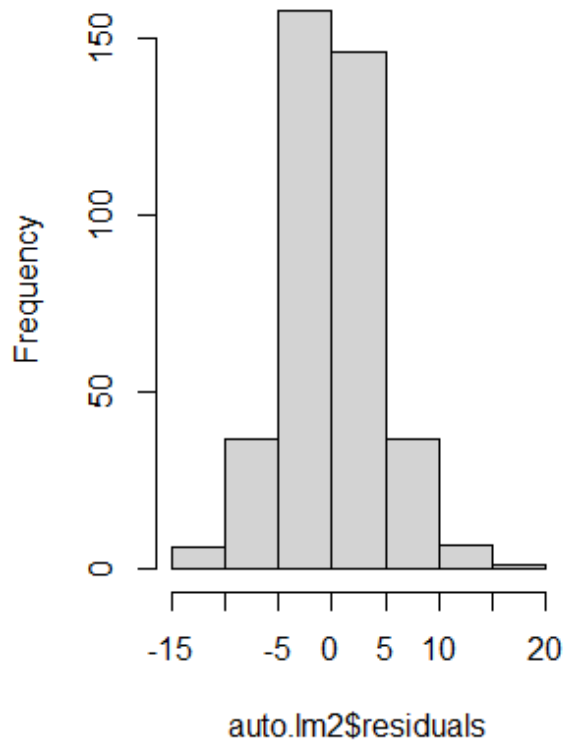
```
plot(auto.lm2,which=2,pch=15)
```



Colas pesadas
leve

```
par(mfrow=c(1,2))  
hist(auto.lm2$residuals)  
boxplot(auto.lm2$residuals)
```

Histogram of auto.lm2\$residuals



Confirmamos lo observado en el qqplot-

Test de Kolmogorov-Smirnov

- ▶ es un test de bondad de ajuste no paramétrico.
- ▶ se testea si una muestra sigue determinada distribución.
- ▶ se puede extender a testear si dos muestras siguen la misma distribución.
- ▶ puede ser unilateral o bilateral, hay que entender que quiere decir esto, pero excede los temas de la materia.
- ▶ como todo test de bondad de ajuste *queremos* no rechazar H_0 , es decir que lo que se quiere probar va en la hipótesis nula (al revés de lo que hacemos en general!)

Test de Kolmogorov-Smirnov

Sea X una variable aleatoria, queremos testear si su distribución tiene cdf F . Sea X_1, \dots, X_n , una muestra aleatoria de F . Notamos F_n a la cdf empírica, i.e. $F_n = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}$.

$$H_0 : X \sim F \text{ vs. } H_0 : X \not\sim F.$$

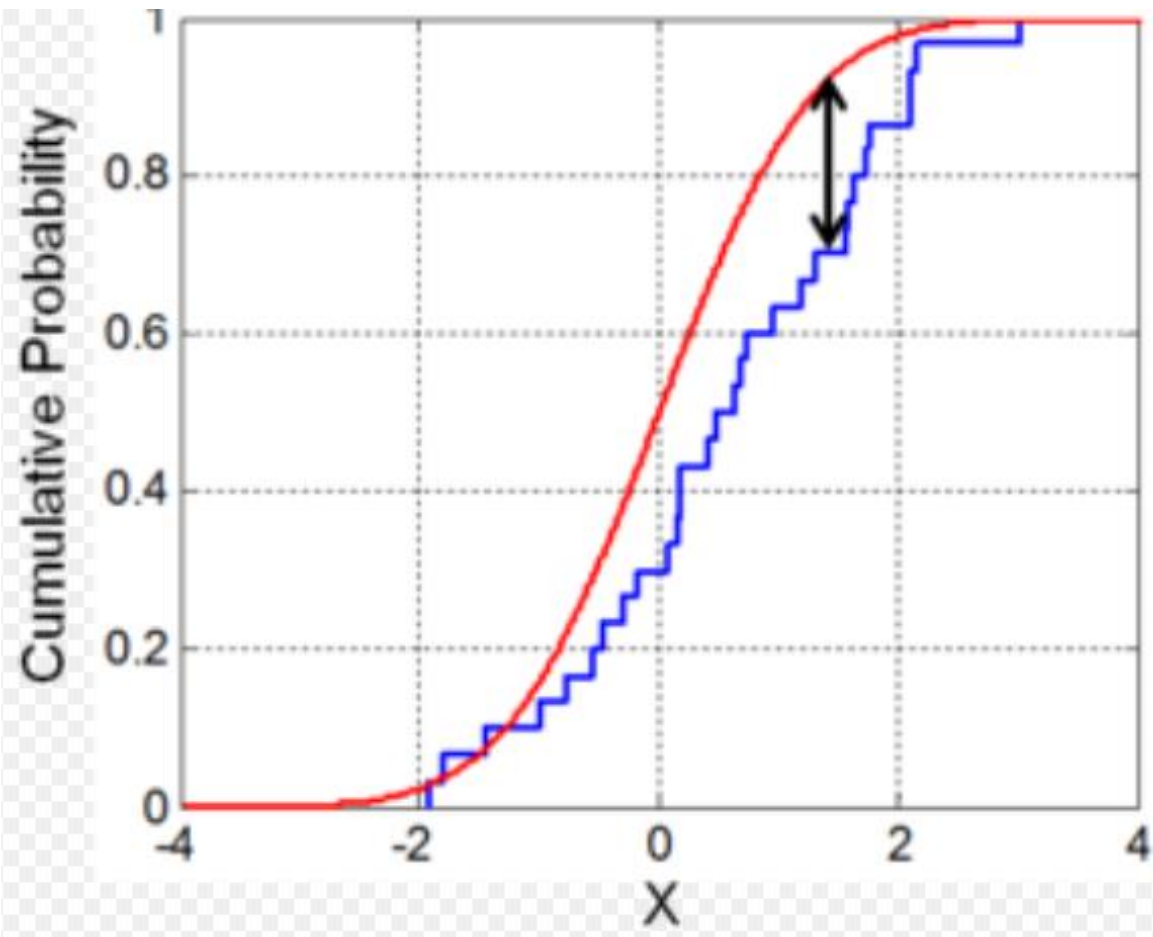
El estadístico del test es

$$D_n = \max_x (|F(x) - F_n(x)|),$$

bajo determinadas condiciones en F , H_0 tiene distribución conocida D (supremo del puente browniano)

Rechazo H_0 si $D_n > D_{1-\alpha}$

Observación: Si la distribución no es conocida el estadístico del test se calcula via bootstrap.



Adecuación del modelo

Hacemos el test de bondad de ajuste para testear si los residuos del modelo lineal siguen una distribución normal.

```
ks.test(auto.lm1$residuals,"pnorm",mean(auto.lm1$residuals),sd(auto.lm1$residuals))
```

One-sample Kolmogorov-Smirnov test

data: auto.lm1\$residuals

D = 0.060525, p-value = **0.1131**

alternative hypothesis: two-sided

A nivel 10% no rechazamos la hipótesis nula, de todos modos es deseable tener p-valor mayor a 0.2

Repetimos el test para el modelo de segundo orden.

```
ks.test(auto.lm2$residuals,"pnorm",mean(auto.lm2$residuals),sd(auto.lm2$residuals))
```

One-sample Kolmogorov-Smirnov test

data: auto.lm2\$residuals

D = 0.058508, **p-value = 0.1366**

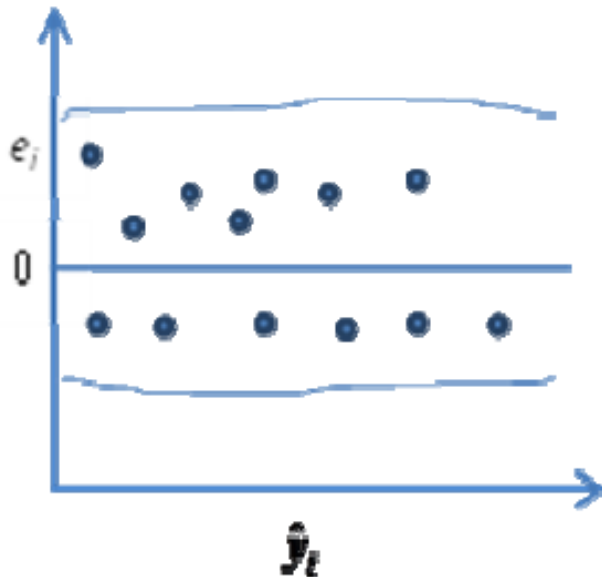
alternative hypothesis: two-sided

La conclusión es similar.

Adecuación del modelo

Gráficos de los residuos versus los valores ajustados

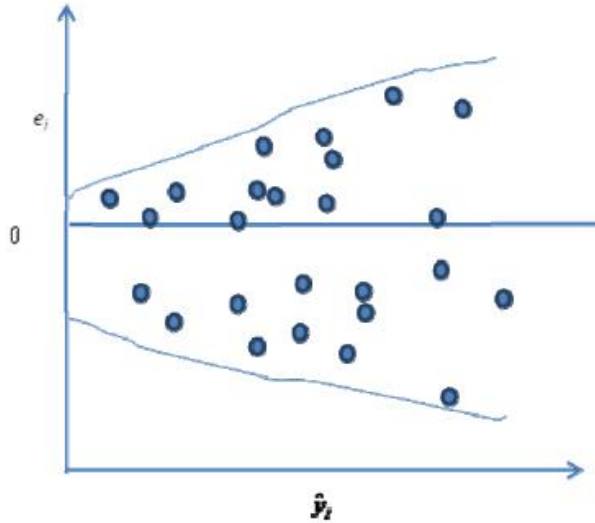
Es útil graficar los residuos r_i (o alguna de sus estandarizaciones) versus los \hat{y}_i .



Residuos fluctúan sin estructura, en una banda horizontal.

No hay efectos visibles del modelo.

Adecuación del modelo

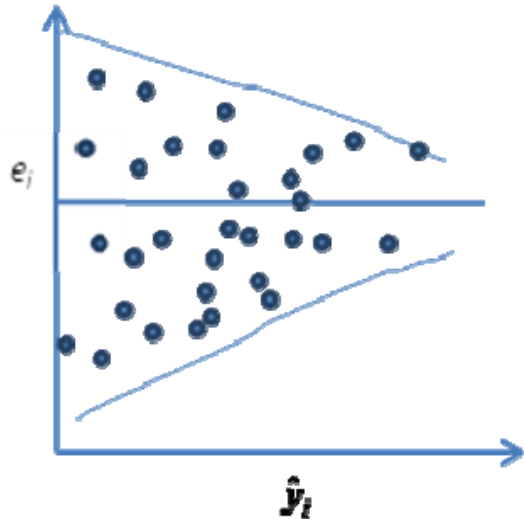


Cono que se abre.

Los errores no son constantes, crecen con y

Soluciones posibles:

- Transformaciones de las x 's
- Transformar la y
- Mínimos cuadrados pesados



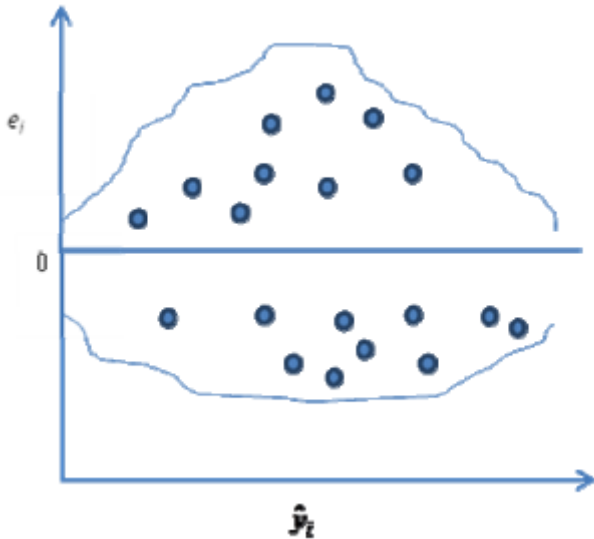
Cono que se cierra.

Los errores no son constantes, decrecen con y

Soluciones posibles:

- Transformaciones de las x 's
- Transformar la y
- Mínimos cuadrados pesados

Adecuación del modelo



Los errores no son constantes, son una proporción de entre 0 y 1.

La y puede tener distribución binomial.

Soluciones posibles:

- Transformaciones de las x 's
- Transformar la y
- Mínimos cuadrados pesados

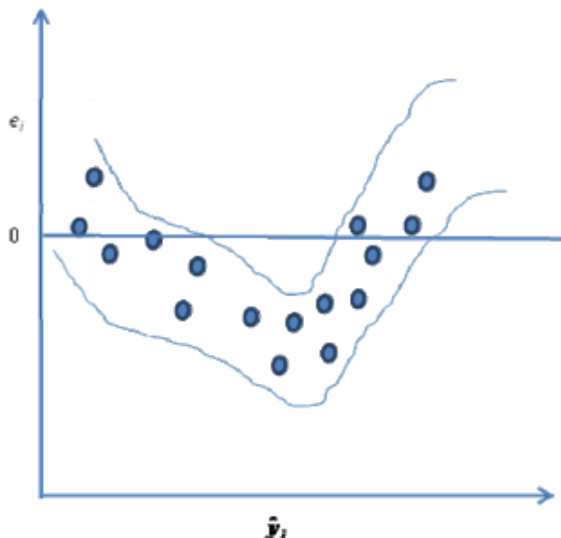


Gráfico curvilíneo indica que la relación entre la y y la X es no lineal.

Podría indicar que otras variables explicativas

Son necesarias en el modelo.

Soluciones posibles:

- Transformaciones de las x 's
- Transformar la y

Adecuación del modelo

Gráficos de los residuos versus los valores ajustados

Es útil graficar los residuos r_i (o alguna de sus estandarizaciones) versus los \hat{y}_i .

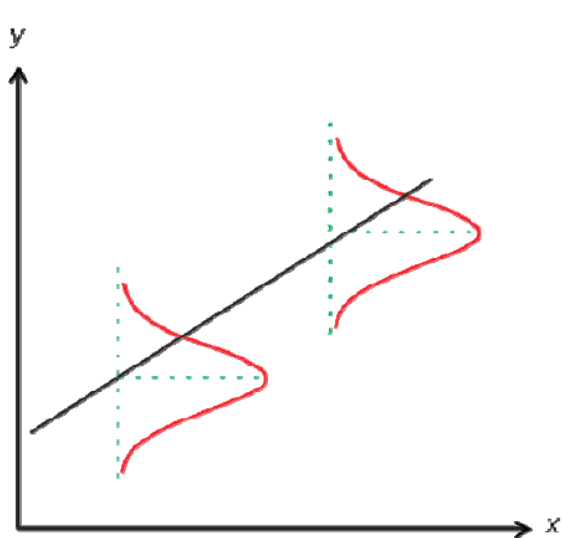
En general residuos altos ocurren con valores extremos de \hat{y}_i . Esto pueden indicar:

- ▶ varianza no es constante.
- ▶ la relación entre y y X no es constante.

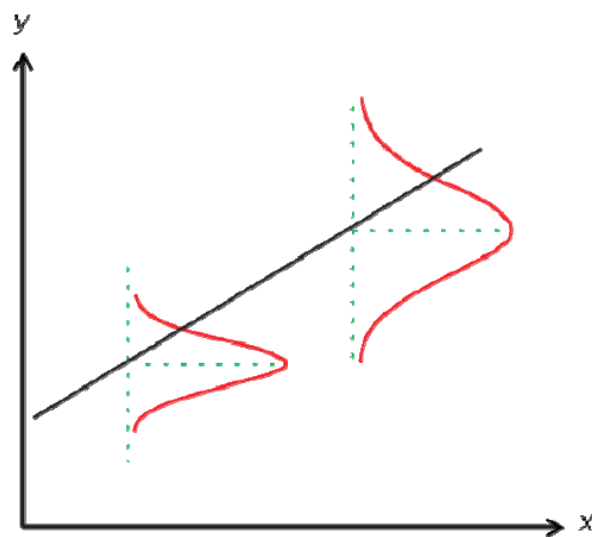
Hay que explorar estas alternativas antes de determinar que son outliers.

Heterocedasticidad

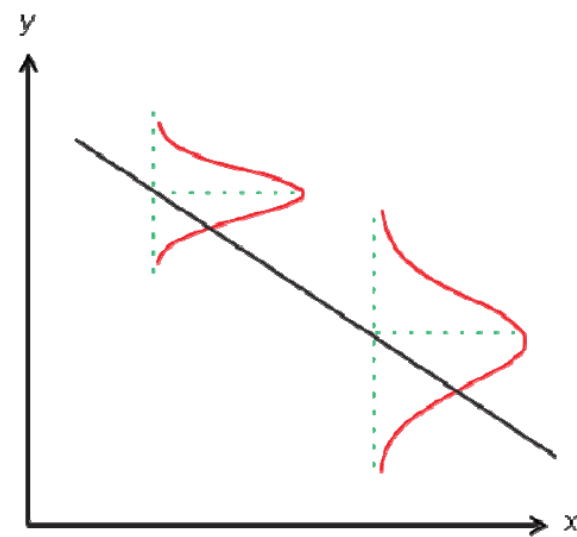
Uno de los supuestos fundamentales que hicimos es que en un modelo lineal $var(\epsilon_i) = \sigma^2$ para todo $i = 1, \dots, n$ y además que las $cor(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$.



Homocedástico



Heterocedástico



Heterocedástico

Posibles causas:

- ▶ Propios de la naturaleza del fenómenos. A mayor ingreso mayor gasto en comida. Más horas de práctica, menor cantidad de errores cometidos.
- ▶ Si las observaciones son promedios, este procedimiento introduce heterocedasticidad. Por ejemplo en lugar de calcular el gasto en ropa de una familia se considera cuanto gasta cada individuo en promedio. En este caso tendríamos $\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{\epsilon}_i$, donde \bar{x}_i es el gasto medio de la familia i -ésima. Luego $\bar{\epsilon}_i$ es el correspondiente error, que tiene esperanza cero pero su varianza será σ^2/m donde m es el tamaño del grupo familiar, la varianza no es constante.

Posibles causas:

- ▶ Especificaciones teóricas del modelo que deben ser corregidas en el mismo. Por ejemplo, poner fertilizante hace que aumente la cosecha hasta un punto que satura y la cosecha se quema. Luego, para predecir la cosecha en función del fertilizante no podemos poner un modelo lineal con varianza constante.
- ▶ Distribuciones asimétricas en alguna de las variables regresoras.
- ▶ Utilizar un modelo funcional no adecuado para el problema.

Test de Rangos de Spearman

Testeamos

$$H_0 : \sigma_1^2 = \dots = \sigma_n^2 \text{ vs } H_A : \exists i, j \text{ tq } \sigma_i^2 \neq \sigma_j^2.$$

Para ello, considerar el modelo $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

1. Hacer la regresión y obtener los residuos, r_i
2. Considerar el valor absoluto, $|r_i|$.
3. Asignar los rangos en orden ascendente a $|r_i|$ y a \hat{y}_i .
4. Calcular la correlación de Spearman entre ambos, i.e.,

$$\rho_{|r|, \hat{y}} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

donde d_i es la diferencia de rangos entre $|r_i|$ y \hat{y}_i .

Test de Rangos de Spearman

5. Asumimos que la correlación poblacional es cero y $n \geq 8$, considerar el estadístico,

$$t_0 = \frac{\rho_{|r|,\hat{y}} \sqrt{n-2}}{\sqrt{1 - \rho_{|r|,\hat{y}}^2}},$$

que bajo la hipótesis nula tiene una distribución t_{n-2} .

