Regresi—n II

© Ricardo A. Queralt

MSDF: Sesi—n 2

Bibliograf'a

- An Introduction to Statistical Learning with Applications in R.

  Springer

  2013 (Corrected at 6th printing 2015)

  Gareth James ¥ Daniela Witten ¥ Trevor Hastie ¥ Robert Tibshirani

  http://www-bcf.usc.edu/~gareth/ISL/

  Cap'tulos: 3, 5 y 6

- R in Action (SECOND EDITION)

  Data analysis and graphics with R

  ROBERT I. KABACOFF

  MANNING

  https://github.com/kabacoff/RiA2

  Cap'tulo 8

- Practical Data Science with R

  NINA ZUMEL y JOHN MOUNT

  MANNING

  https://github.com/WinVector/zmPDSwR

  Cap'tulo 7

Diagnosis

Processing math: 38%

Table 8.4 Useful functions for regression diagnostics (car package)

Function	Purpose
`qqPlot()`	Quantile comparisons plot
`durbinWatsonTest()`	Durbin–Watson test for autocorrelated errors
`crPlots()`	Component plus residual plots
`ncvTest()`	Score test for nonconstant error variance
`spreadLevelPlot()`	Spread-level plots
`outlierTest()`	Bonferroni outlier test
`avPlots()`	Added variable plots
`influencePlot()`	Regression influence plots
`scatterplot()`	Enhanced scatter plots
`scatterplotMatrix()`	Enhanced scatter plot matrixes
`vif()`	Variance inflation factors

- Normalidad
- Linealidad
- Varianza Constante. Homocedasticidad
- Validaci—n Global
- Multicolinealidad
- Observaciones an—malas

Ejemplo: Advertising

> En el fichero `Advertising.csv` se encuentran los resultados de 200 campa–as de publicidad de una entidad financiera.

> - La variable `Sales` representa el nœmero de productos que han vendido en la campa–a en miles (p.e. el nœmero de fondos de inversi—n, de dep—sitos o de cuentas corrientes).

> Las variables `TV`, `Radio` y `Newspaper` los gastos en miles de euros de las respectivas campa–as de publicidad.

> Hay importantes preguntas que queremos contestar mediante el modelo de regresi—n:

> - ÀHay una relaci—n entre el presupuesto de publicidad y las Ventas?
> - ÀC—mo de fuerte es la relaci—n entre los gastos de publicidad y las ventas?
> - ÀCu‡l de los los medios contribuye m‡s a las ventas?
> - ÀC—mo de precisos se pueden estimar los