6. Rechazamos H_0 si $t_0 > t_{n-2,1-\alpha}$.

Volvemos al ejemplo de los autos

```
cor.test(abs(auto.lm2$residuals), auto.lm2$fitted.values, method="spearman")
```

Spearman's rank correlation rho

data: abs(auto.lm2\$residuals) and auto.lm2\$fitted.values

S = 6439957, p-value = **2.469e-13** → hay heterocedasticidad

alternative hypothesis: true rho is not equal to 0

sample estimates:

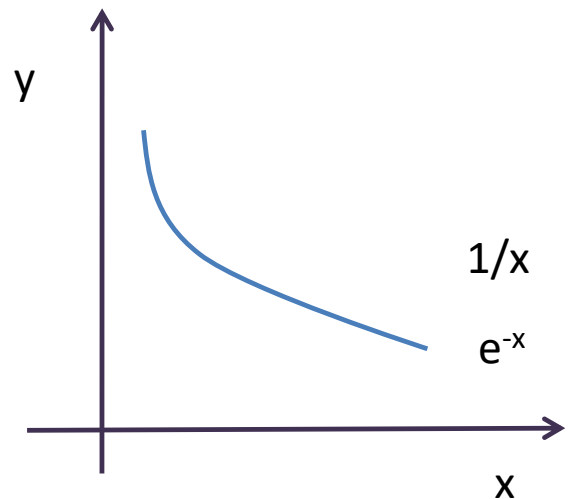
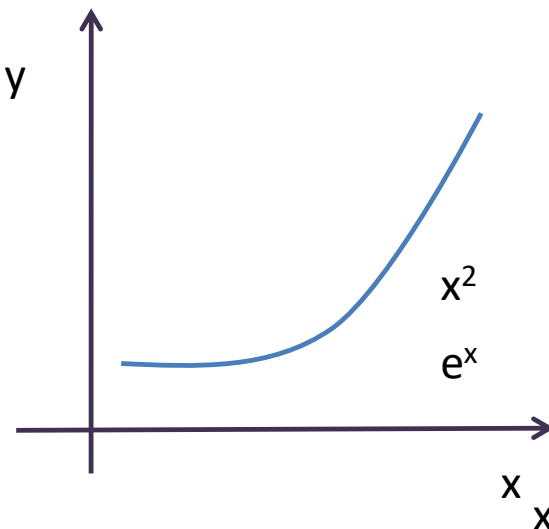
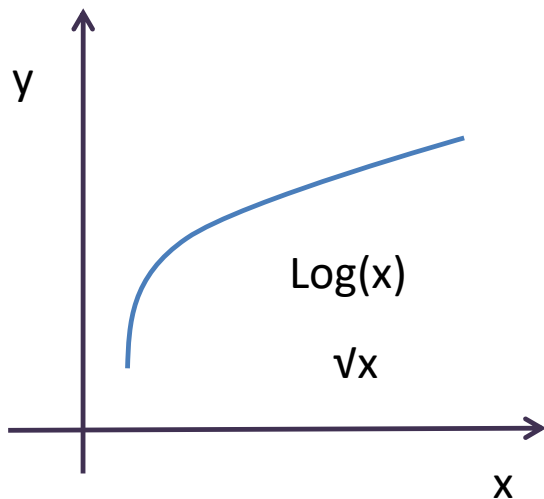
rho

0.3585263 → correlación estimada

Transformaciones de las variables

Comencemos por casos donde los **residuos** son aproximadamente **normales** y **varianza constante**.

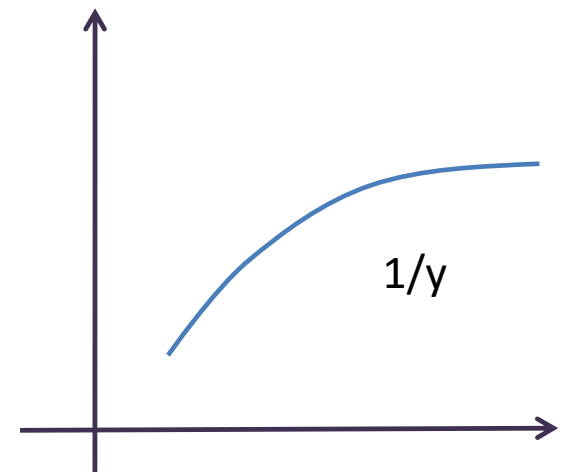
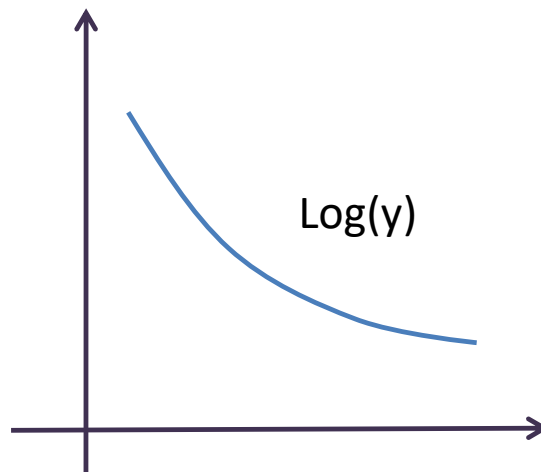
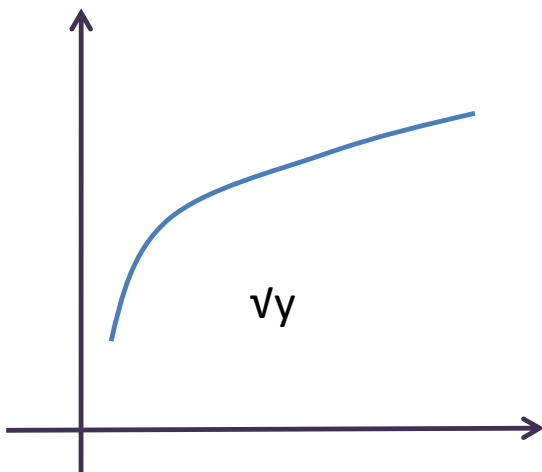
En estos casos no es recomendable transformar la y , porque pueden modificar la forma de la distribución de los errores. En el caso de regresión simple se sugiere



Transformaciones de las variables

Los casos de **residuos no normales** y **varianza no constante** suelen ocurrir simultáneamente.

En estos casos es recomendable transformar la y si bien esto apunta a modificar la forma y dispersión de los errores en ocasiones también corrige la relación curvilínea. En otras oportunidades sigue siendo necesario transformar la x .



Gráficos de residuos

Transformaciones de Box y Cox

En ocasiones es difícil determinar que transformación hay que hacer para corregir asimetrías en los términos de errores sobre la variable y . El procedimiento consiste en considerar la familia de transformaciones $y' = y^\lambda$ y proponemos el modelo

$$y^\lambda = X\beta + \epsilon,$$

Donde por ejemplo

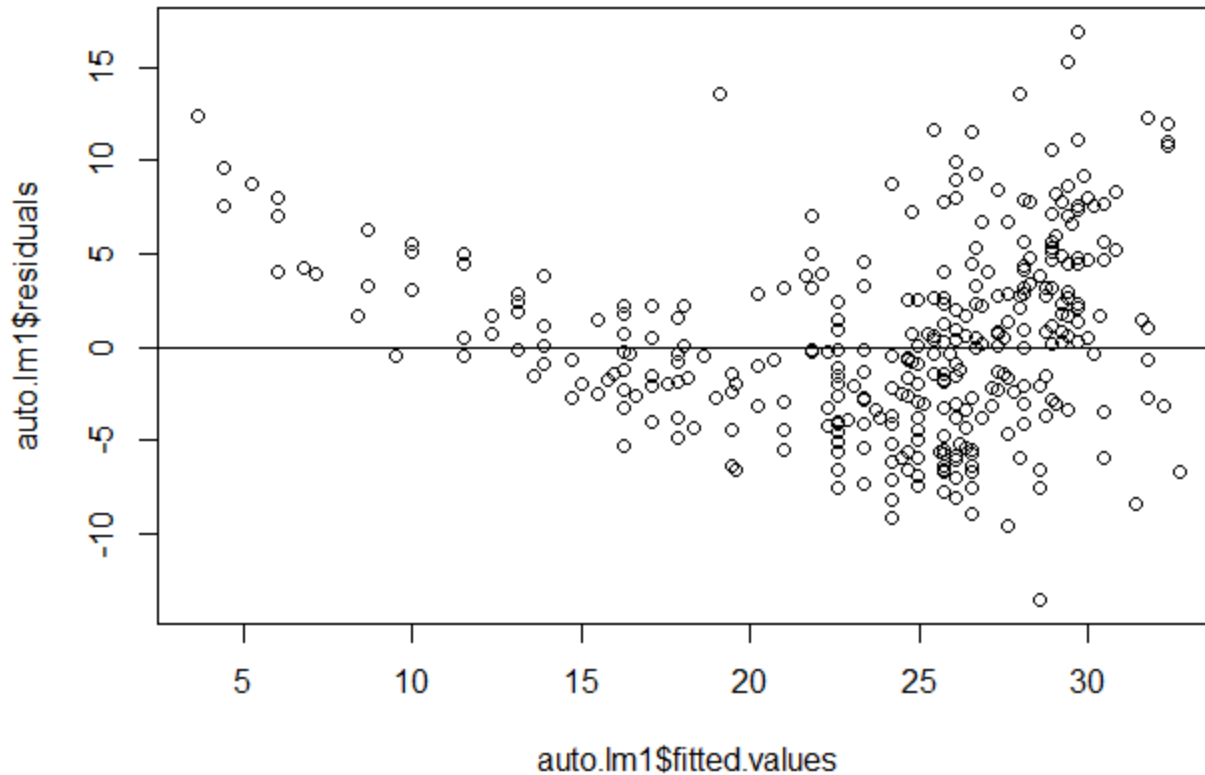
λ	y^λ
2	y^2
0.5	\sqrt{y}
0	$\ln y$
-0.5	$\frac{1}{\sqrt{y}}$
-1	$\frac{1}{y}$

El nuevo modelo tiene un parámetro mas para estimar λ , que se ajusta por máxima verosimilitud.

En general conviene utilizar parámetros que se puedan interpretar con facilidad.

Grafico los residuos vs los valores predichos para la regresión lineal de orden uno de Auto

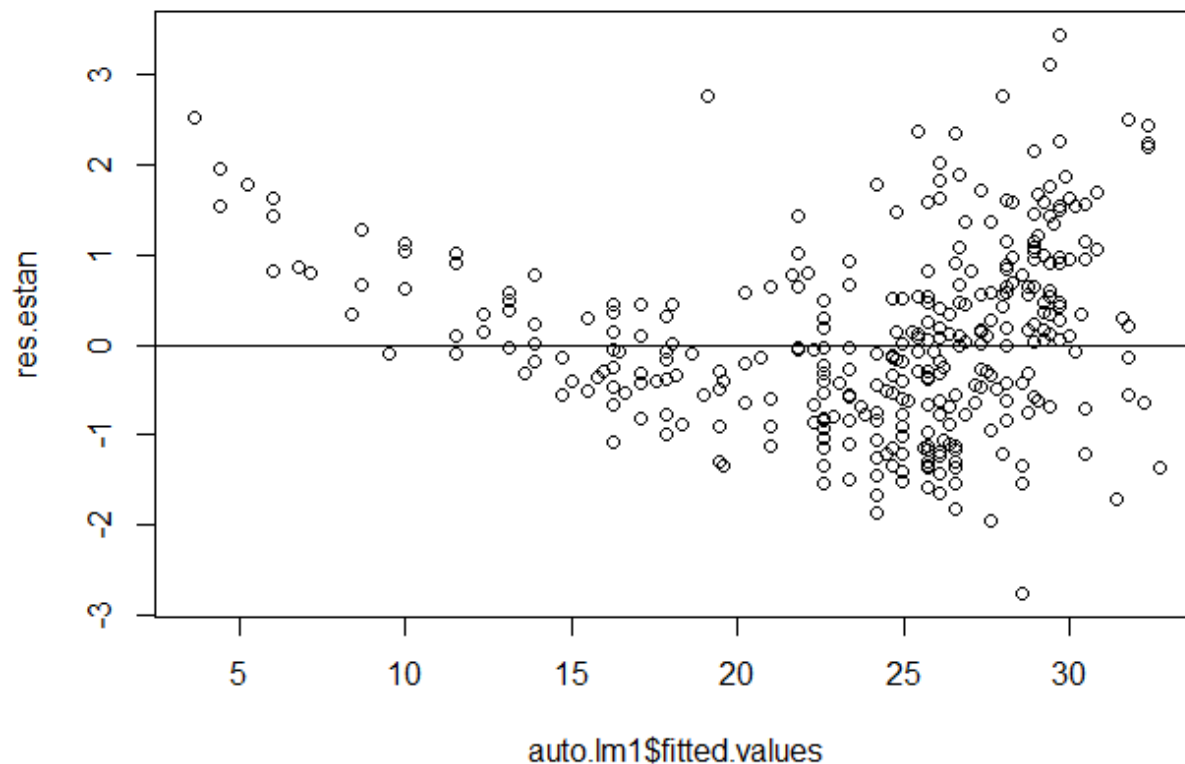
```
plot(auto.lm1$fitted.values,auto.lm1$residuals)  
abline(1,0)
```



- Relación curvilínea
- Inflación de varianza

Repito con los residuos estandarizados

```
res.estan=(auto.lm1$residuals-mean(auto.lm1$residuals))/sd(auto.lm1$residuals)  
plot(auto.lm1$fitted.values,res.estan)  
abline(0,0)
```



Solo cambia la escala
del eje y.

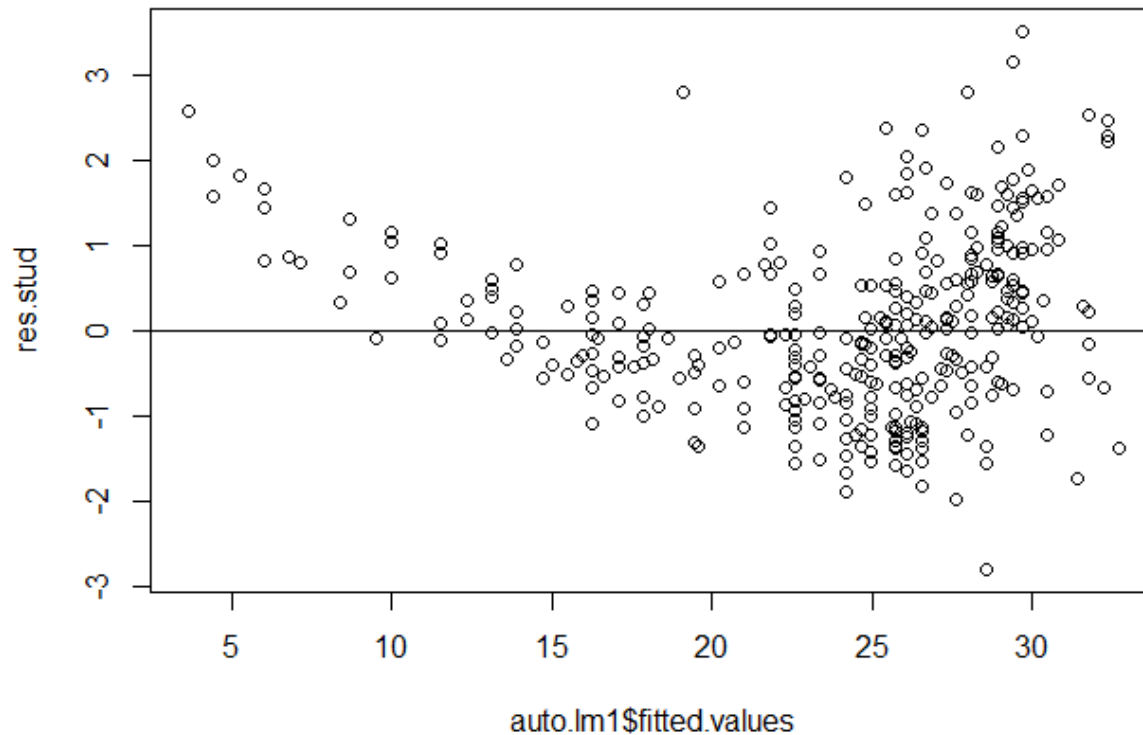
Gráfico de los residuos R-student

```
library(MASS)
```

```
res.stud=studres(auto.lm1)
```

```
plot(auto.lm1$fitted.values,res.stud)
```

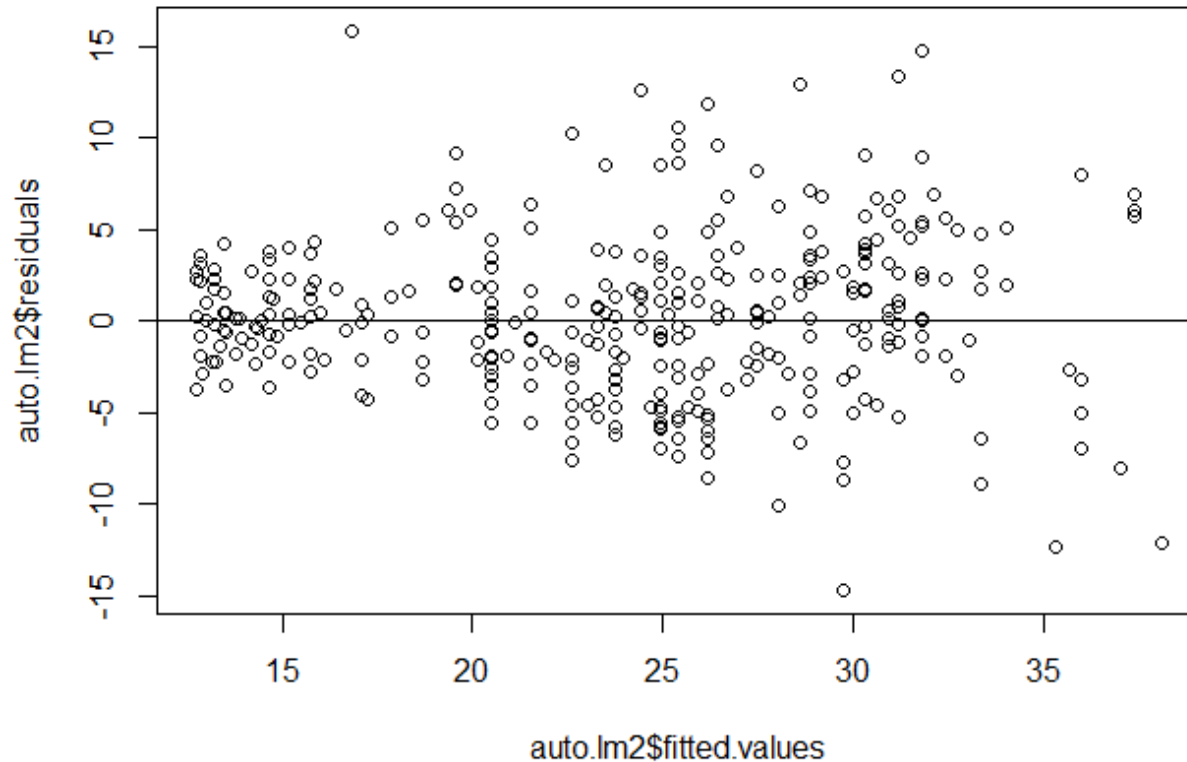
```
abline(0,0)
```



En este caso prácticamente no se ven modificaciones.

Grafico los residuos vs los valores predichos para la regresión lineal de orden dos de Auto

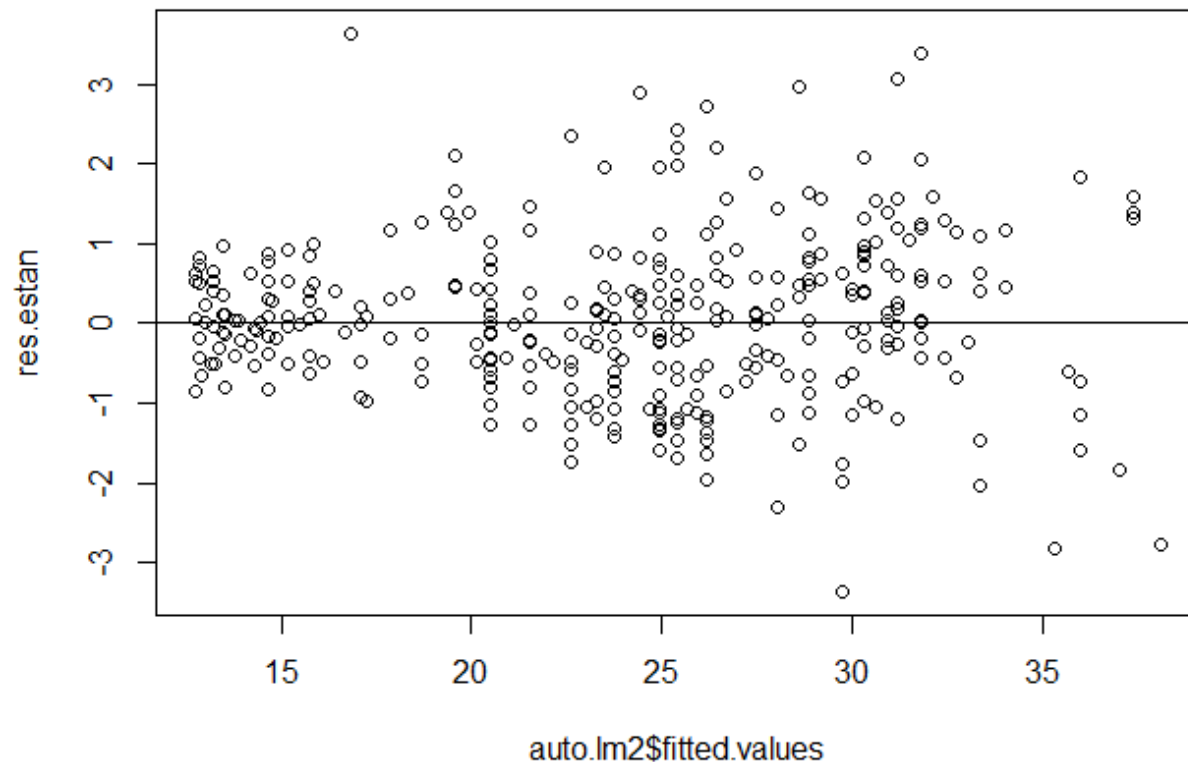
```
plot(auto.lm2$fitted.values,auto.lm2$residuals)  
abline(1,0)
```



- No hay más relación curvilínea
- Persiste la inflación de varianza

Repito con los residuos estandarizados

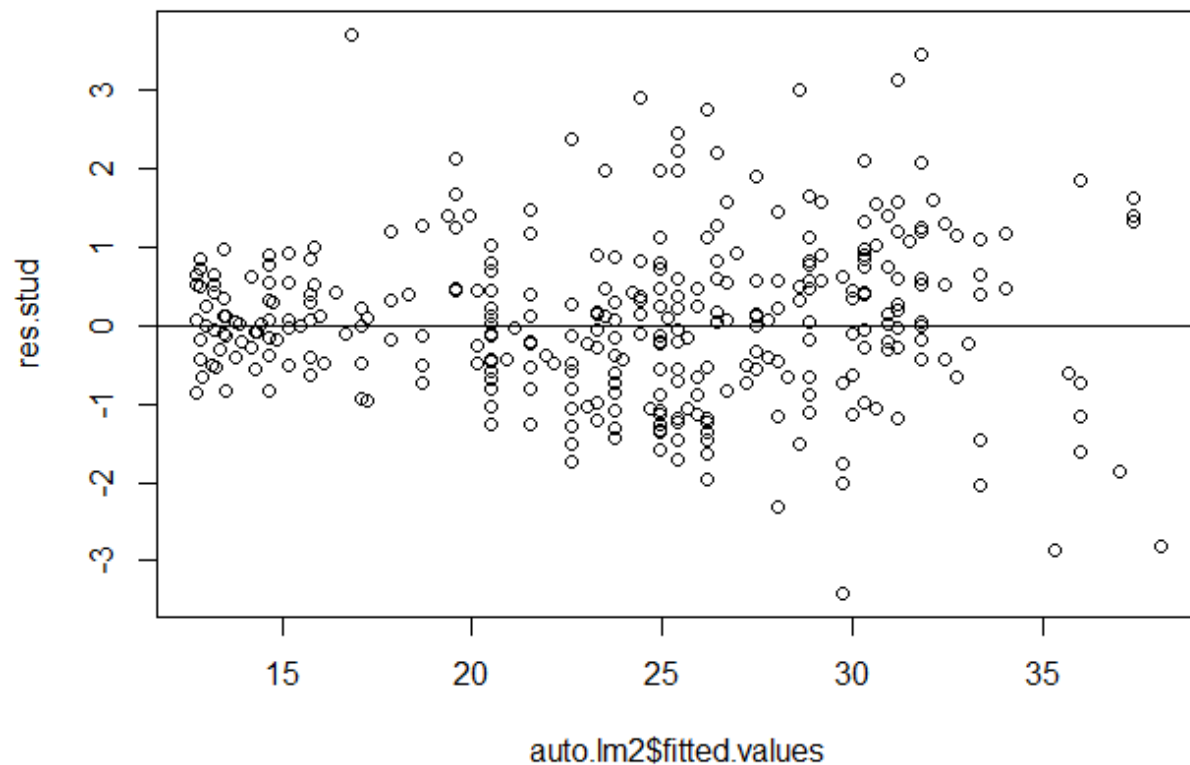
```
res.estan=(auto.lm2$residuals-mean(auto.lm2$residuals))/sd(auto.lm2$residuals)  
plot(auto.lm2$fitted.values,res.estan)  
abline(0,0)
```



Solo cambia la escala
del eje y.

R-student

```
res.stud=studres(auto.lm2)  
plot(auto.lm2$fitted.values,res.stud)  
abline(0,0)
```



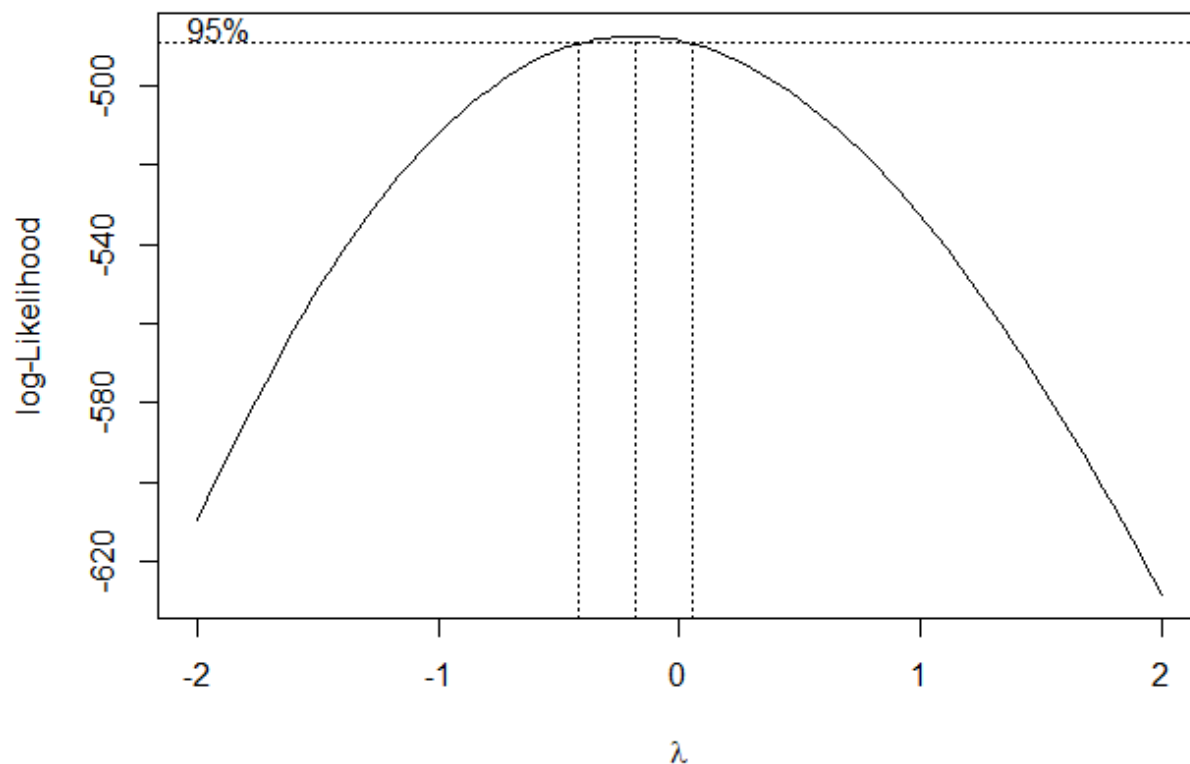
Nuevamente en este caso
no se registran cambios.

Hacemos la transformación de box y cox

```
auto.lm2.bc=boxcox(auto.lm2)
```

```
auto.lm2.bc$x[which.max(auto.lm2.bc$y)]
```

```
[1] -0.1818182
```



El estimador de máxima verosimilitud empírico de λ es -0.1818
vemos el intervalo de confianza señalado por el gráfico y preferimos un valor donde la regresión sea más sencilla de interpretar.
Luego, consideramos $\lambda = 0$, que indica considerar como variable de respuesta $\log(y)$.


```
lnmpg=log(Auto$mpg)
auto.lm3<-lm(lnmpg ~ Auto$horsepower + hp2)
summary(auto.lm3)
```

Call:

```
lm(formula = lnmpg ~ Auto$horsepower + hp2)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.66460	-0.12041	0.00316	0.11349	0.66376

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.402e+00	7.260e-02	60.639	< 2e-16

Auto\$horsepower	-1.711e-02	1.255e-03	-13.632	< 2e-16

hp2	3.901e-05	4.922e-06	7.925	2.44e-14

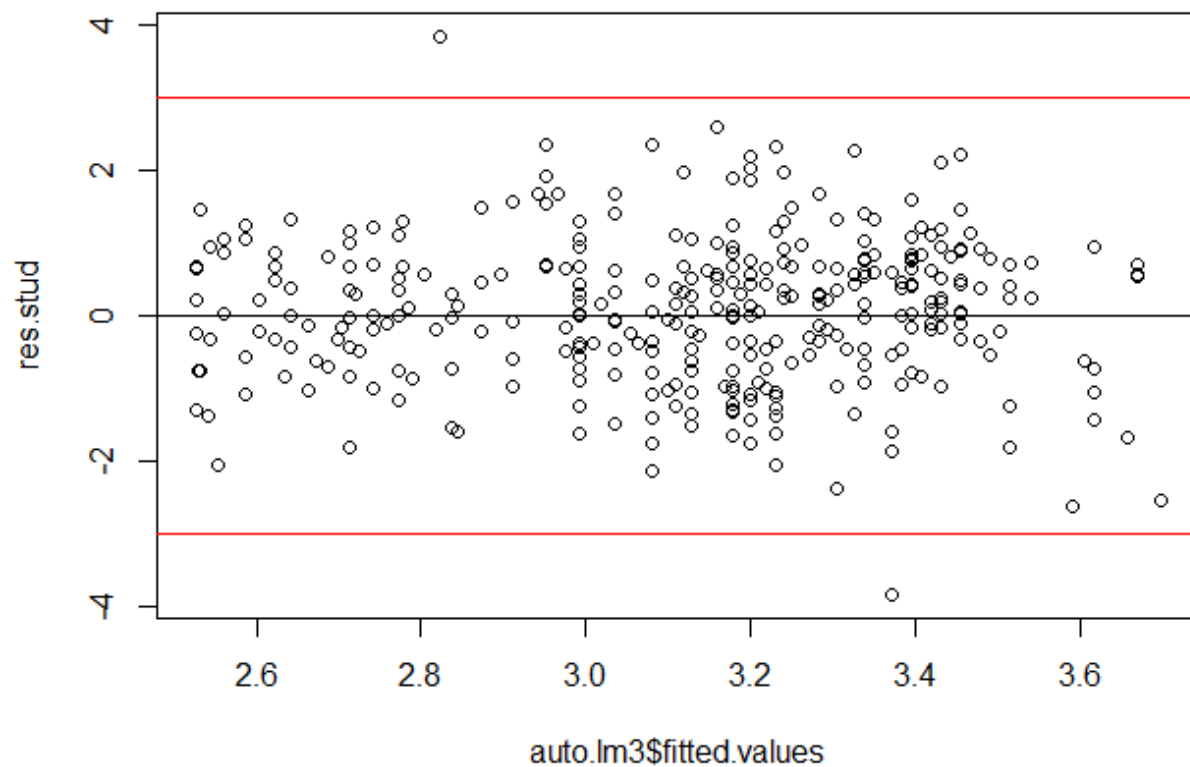
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1764 on 389 degrees of freedom

Multiple R-squared: 0.7324, Adjusted R-squared: **0.731**  Aumentó

F-statistic: 532.2 on 2 and 389 DF, p-value: < 2.2e-16

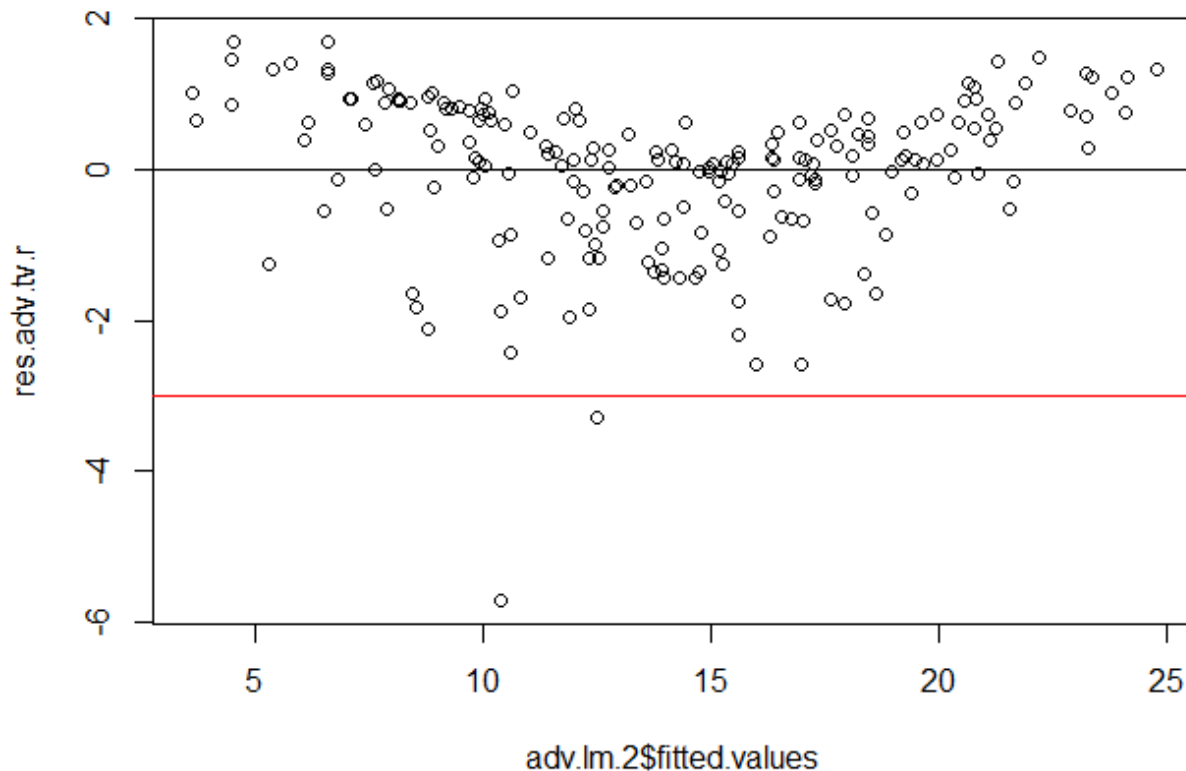
Gráfico de los residuos studentizados.



La varianza se estabilizo.
Hay dos valores atípicos.

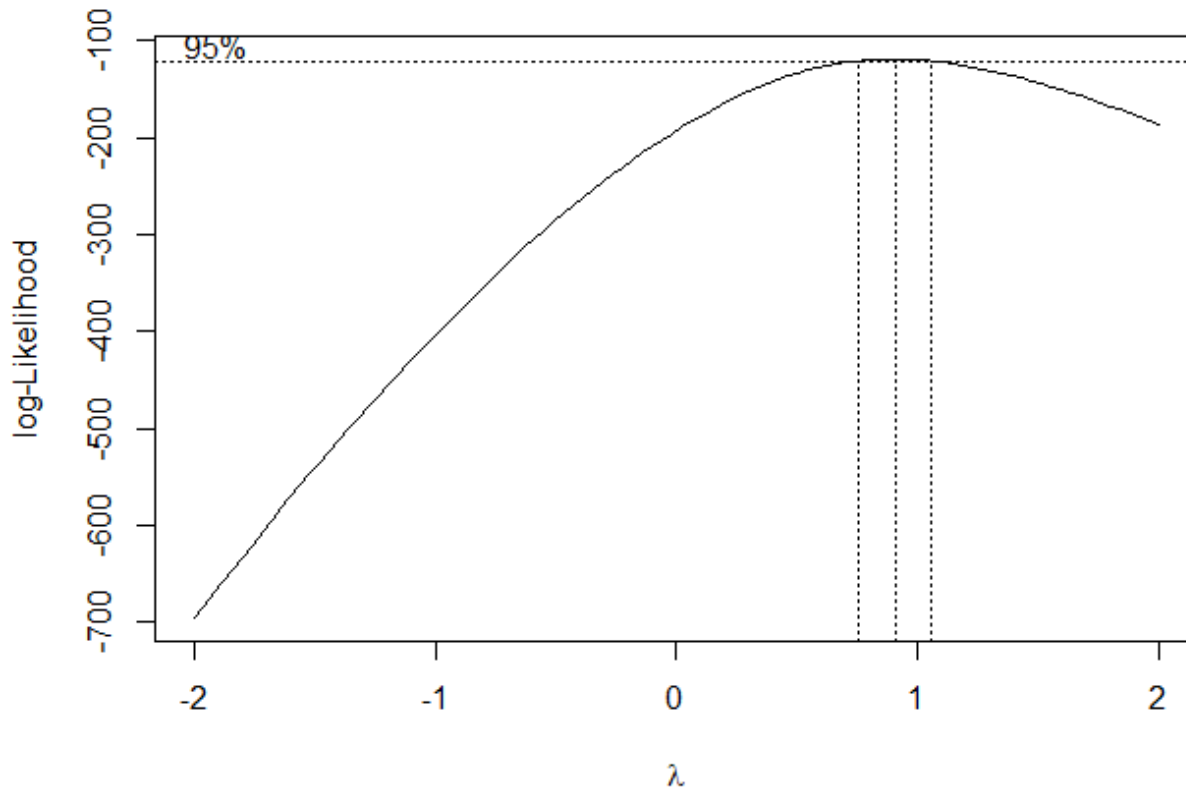
Retomamos el ejemplo de advertising.

La variable de respuesta es ventas y las regresoras son inversión publicitaria en TV y Radio
Graficamos los residuos studentizados.



- La relación entre las variables es no lineal.
- Se destaca la presencia de un outlier pronunciado.
- no esta claro que la varianza sea constante pero esto es menos pronunciado.

Hacemos la transformación de Box y Cox para el ejemplo de Advertising.



Como la varianza no varía vemos que al hacer la transformación de Box y Cox el intervalo de confianza de nivel 95% contiene al uno, esto indica que no es necesario realizar ninguna transformación en esa variable.

Gráficos de los residuos versus las variables regresoras.

En el eje Y graficar los residuos y en el eje X cada una de las X_j . La interpretación es análoga a la estudiada anteriormente:

- ▶ Si los residuos están en una banda dispersos uniformemente, no hay estructura.
- ▶ Si están en un cono, la varianza no es constante.
- ▶ Si se disponen en forma de *huevo*, hay que modelar de otro modo la relación entre y y X_j , puede ser introduciendo órdenes superiores en X_j u otras transformaciones.

Adecuación del modelo

Observaciones:

- ▶ En regresión simple este gráfico no tiene sentido \hat{y} es una función lineal de x .
- ▶ Puede ser útil graficar los residuos vs variables que podrían entrar potencialmente en el modelo. Si se observa estructura la incorporación de esta variable podría servir.
- ▶ Los gráficos de residuos parciales son mejores que estos para diagnosticar.

Transformaciones de Box y Tidwell

Estas transformaciones son análogas a las de Box y Cox pero se utilizan para detectar transformaciones en las variables regresoras. Asumimos que la variable de respuesta y está asociada a una variable regresora x a través de la transformación

$$\begin{aligned} x^\alpha & \quad \text{cuando } \alpha \neq 0 \\ \log(x) & \quad \text{cuando } \alpha = 0. \end{aligned}$$

El parámetro α se estima por máxima verosimilitud, nuevamente considerando la RSS, donde ahora los parámetros a estimar son β y α .

Me fijo en el ejemplo de Advertising.

```
library(car)
```

```
rad=radio+0.001
```

Las variables tienen que ser positivas como radio tiene
El mínimo en cero lo desplazo un poquito

```
boxTidwell(sales ~ TV + rad)
```

	MLE of lambda	Score	Statistic (z)	Pr(> z)
TV	0.35151	-8.5182	<2e-16	***
rad	1.15019	1.5904	0.1117	

Sugiere transformar TV,
la transformación más
cercana fácil de interpretar
es tomar raíz cuadrada

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

Coeficiente cercano
a uno, sugiere no
transformar

```
iterations = 4
```

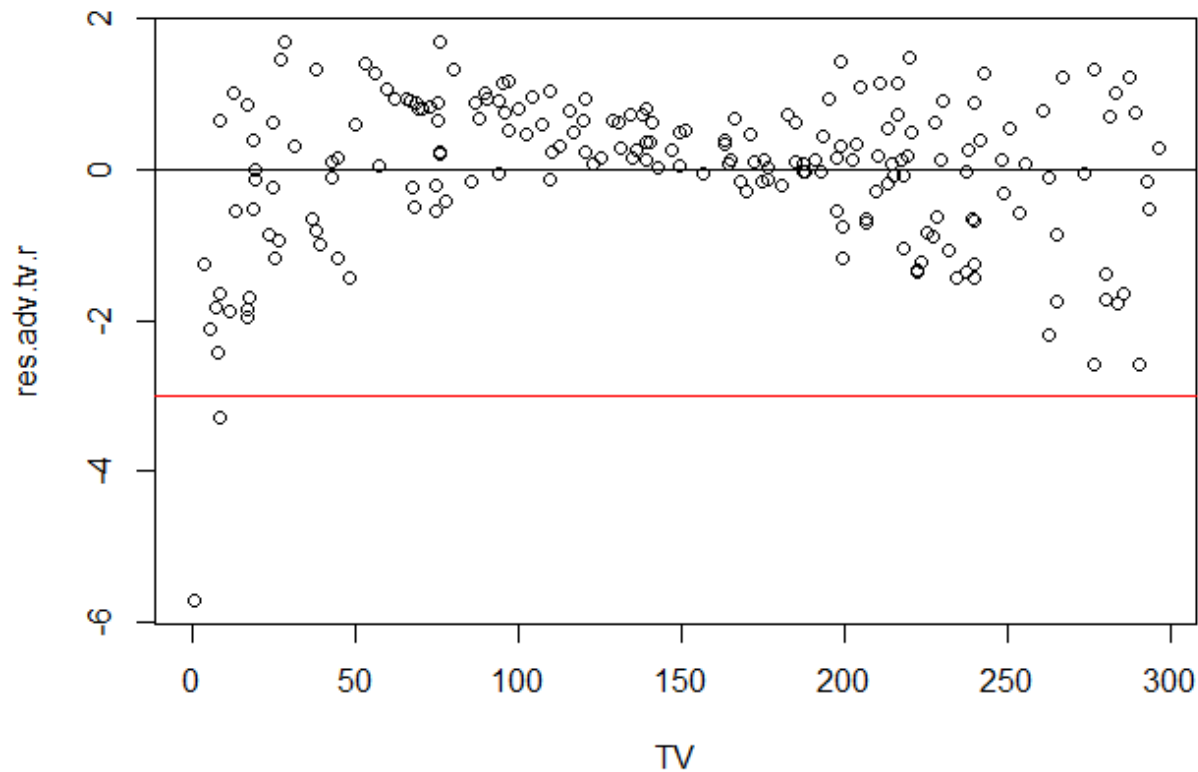
Una aclaración importante es que esta salida no da un intervalo de confianza a diferencia de las transformaciones de Box y Cox donde si tenemos esos datos.

Gráfico de los residuos studentizados versus la inversión publicitaria en TV, ej Advertising.

```
plot(TV,res.adv.tv.r)
```

```
abline(0,0)
```

```
abline(-3,0,col="red")
```



Se observa estructura.

La varianza en la región central es menor y en general los residuos positivos.

En las inversiones bajas y alta, hay mas dispersión y los residuos negativos predominan Levemente.

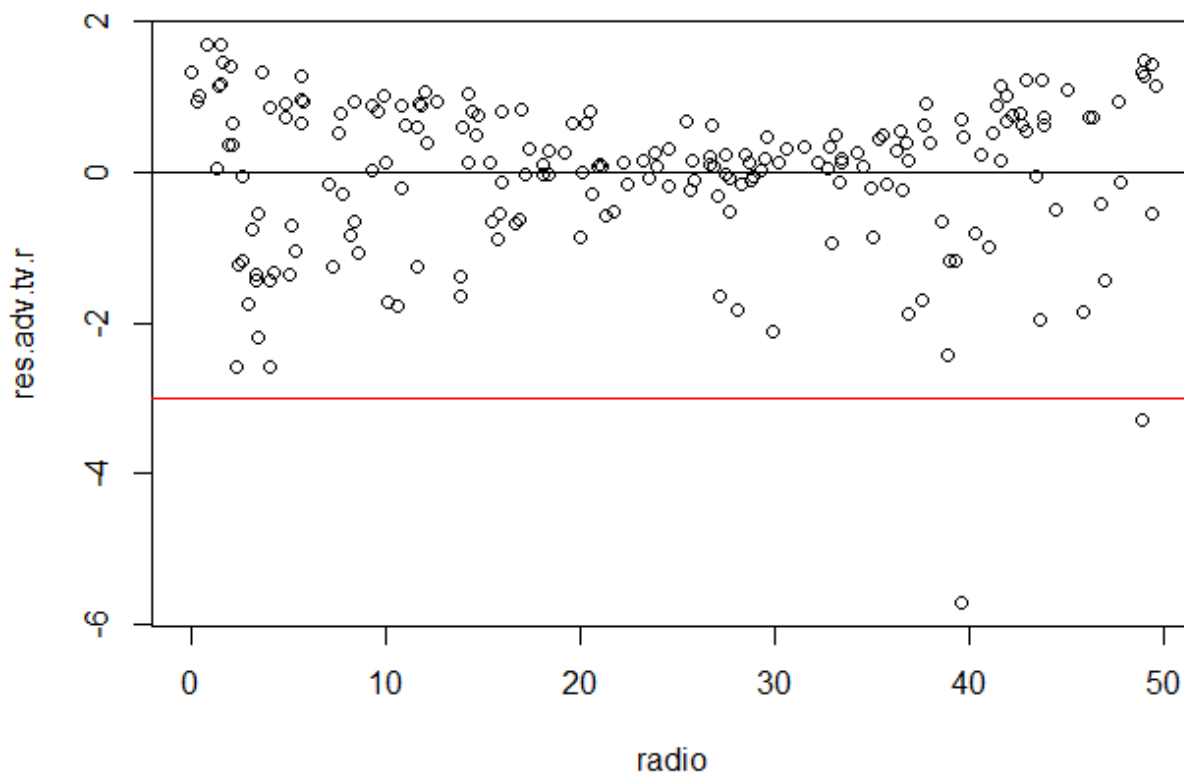
Se siguen observando los atípicos.

Repetimos el gráfico para la radio

```
plot(radio,res.adv.tv.r)
```

```
abline(0,0)
```

```
abline(-3,0,col="red")
```



Se observa una estructura similar aunque menos pronunciada.

Se siguen viendo dos atípicos, Habría que verificar si son los mismos.

```
sqTV=sqrt(TV)
adv.lm.3<-lm(sales~sqTV+radio)
summary(adv.lm.3)
```

```
Call:
lm(formula = sales ~ sqTV + radio)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.2997	-0.8514	0.0371	0.8599	3.4128

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.617931	0.325651	-4.968	1.46e-06 ***
sqTV	0.974854	0.023962	40.683	< 2e-16 ***
radio	0.194496	0.006677	29.131	< 2e-16 ***

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

Residual standard error: 1.398 on 197 degrees of freedom

Multiple R-squared: 0.929, Adjusted R-squared:
0.9282

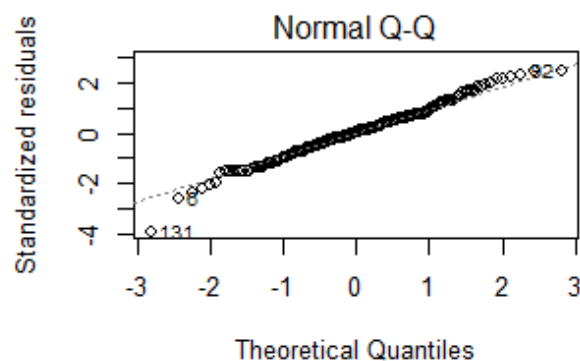
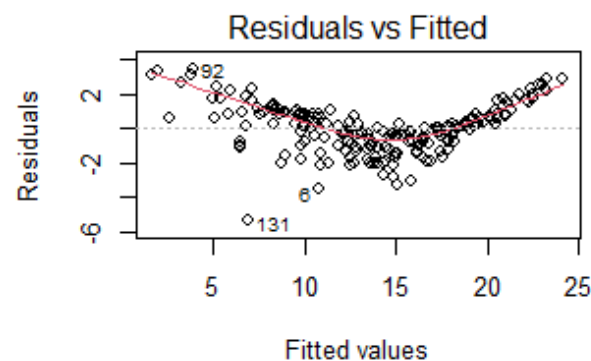
F-statistic: 1288 on 2 and 197 DF, p-value: < 2.2e-16

Este modelo es más complejo, la cantidad de variables es la misma pero al haber transformado más es difícil de analizar.

El R² aumenta.

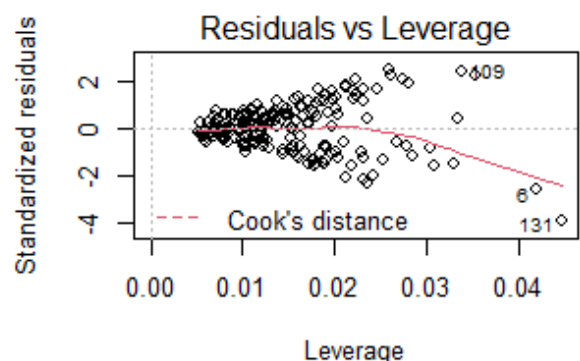
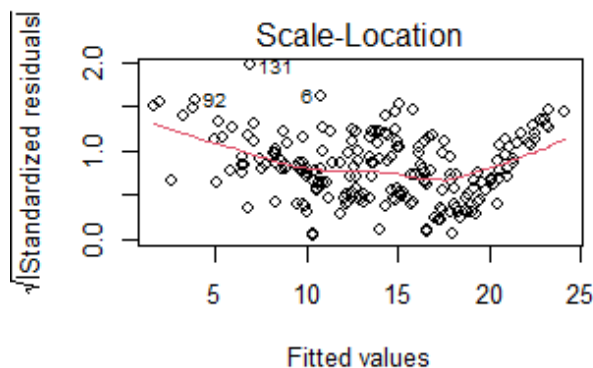
sin embargo como veremos en los gráficos la estructura en los residuos persiste.

Más adelante veremos otro enfoque.

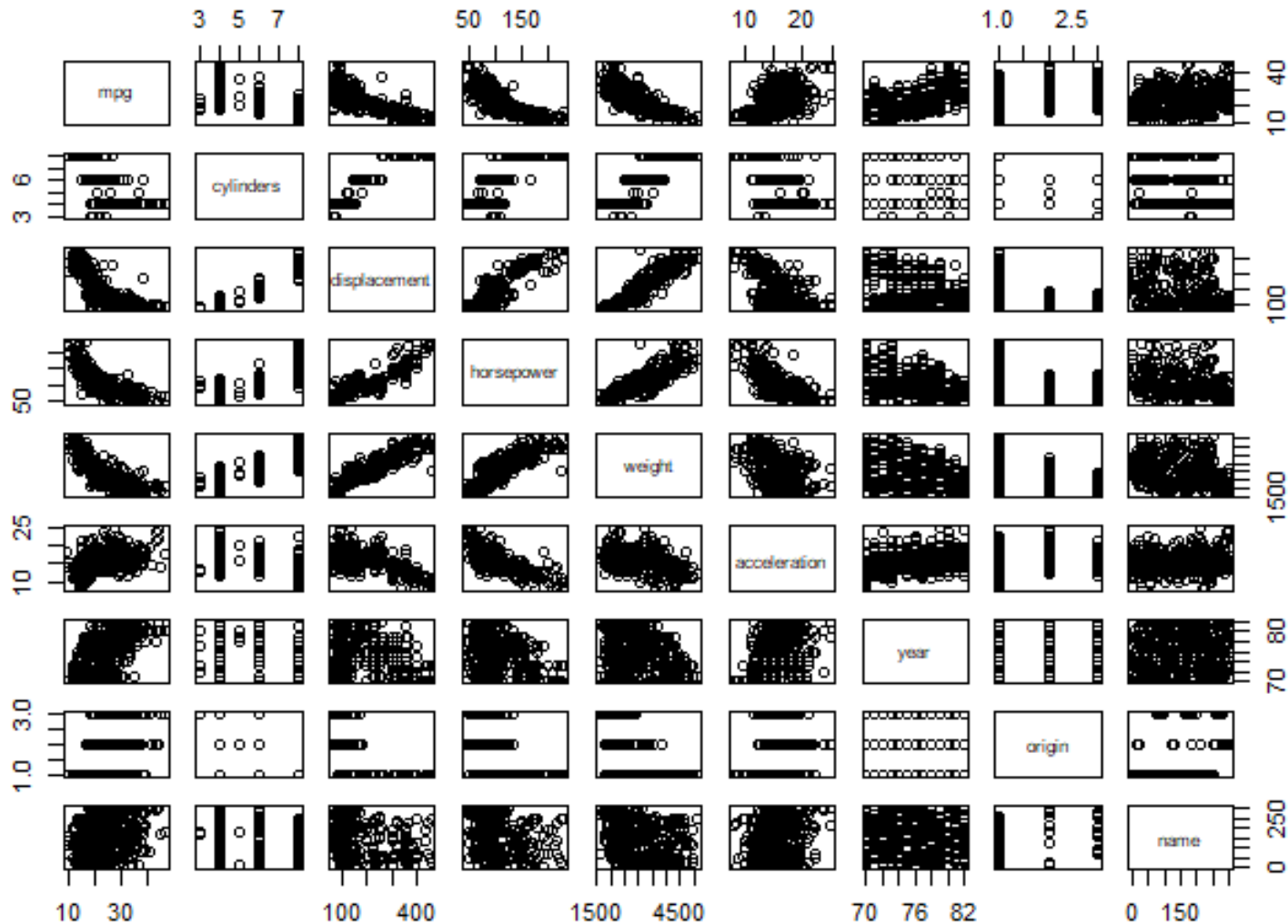


Los gráficos de los residuos no mejoraron.

Probé muchas otras transformaciones y todas eran peores.



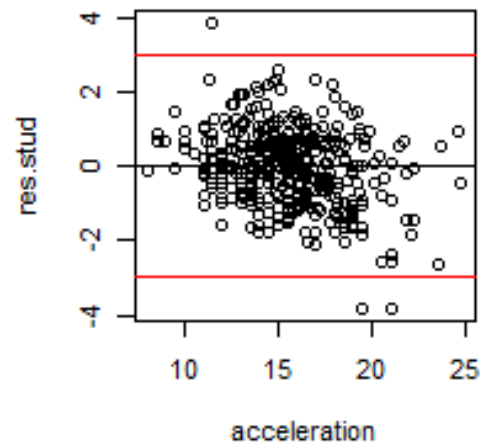
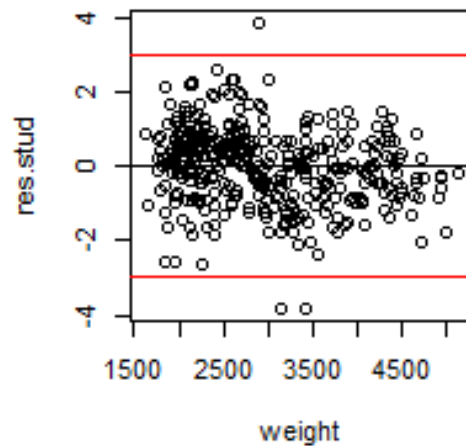
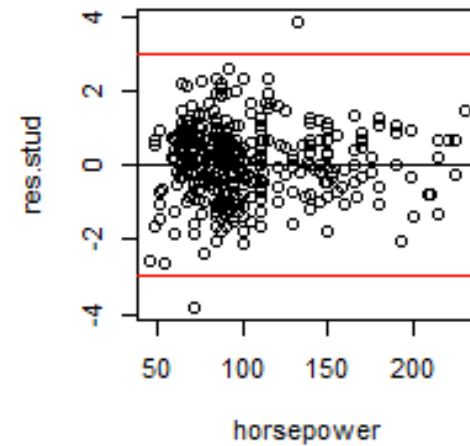
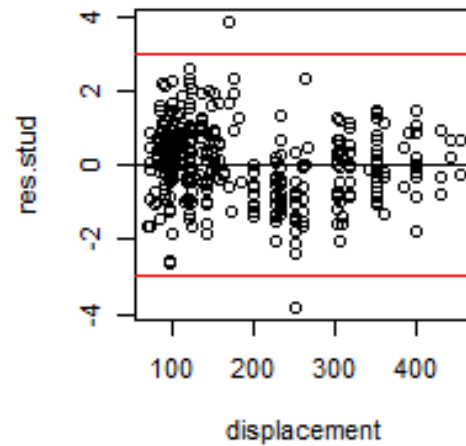
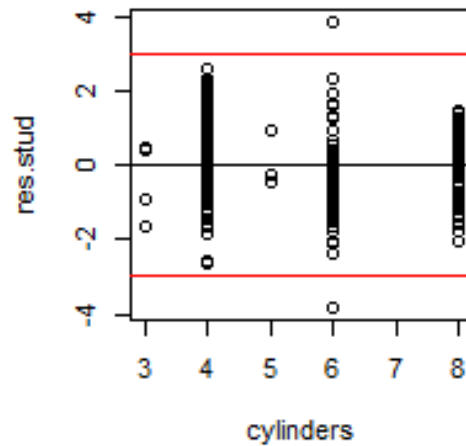
Volvemos al ejemplo de Auto
`pairs(Auto)`



Hay variables que parece tener información similar a horsepower, por ejemplo displacement, weight y acceleration, tienen correlación alta.

Otras que no parecen estar relacionada

Gráfico de los residuos estudentizados vs otras variables regresoras.



La variable *weight* es la que parece tener algo de estructura.

```
auto.lm4<-lm(lnmpg ~ Auto$horsepower + hp2+ weight)
summary(auto.lm4)
```

Call:

```
lm(formula = lnmpg ~ Auto$horsepower + hp2 + weight)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.47181	-0.09256	0.00515	0.10195	0.49862

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.358e+00	6.329e-02	68.849	< 2e-16 ***
Auto\$horsepower	-8.196e-03	1.350e-03	-6.069	3.07e-09 ***
hp2	2.012e-05	4.602e-06	4.372	1.58e-05 ***
weight	-2.191e-04	1.953e-05	-11.217	< 2e-16 ***

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

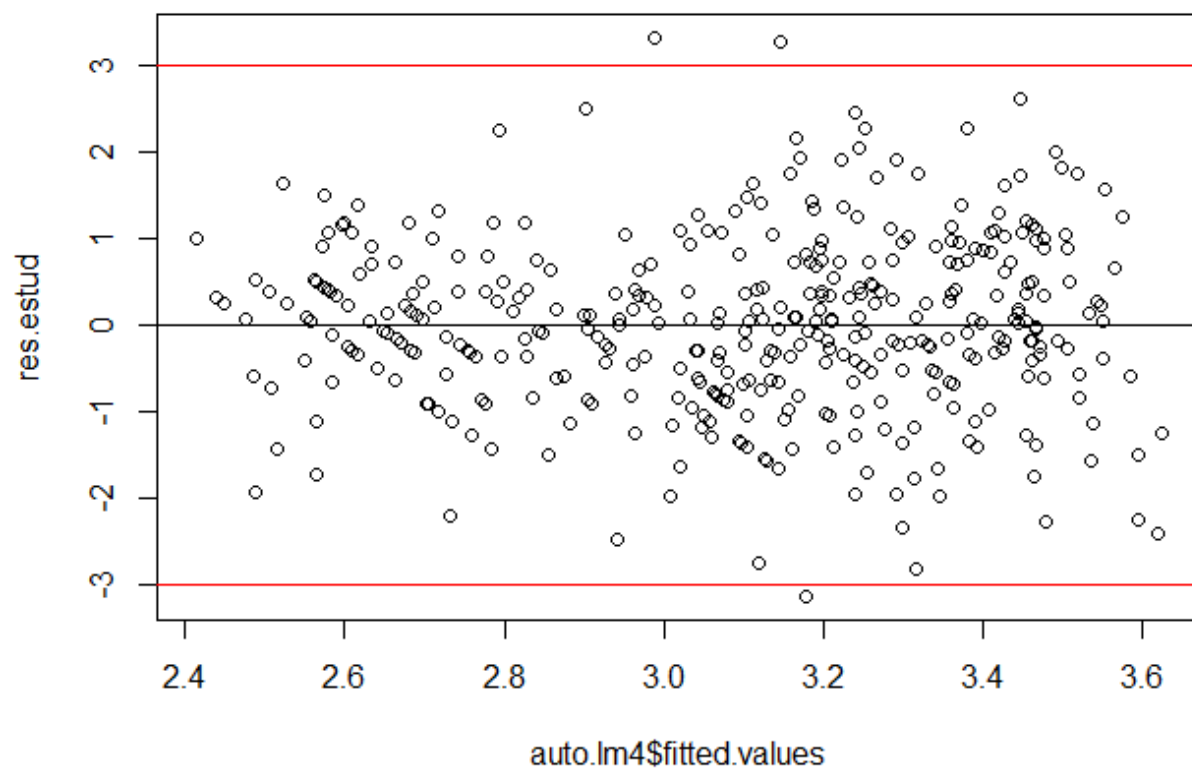
Residual standard error: 0.1535 on 388 degrees of freedom

Multiple R-squared: 0.7979, Adjusted R-squared: **0.7963**

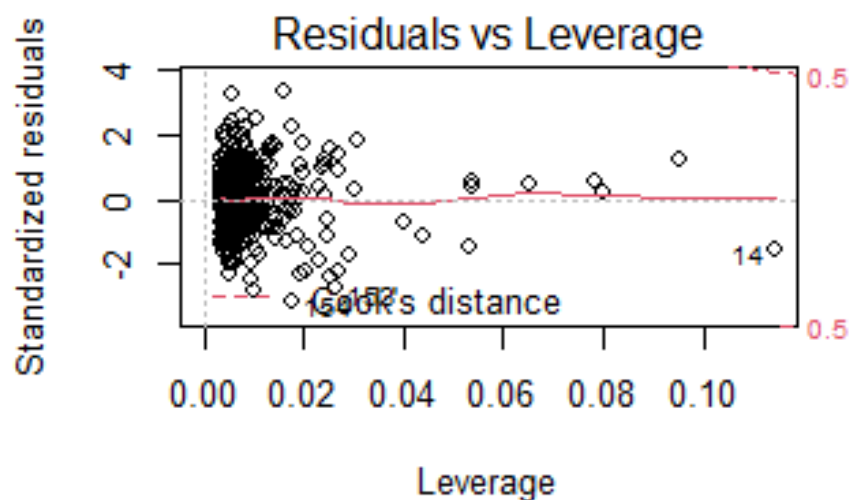
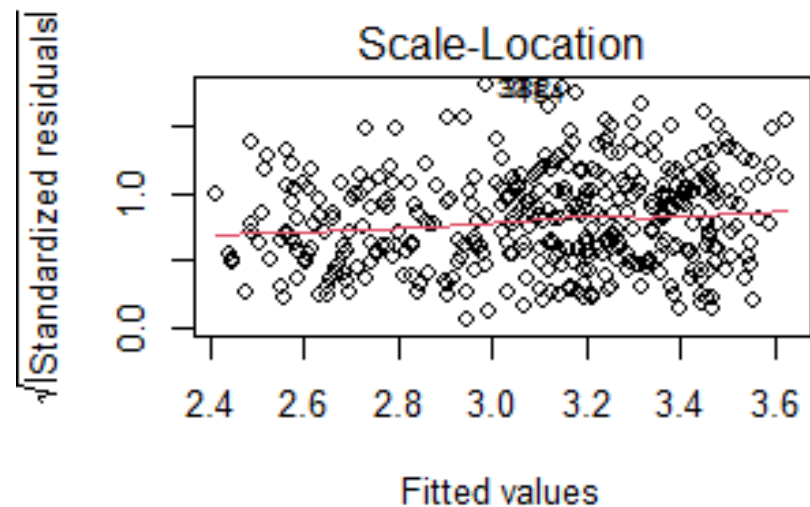
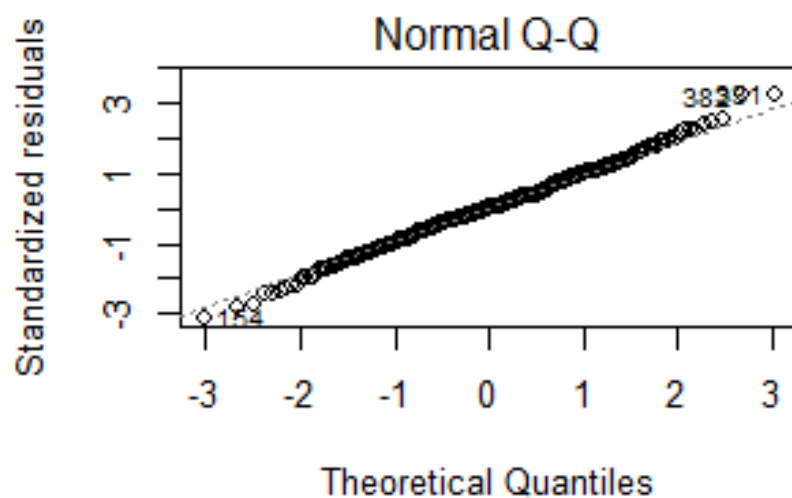
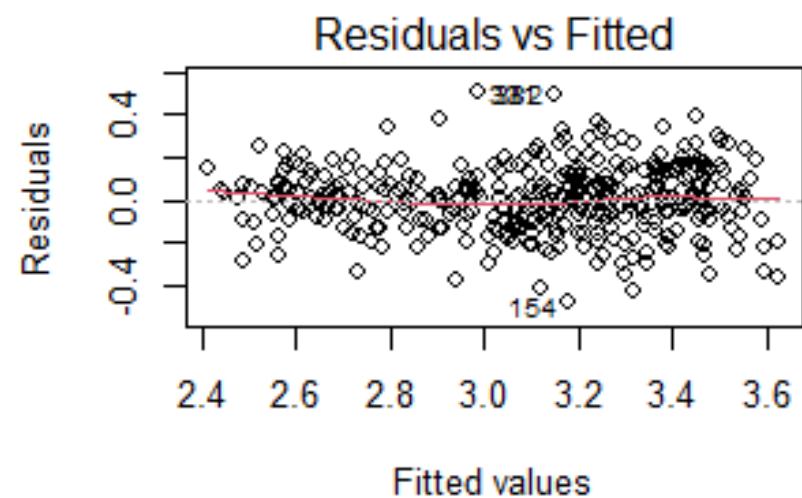
F-statistic: 510.6 on 3 and 388 DF, p-value: < 2.2e-16

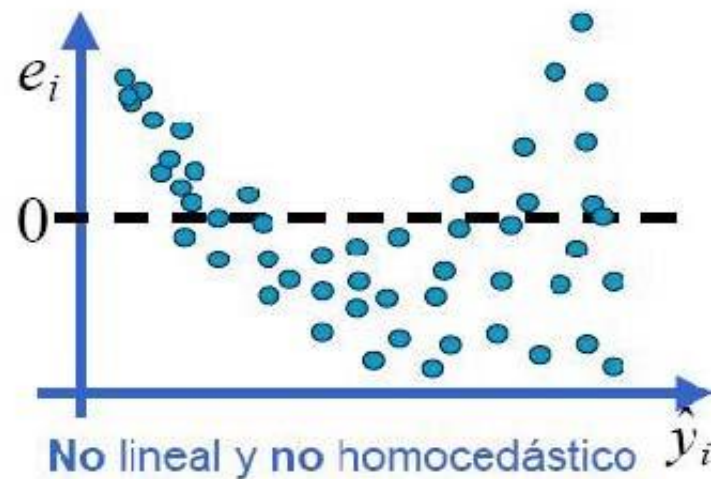
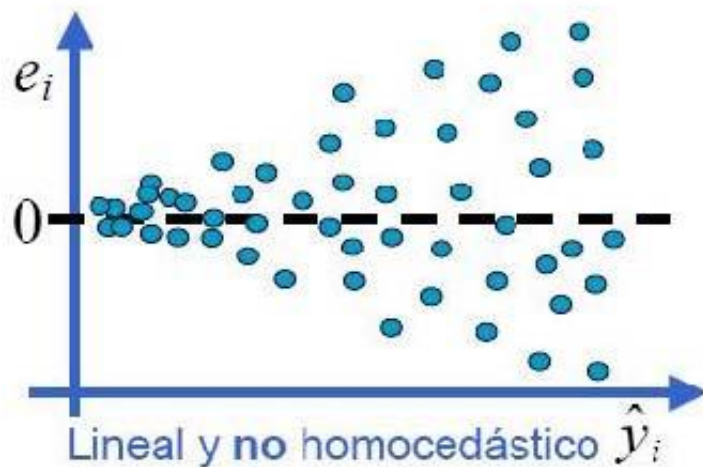
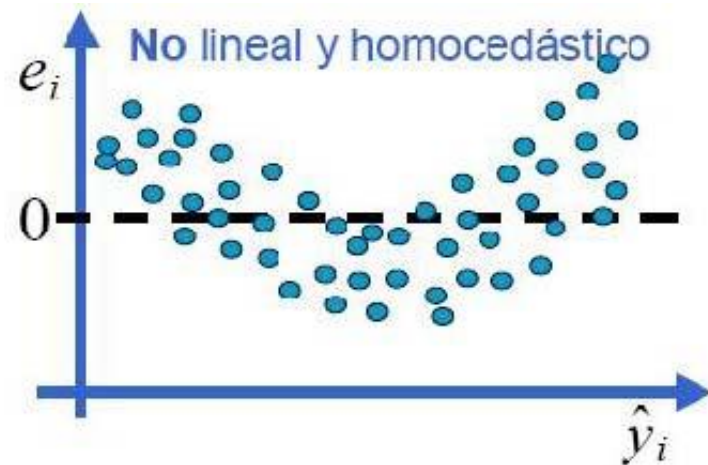
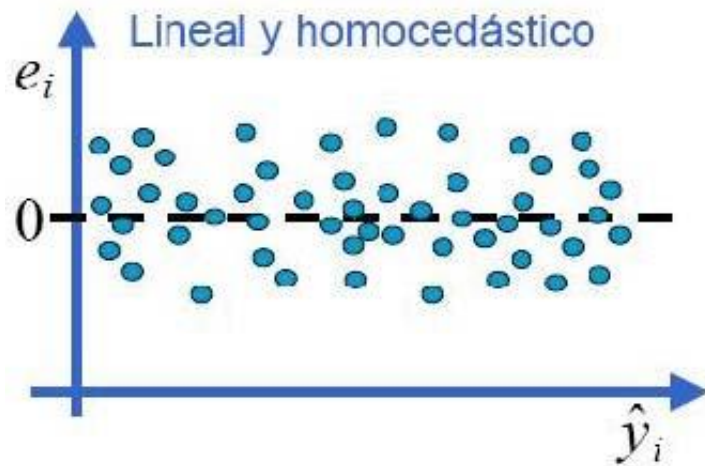
Aumentó
de 0.731 a 0.7963, es
decir.
 $(0.7963 - 0.731) / (1 - 0.731) =$
0.24,
Explica un 24% de la
varianza que faltaba
explicar.

Gráfico de los residuos studentizados



Se puede ver que ya no
Hay estructura y que los
outliers son moderados





Gráficos de los residuos en función del tiempo.

En general la idea en estos gráficos es graficar el r_i versus el índice i .

Cúal es la consecuencia de la presencia de correlación entre los errores?

Subestimar los desvíos estándares de los estimadores de los parámetros. Luego intervalos de predicción y confianza serán menores. También los p-valores correspondientes. Tal vez consideremos que determinados coeficientes son significativos cuando no lo son.

Adecuación del modelo

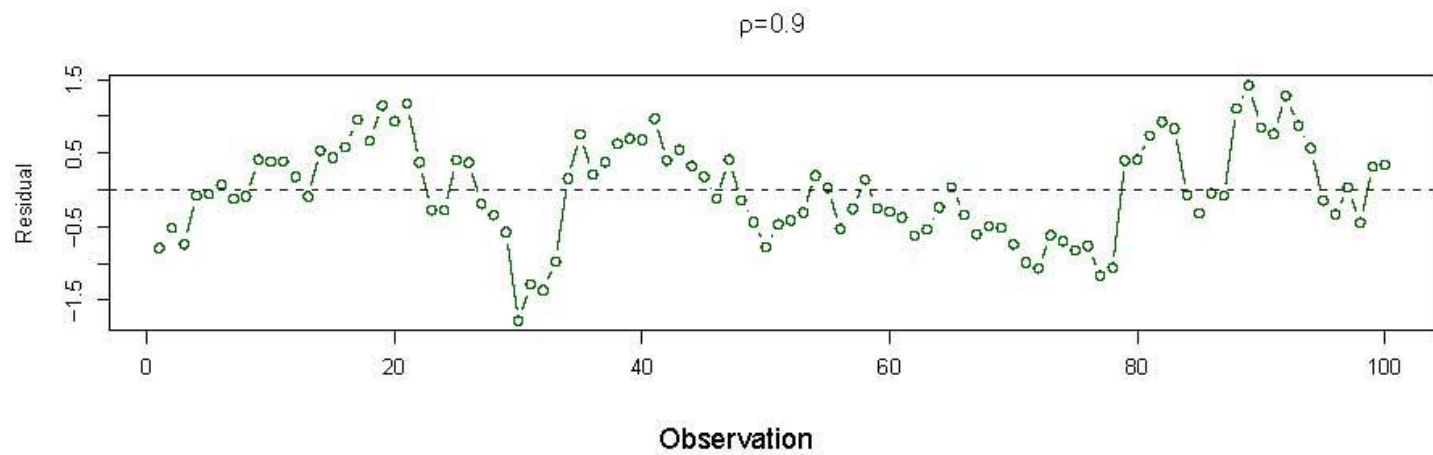
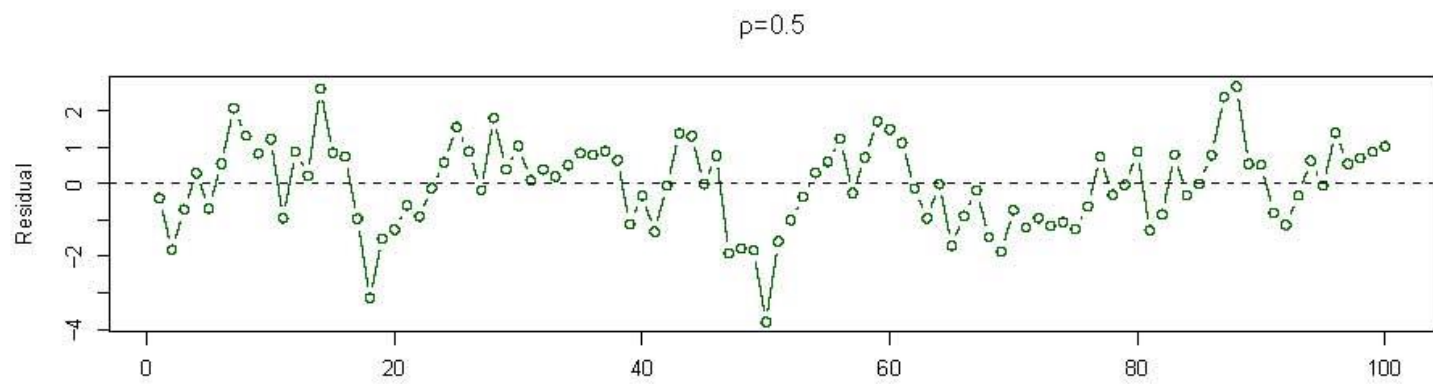
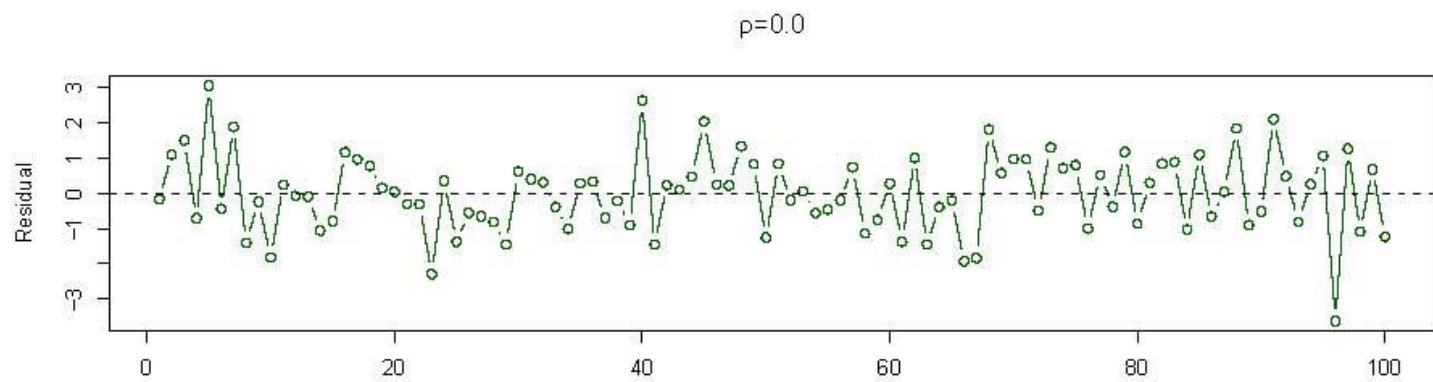
Dónde suelen aparecer estos problemas?

- ▶ Series de tiempo (si tenemos tiempo diremos algo más adelante).
- ▶ Mal diseño de experimentos. Supongamos que se quiere predecir la altura de individuos en base a su peso. El supuesto de independencia se viola si los individuos siguen la misma dieta, pertenecen a la misma familia o están expuestos a los mismos factores ambientales.

Adecuación del modelo

Observaciones:

- ▶ Si los residuos están en una banda dispersos uniformemente, no hay estructura.
- ▶ Si están en un cono, la varianza no es constante y varía con el tiempo.
- ▶ Si se disponen en forma de *huevo*, la relación es no lineal. Hay que incorporar términos lineales o cuadráticos en el tiempo.



Test de Durbin-Watson.

Este test busca establecer si los residuos de una regresión son independientes, específicamente si $E(\epsilon_i \epsilon_{i-1}) = 0$.

Este test se basa en suponer que los errores son autorregresivos de orden 1, es decir,

$$\epsilon_i = \rho \epsilon_{i-1} + \eta_i,$$

donde $\epsilon_i, \epsilon_{i-1}$ son los errores de las observaciones i e $i - 1$ respectivamente y η_i es una variable aleatoria $\eta_i \sim N(0, \sigma_\eta^2)$ y ρ es el coeficiente de autocorrelación. Si $\rho = 0$ entonces los errores (ϵ) son independientes.

Luego,

$$H_0 : \rho = 0 \text{ vs } H_A : \rho \neq 0.$$

El estadístico del test es,

$$d = \frac{\sum_{i=2}^n (\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^n (\epsilon_i)^2},$$

se rechaza H_0 si $d < d_\alpha$.

Veamos si hay correlación lineal en el último modelo del ejemplo Advertising.

```
library(lmtest)
```

```
dwtest(adv.lm.3)
```

Durbin-Watson test

data: adv.lm.3

DW = 2.1294, p-value = **0.8216** → No hay correlación lineal

alternative hypothesis: true autocorrelation is greater than 0

Adecuación del modelo

Observaciones:

- ▶ Hay que ser cautelosos a la hora de utilizar estos gráficos ya que sólo sugieren una posible relación entre la variable de estudio y el resto de las variables regresoras.
- ▶ No detectan interacciones entre variables.
- ▶ Los residuos parciales suelen estar afectados por problemas de multicolinealidad.

Detección de outliers

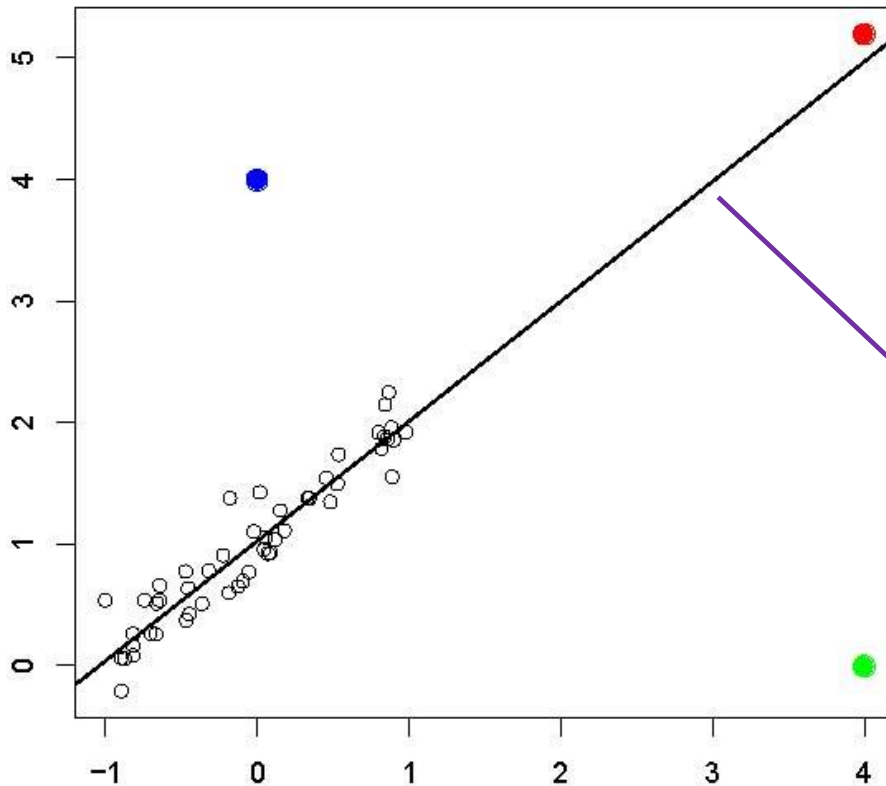
- ▶ Los outliers son puntos extremos.
- ▶ Para identificarlos usaremos los residuos R-student o el leverage.
- ▶ Dependiendo de la posición pueden tener efecto severo o moderado en la regresión.
- ▶ Los gráficos de los residuos y de normalidad ayudan a detectarlos, si la contaminación está en la coordenada y .
- ▶ Si la contaminación es en la coordenada X también pueden tener efecto en la regresión.
- ▶ En muchas ocasiones los outliers son producto de datos mal medidos o cargados, en esos casos hay que descartarlos.
- ▶ En otros casos, son casos que ocurren con probabilidad baja, quitarlos sería un error.

Adecuación del modelo

Detección de outliers

- ▶ Los outliers al afectar la estimación de los parámetros afectan los test t , F , el coeficiente de determinación y los intervalos de confianza y predicción.
- ▶ En regresión alta son difíciles de identificar, sobre todo si son varios, enmascaramiento.
- ▶ Para detectar un outlier podemos estudiar los residuos R -student y señalar como potencial outlier aquel que en módulo sea mayor a 3.

Adecuación del modelo



Hay 3 observaciones alejadas de la nube de puntos:

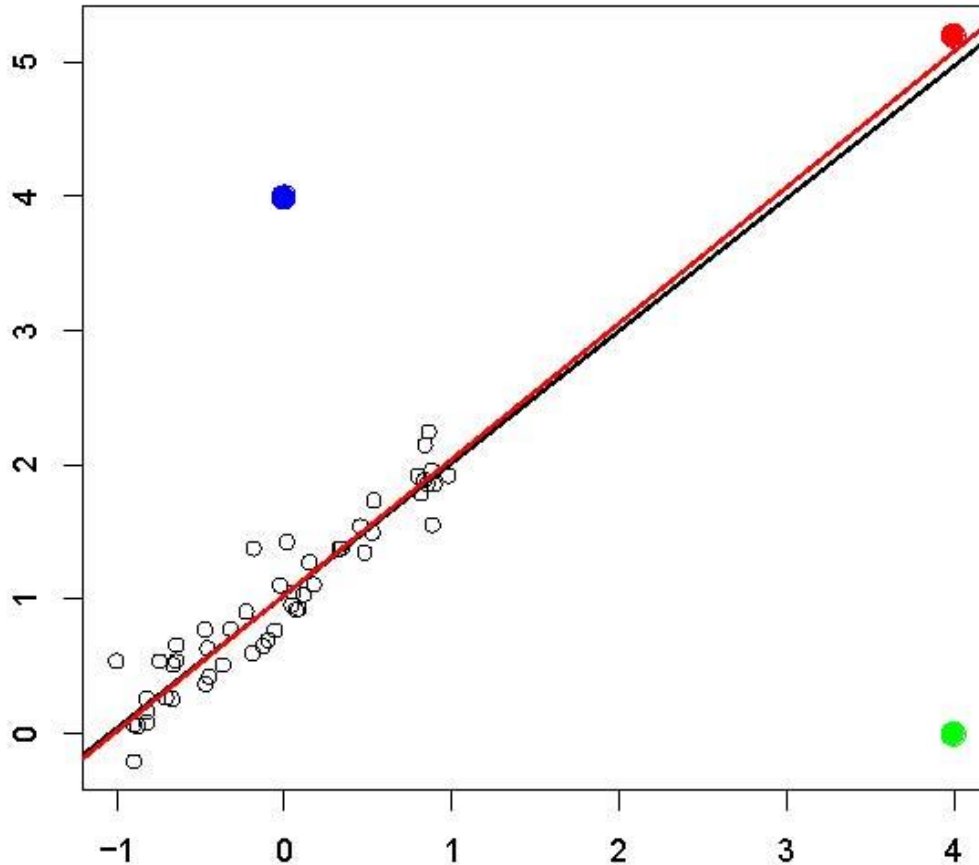
•Azul alejada en y

•Roja alejada en x

•Verde alejada en x e y

La recta negra es la recta con la que fueron generados los datos.

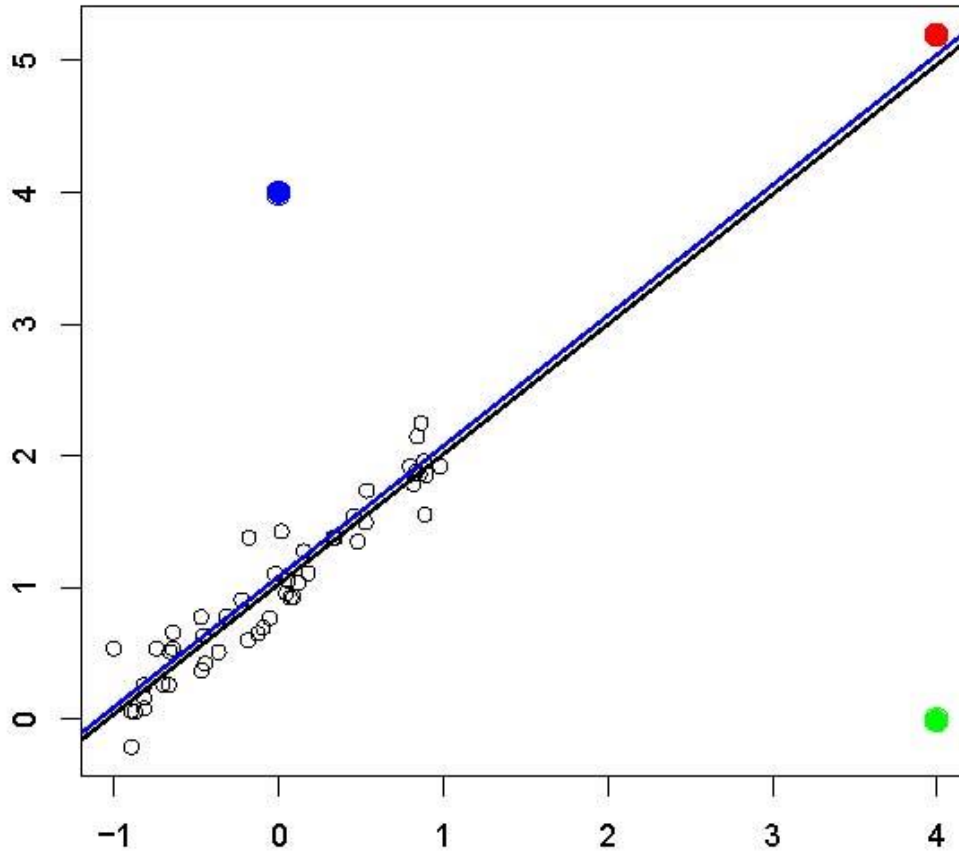
Adecuación del modelo



Recta de regresión con datos negros y rojo.

Prácticamente no afecta al modelo, es una observación muy influyente.

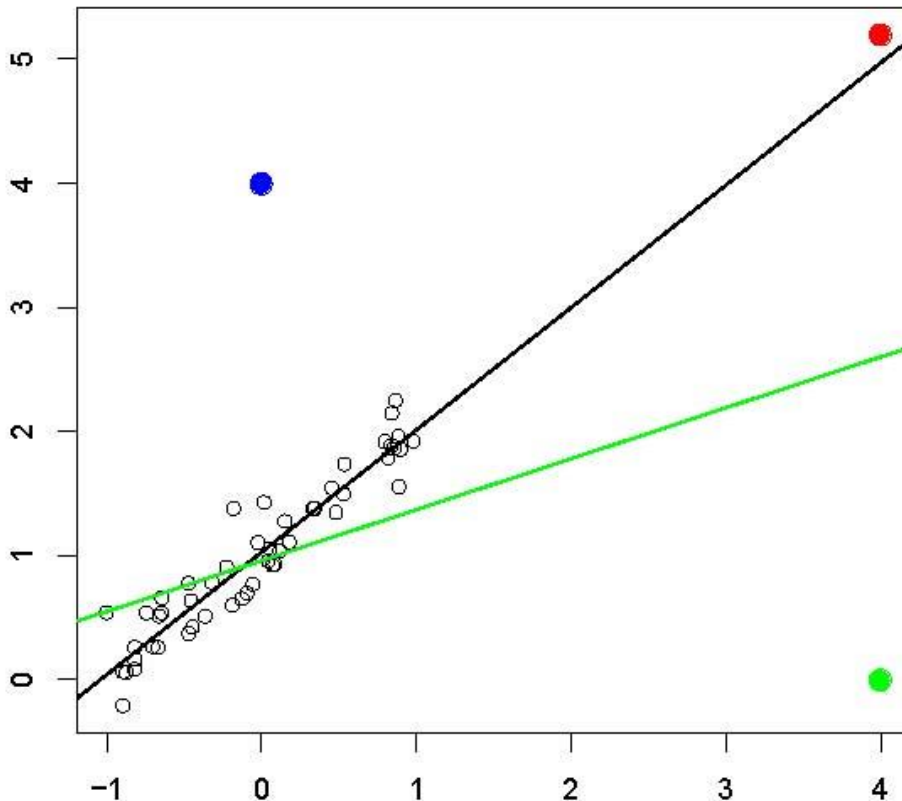
Adecuación del modelo



Recta de regresión con datos negros y azul.

Prácticamente no afecta al modelo, tiene residuo alto

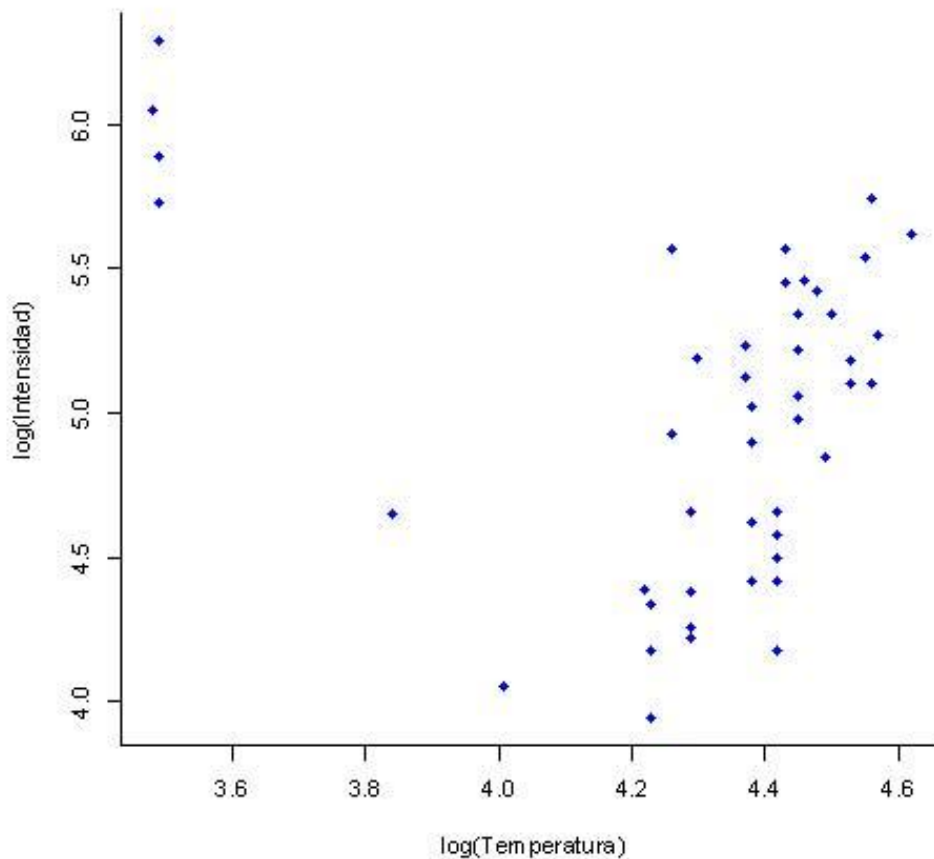
Adecuación del modelo



Recta de regresión con datos negros y verde.

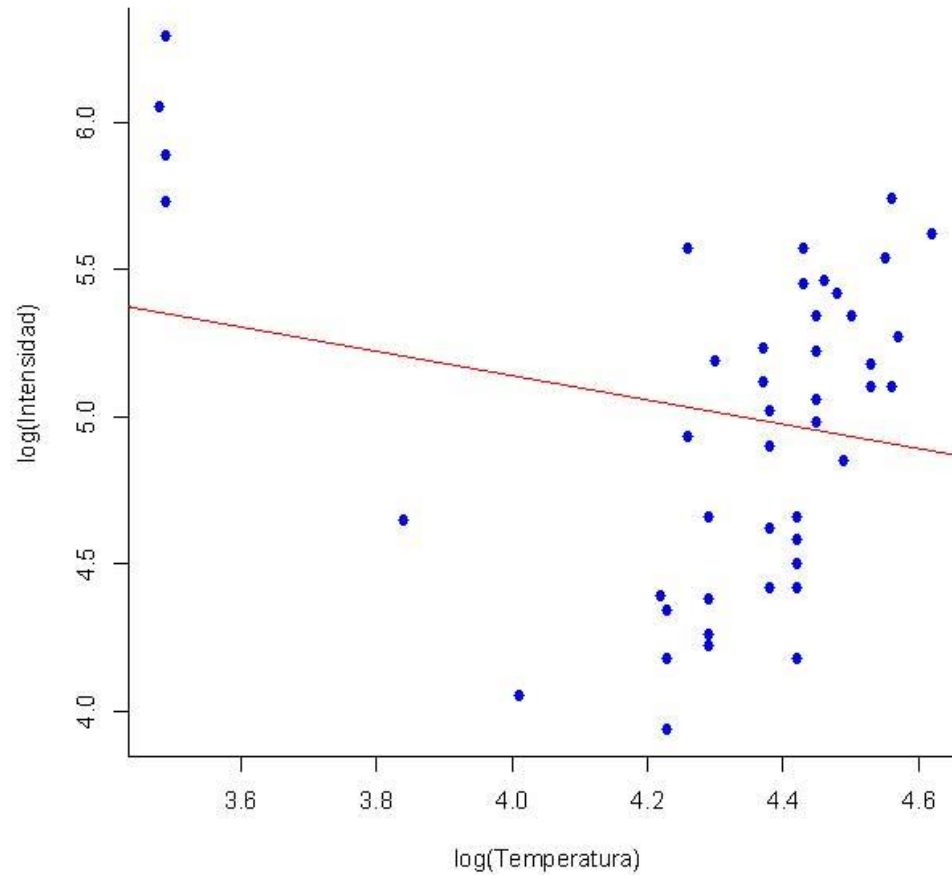
Afecta mucho la recta de regresión, es una observación muy influyente y además tiene residuo grande.

Adecuación del modelo

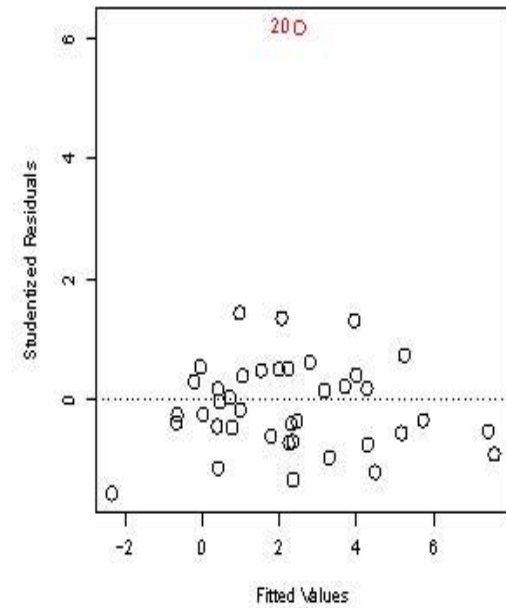
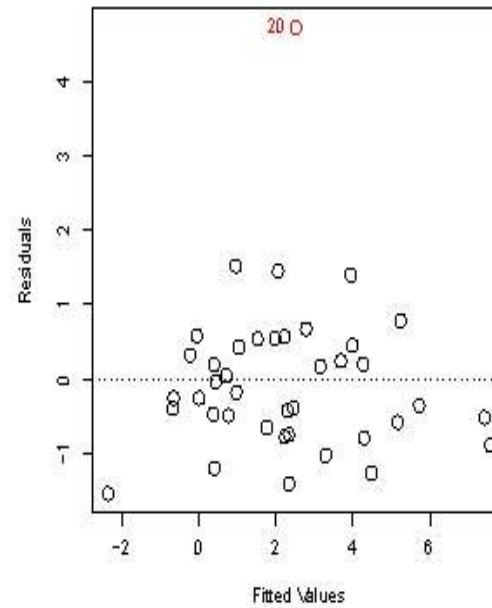
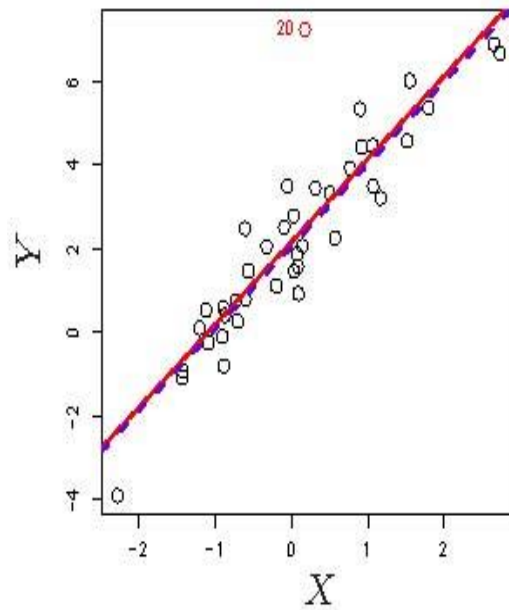


Para 47 estrellas se han registrado el logaritmo de la temperatura efectiva en la supercie (Temp) y el log de la intensidad de su luz (Intens).

Adecuación del modelo



Adecuación del modelo



Adecuación del modelo

¿Cómo detectarlos?

- La estimación es sensible a su presencia. Incluirlos o no pueden modificar drásticamente el RSE o R^2 , aún en los casos donde haya una gran variación en la recta de regresión.
- En el ejemplo el punto rojo es un outlier.
 - RSE (datos completos) 1.09
 - RSE (datos sin outliers) 0.77
 - R^2 (datos completos) 80.5%
 - R^2 (datos sin outliers) 89.2%

Ajuste robusto

- ▶ OLS es muy sensible a la presencia de datos atípicos.
- ▶ Basta con una observación muy alejada del resto para dañar al estimador.
- ▶ Si se tienen varias observaciones es difícil detectar outliers con criterios *leave one out*.
- ▶ En dimensiones altas es difícil detectar outliers, ya que una observación puede no ser atípica en ninguna coordena y sin embargo serlo como observación propiamente dicha.
- ▶ El problema no está en el modelo (lineal) sino en la propuesta que consideramos para estimar los parámetros.

Se propone un método de regresión que no sea sensible a la presencia de datos atípicos.

Propuesta: Modificar la función de pérdida de forma tal que:

- ▶ le de poco peso a las observaciones con residuos grandes.
- ▶ que tenga alta eficiencia, es decir que bajo los supuestos de normalidad el comportamiento de los estimadores sea similar al de OLS.

Recordemos que el estimador de mínimos cuadrados

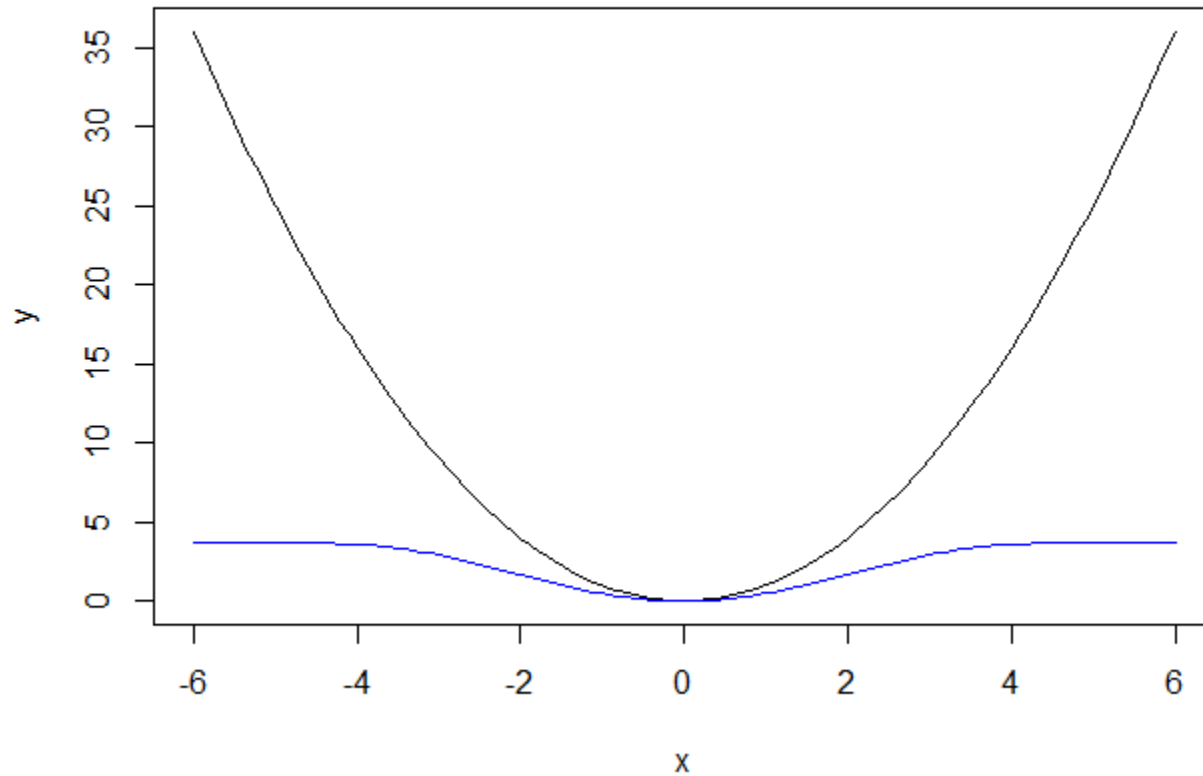
$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^2,$$

proponemos

$$\hat{\beta}_R = \arg \min_{\beta} \rho \left(\frac{y - X\beta}{\hat{\sigma}} \right).$$

Donde la función $\rho : \mathbb{R} \rightarrow \mathbb{R}$ es una función que cerca del cero se comporta como una cuadrática, pero a partir de un punto se mantiene constante. De este modo le da el mismo peso a todos los residuos elevados. $\hat{\sigma}$ es un estimador robusto de la escala de los residuos.

Adecuación del modelo



Adecuación del modelo

MM-estimadores de regresión

Este es un estimador en tres pasos:

1. Calcular un estimador inicial $\hat{\beta}_0$ que sea robusto (aunque tal vez no eficiente).
2. Calcular un estimador robusto de la escala de los residuos $\hat{\sigma}$
 $\hat{r}_i(\hat{\beta}_0) = y_i - X_i\hat{\beta}_0$.
3. Encontrar $\hat{\beta}_R$ mediante un procedimiento iterativo que tenga como estimadores iniciales los hallados en los puntos anteriores

$$\hat{\beta}_R = \arg \min_{\beta} \rho \left(\frac{y - X\beta}{\hat{\sigma}} \right).$$

Adecuación del modelo

Consideremos el siguiente ejemplo:

Consideremos los datos **wood** de la librería *robustbase* de R. El mismo tiene datos que buscan determinar el peso específico de la madera a partir de datos anatómicos de la madera. Los datos originales son del libro de Draper y Smith (1966, p. 227). El conjunto de datos tiene 20 observaciones y 5 covariables. Cuatro datos fueron reemplazados por outliers (son las observaciones 4, 6, 8 y 19).

Adecuación del modelo

```
library(RobStatTM) #esta libreria hay que bajarla de internet.  
library(robustbase)  
data(wood, package='robustbase')  
help(wood)  
cont <- lmrobdet.control(bb = 0.5, efficiency = 0.85, family = "bisquare")  
#MM fit  
woodMM <- lmrobdetMM(y ~ ., data=wood, control=cont)  
#LS fit  
woodLS <- lm(y ~ ., data=wood)
```

summary(woodMM)

Call:

lmrobdetMM(formula = y ~ ., data = wood, control = cont)

Residuals:

Min	1Q	Median	3Q	Max
-0.2533249	-0.0079794	-0.0009577	0.0012654	0.0133343

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.37937	0.05583	6.795	8.66e-06 ***
x1	0.21576	0.04343	4.968	0.000207 ***
x2	-0.07674	0.20364	-0.377	0.711941
x3	-0.56358	0.04446	-12.676	4.62e-09 ***
x4	-0.39615	0.06762	-5.858	4.16e-05 ***
x5	0.60202	0.08172	7.367	3.53e-06 ***

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

Robust residual standard error: 0.01372

Multiple R-squared: 0.8499,

Adjusted R-squared: 0.7963

Convergence in 8 IRWLS iterations

summary(woodLS)

Call:

lm(formula = y ~ ., data = wood)

Residuals:

Min	1Q	Median	3Q	Max
-0.030415	-0.012318	-0.003494	0.012760	0.047892

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.42178	0.16912	2.494	0.02576 *
x1	0.44069	0.11688	3.770	0.00207 **
x2	-1.47501	0.48692	-3.029	0.00901 **
x3	-0.26118	0.11199	-2.332	0.03513 *
x4	0.02079	0.16109	0.129	0.89915
x5	0.17082	0.20336	0.840	0.41505

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

Residual standard error: 0.02412 on 14 degrees of freedom

Multiple R-squared: 0.8084, Adjusted R-squared: 0.74

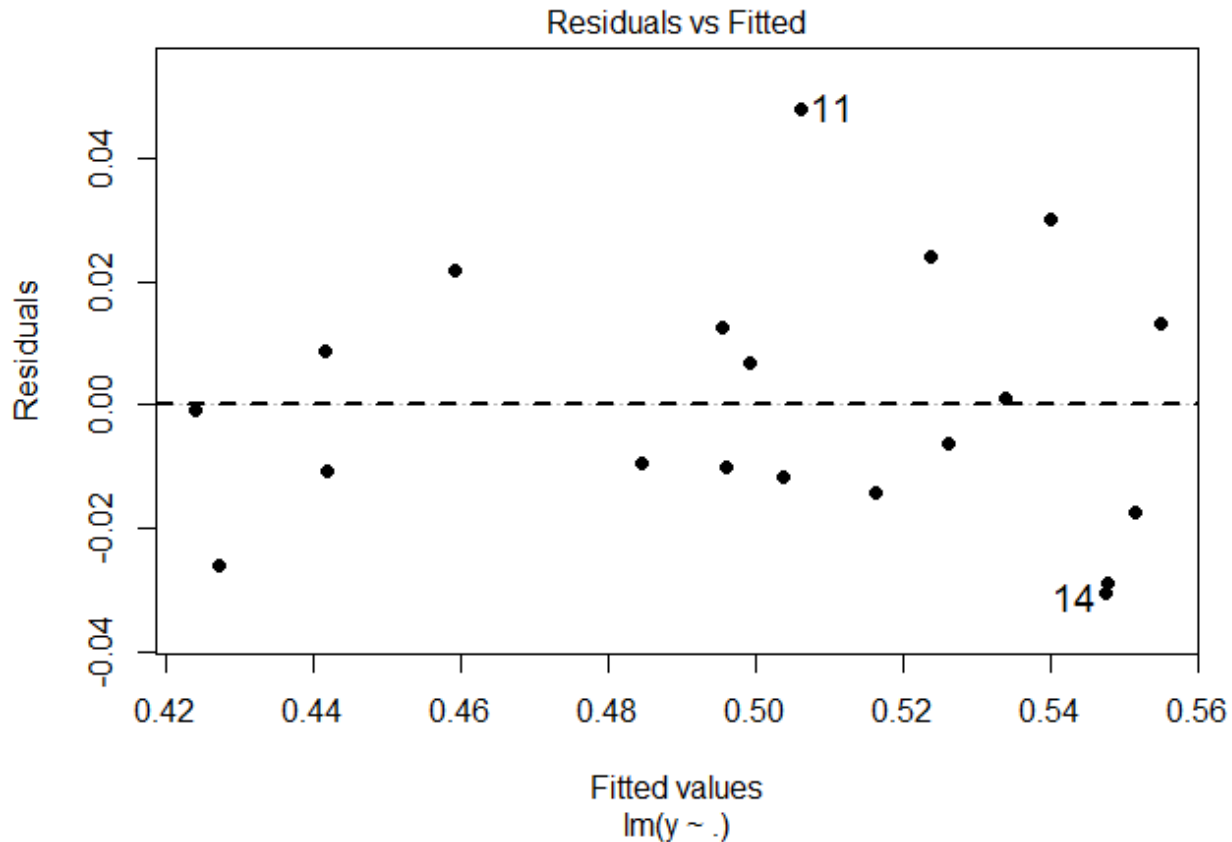
F-statistic: 11.81 on 5 and 14 DF, p-value: 0.0001282

Observaciones:

- Las estimaciones son muy diferentes.
- No coinciden los coeficientes significativos.

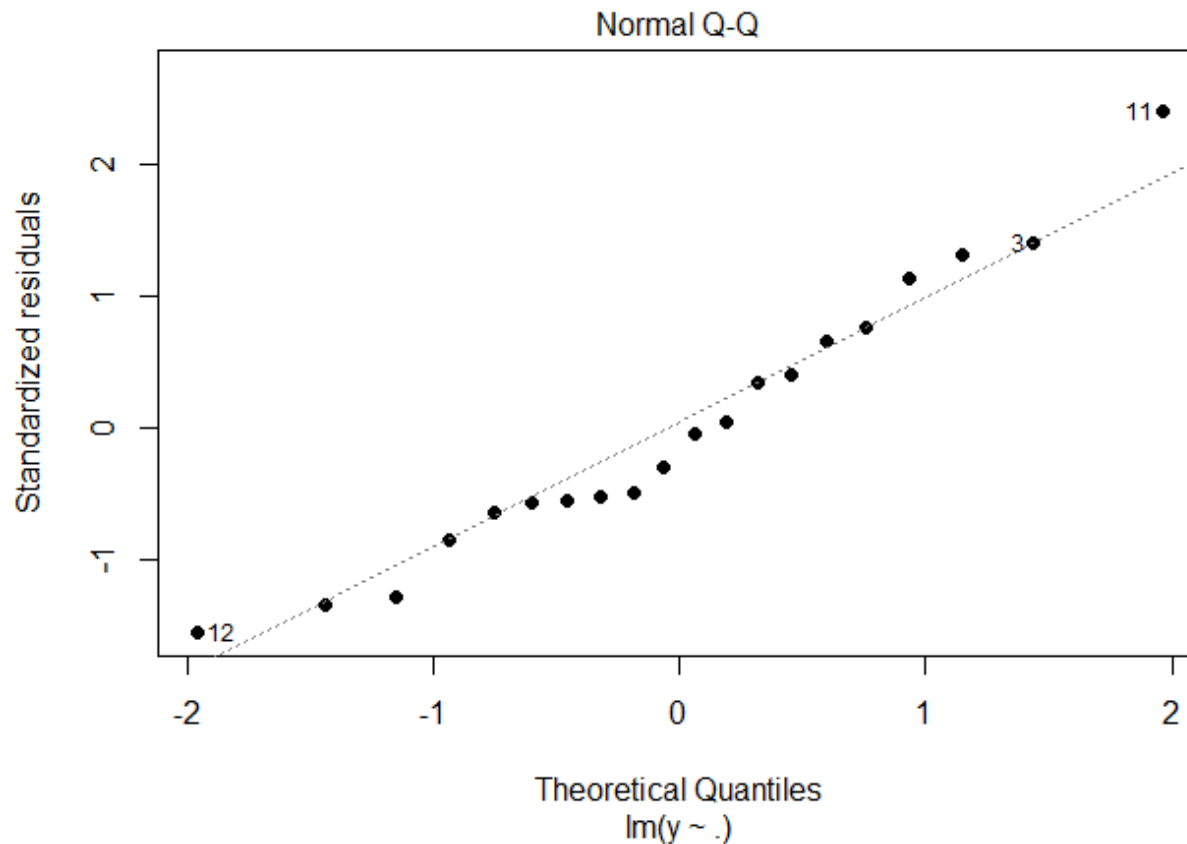
Adecuación del modelo

```
plot(woodLS, which=1, add.smooth=FALSE, pch=19, id.n=2, cex.id = 1.2)  
abline(h=c(-2.5, 0, 2.5) * sigmaLS, lty=2, lwd=2)
```



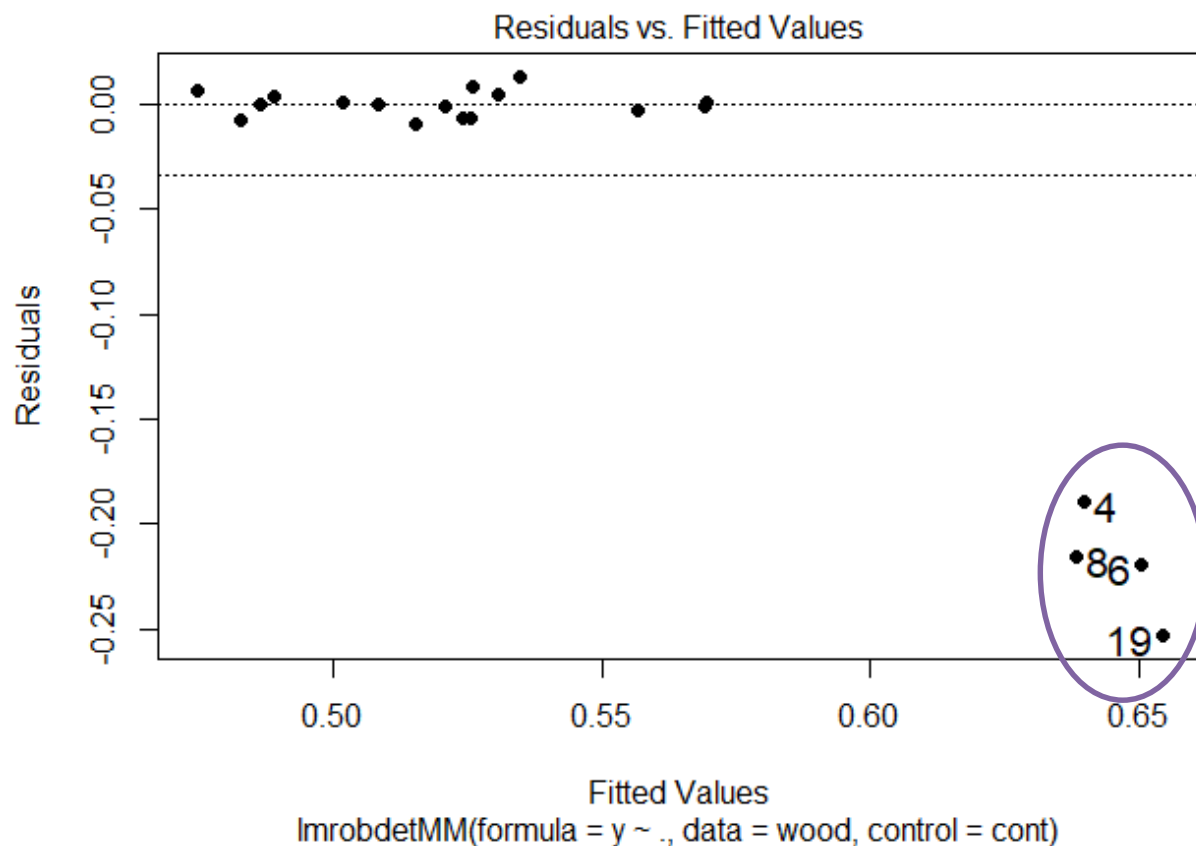
Adecuación del modelo

```
plot(woodLS, which=2, pch=19)
```



Adecuación del modelo

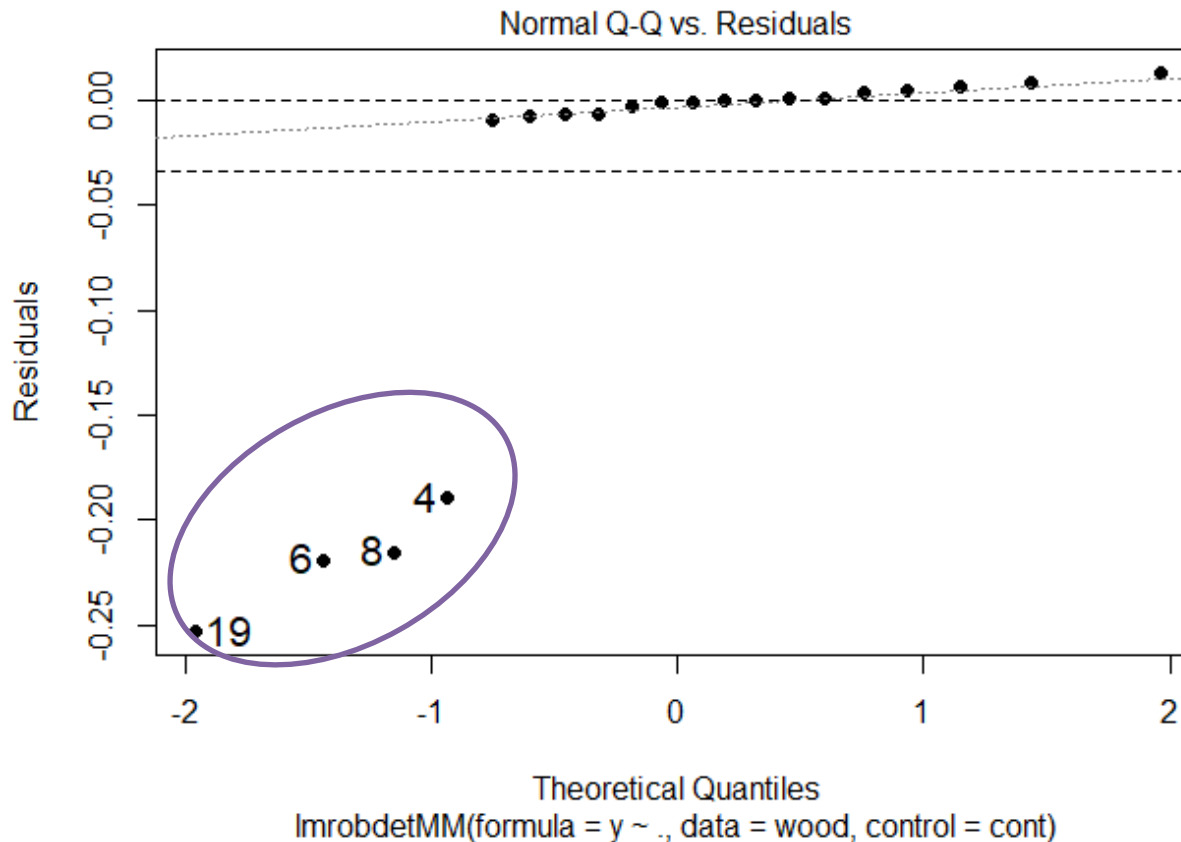
```
plot(woodMM, which=4, add.smooth=FALSE, pch=19, cex.id=1.3, id.n=4)
```



Datos
contaminados

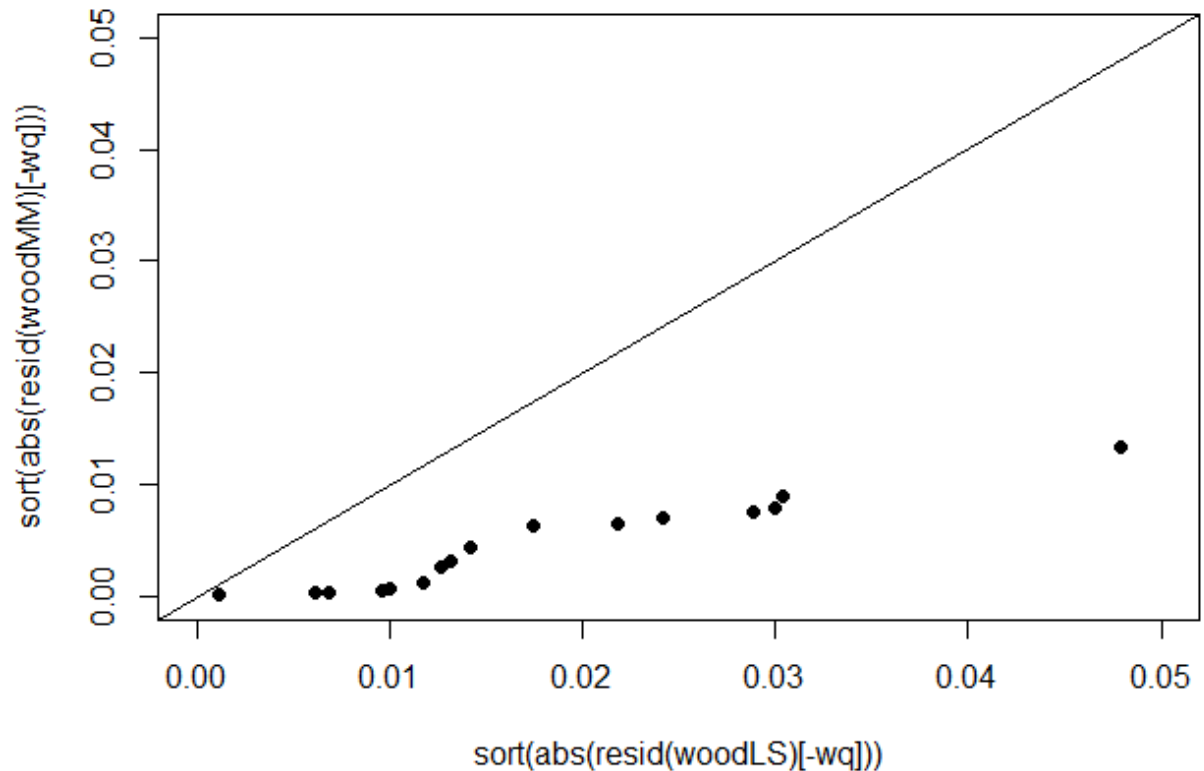
Adecuación del modelo

```
plot(woodMM, which=2, add.smooth=FALSE, pch=19, cex.id=1.3, id.n=4)  
abline(h=c(-2.5, 0, 2.5) * woodMM$scale, lty=2)
```



Adecuación del modelo

Plot de los residuos del MM-estimador vs los residuos de OLS.



Adecuación del modelo

Conclusiones:

- ▶ El ajuste OLS no identifica los outliers.
- ▶ El ajuste por MM-estimador identifica los outliers.
- ▶ Las estimaciones dadas por los dos métodos son muy diferentes, las de OLS están distorcidas por los valores atípicos.
- ▶ Los métodos robustos permiten identificar outliers.
- ▶ El último gráfico nos deja ver que en general (salvo en los outliers) la estimación MM da menores residuos que la OLS.

Adecuación del modelo

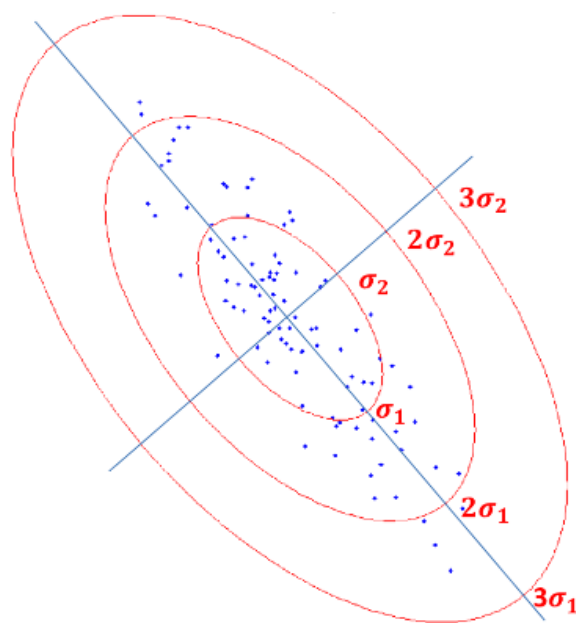
Conclusiones:

- ▶ MM-estimador es un estimador OLS pesado. Si volvemos a correr la regresión OLS pero le ponemos los pesos que da el MM-estimador, obtenemos las estimaciones del MM-estimador.

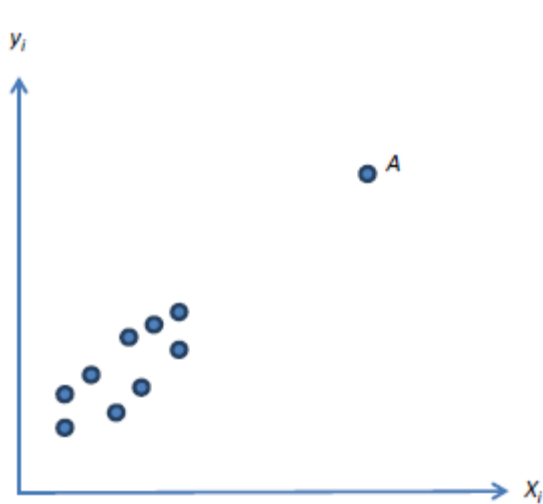
Influencia de las observaciones

Para medir la influencia que tienen las observaciones en la regresión podemos analizar el **leverage**.

Recordemos que el leverage de la observación i está dado por la i -ésima componente de la diagonal de la matriz H , h_{ii} . Sabemos que $0 < h_{ii} < 1$ y además que $tr(H) = \sum_{i=1}^n h_{ii} = p$. El leverage es una medida de la distancia entre los valores de las covariables X de la i -ésima observación respecto del valor del promedio de todas las X observadas en los n casos.

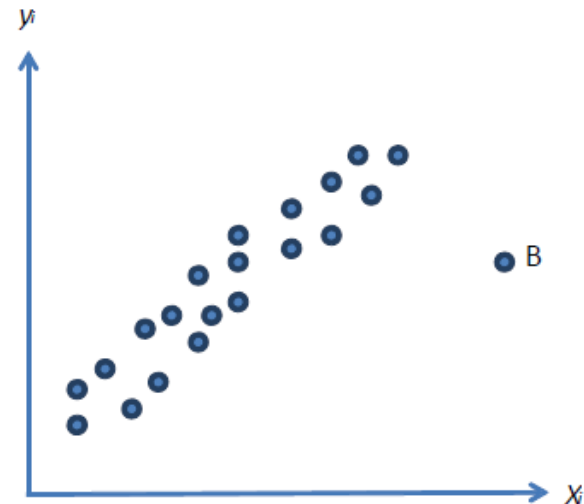


En general tenemos dos casos que nos preocupan



A es un punto alejado del resto, alejado en la coordenada x , pero alienado al resto de los puntos.

- No afecta las estimaciones de los parámetros
- Afecta el R^2 , errores estándares, etc



B es un punto alejado del resto, alejado en la coordenada x así como también en la coordenada y .

- Afecta las estimaciones de los parámetros
- Tuerce la recta de regresión en esa dirección

En ocasiones un conjunto pequeño de datos tiene una influencia desproporcionada sobre los coeficientes del modelo y sus propiedades. Esta situación no es deseable, el ajuste debe representar a la mayoría de los datos y no depender de un pequeño subconjunto.

Es importante, detectar los puntos de alta influencia:

- Determinar si son outliers, en ese caso tratarlos como tales.
- Si no son outliers, comprender en qué modo la presencia de estos puntos afectan la regresión.

Criterios para determinar si una observación tiene alto leverage

- ▶ $h_{ii} > \frac{2p}{n}$.
- ▶ Hacer un histograma o boxplot de los leverage y ver si se detecta alguna estructura, i.e. si hay observaciones con leverage más alto que el resto.

Observación: en ocasiones $\frac{2p}{n} > 1$ y este criterio no se puede aplicar.

Otra medida de influencia es la **distancia de Cook**. Para cada observación x_i se define

$$D_i = \frac{r_i^2}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^2}$$

- ▶ Depende de r_i y de h_{ii}
- ▶ Puede ser alto si tiene
 - ▶ alto residuo y bajo leverage. Es un outlier aunque no modifica la regresión.
 - ▶ bajo residuo y alto leverage. No es un outlier.
 - ▶ alto residuo y leverage. Es un outlier y modifica la regresión.

Una observación tiene D_i alta si es superior $f_{p,n-p,0.5}$.

La salida de estándar de la regresión muestra 3 gráficos por defecto sobre la influencia de los puntos.

Retomemos el ejemplo de Advertising, solo a modo ilustrativo.

```
#calculo las distancias de Cook
```

```
which((cooks.distance(adv.lm.3)>pf(3,193,0.5))==TRUE)
```

```
named integer(0)
```

No hay observaciones con distancia de Cook mayor al percentil 0.5 de una f con 3 y 197 grados de libertad.

```
par(mfrow=c(1,1))
```

```
plot(adv.lm.3,4)
```

```
plot(adv.lm.3,5)
```

```
plot(adv.lm.3,6)
```

Distancias de Cook versus índice.

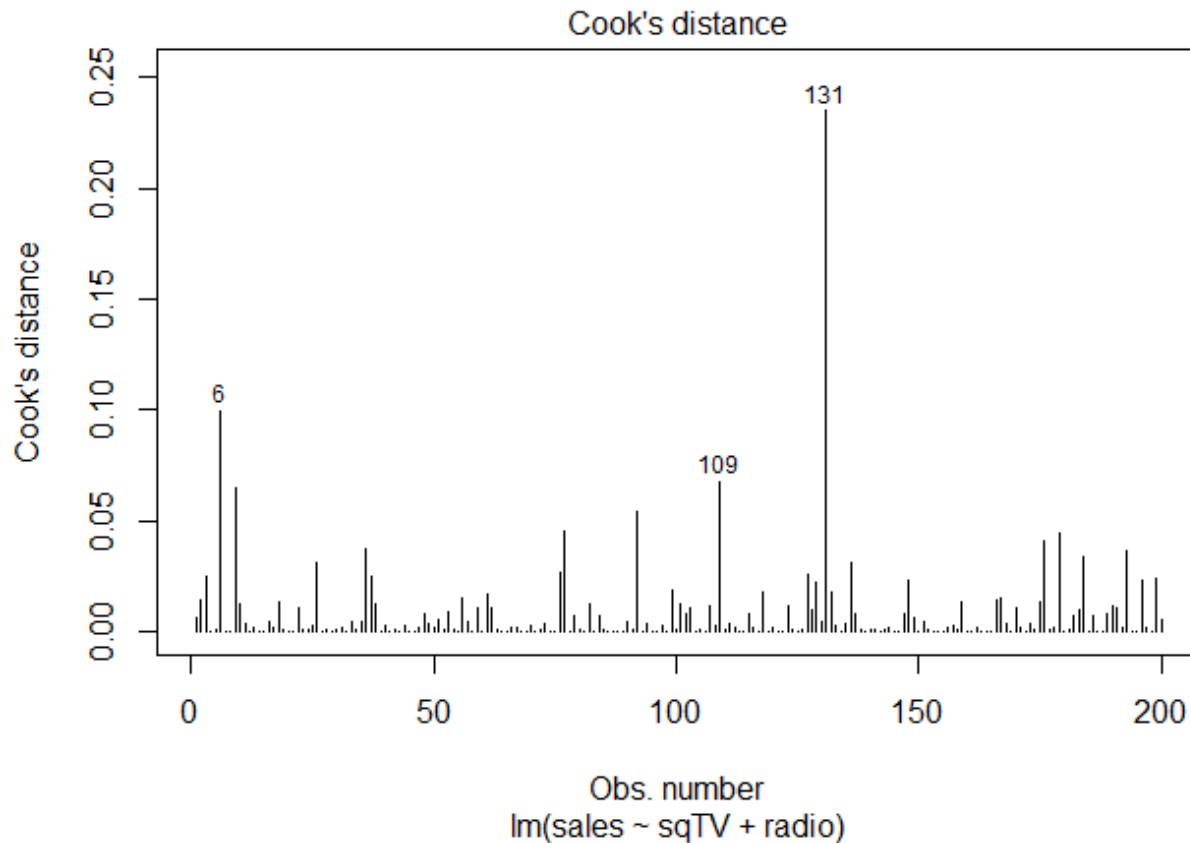


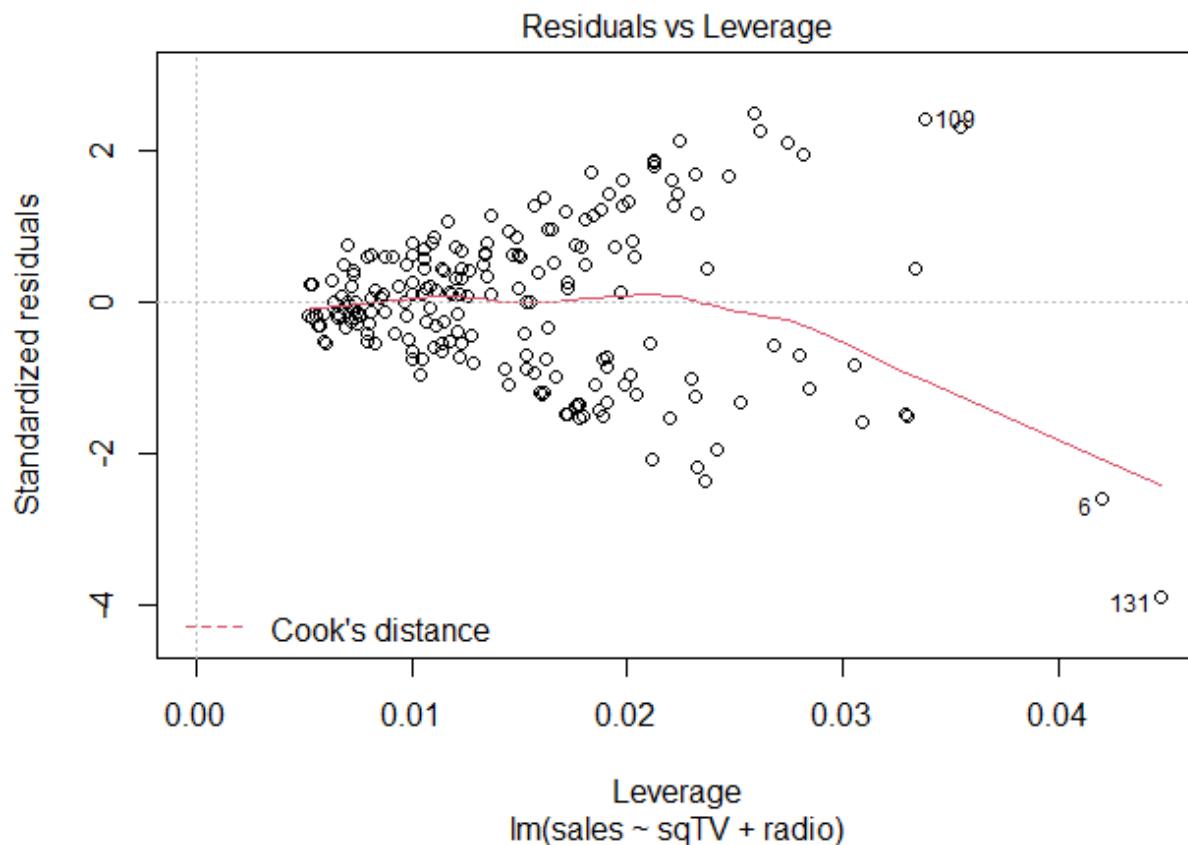
Gráfico de las distancias de Cook:

- Ninguna es superior a 0.42 que es el percentil de la f correspondiente.
- Sin embargo la observación 131 tiene distancia de Cook significativamente más alta que el resto de los puntos.
- También están señaladas las Observaciones 6 y 109.

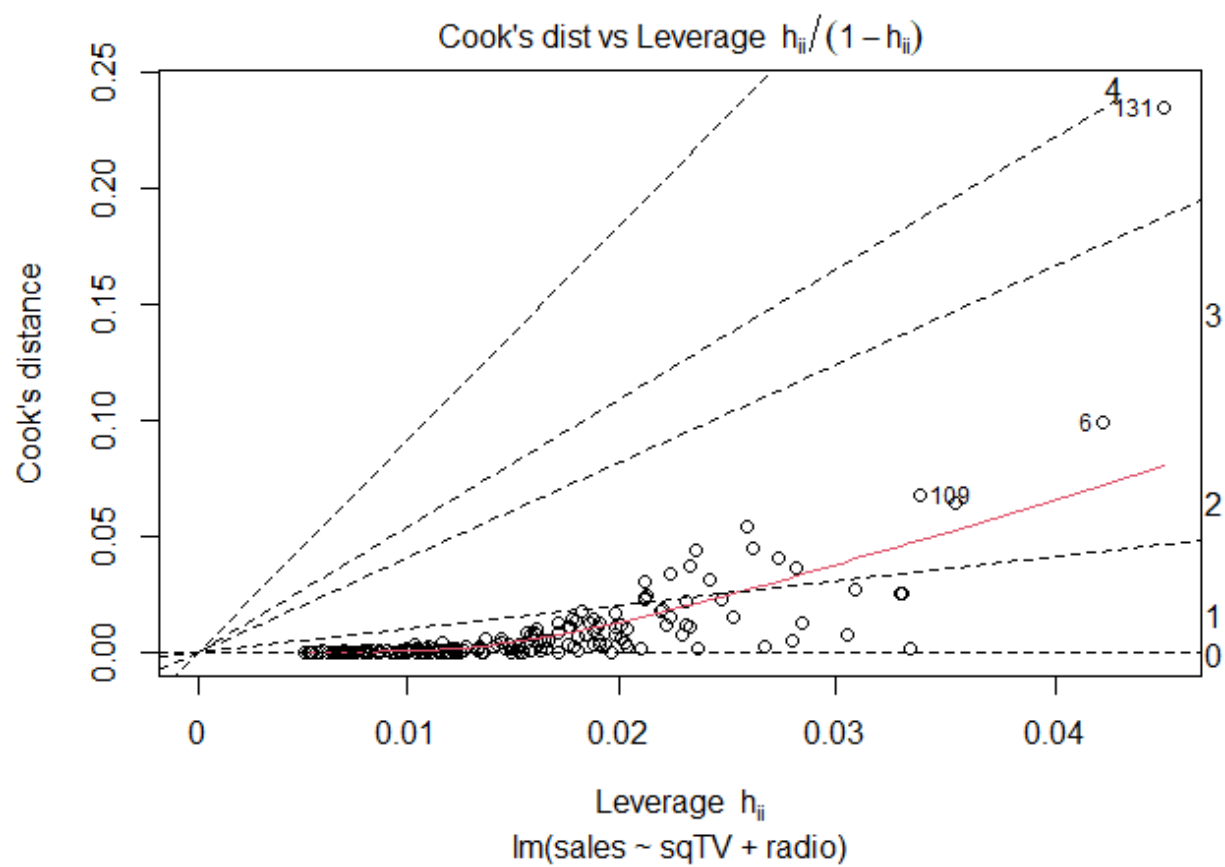
#calculo el leverage

```
which((hatvalues(adv.lm.3)>(2*3/200))==TRUE)
```

3 6 9 76 79 109 127 131 156



- La observación 131 tiene residuo y leverage altos.
- Las otras observaciones con alto leverage no tienen residuo alto.



Otros coeficientes de influencia: DFBetas.

El coeficiente **DFBetas** indica el cambio que produce quitar la observación i sobre la estimación del parámetro β_j .

$$DFBetas_{ij} = \hat{\beta}_j - \hat{\beta}_{j(i)},$$

donde $\hat{\beta}_{j(i)}$ es la estimación de β_j removiendo la i -ésima observación.

Hay que prestar atención si $DFBetas_{ij} > \frac{2}{\sqrt{n}}$.

En R...

`dfbetas(adv.lm.3)`

Otros coeficientes de influencia: DFFits.

El coeficiente **DFFits** indica la cantidad de desvíos estándares produce quitar la observación i sobre la estimación el correspondiente valor predicho \hat{y}_i .

$$DFFits_i = r_{(i)} \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

Hay que prestar atención si $DFFits_i > 2\sqrt{\frac{p}{n}}$.

En R....

```
which(abs(dffits(adv.lm.3))>(2*sqrt(3/200)))
```

Dejo esta página que hace lindos gráficos de residuos

https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html

Bibliografía

- Draper, N. and Smith, H. (1998) Applied Regression Analysis, 3rd Ed., Wiley, Capítulo 2.
- Weisberg, S. (2005) Applied Linear Regression, 3rd Ed., Wiley, Capítulos 7 y 8.
- Maronna, R., Martin, D. and Yohai, V. (2006) Robust Statistics. Wiley. Capítulo 5.5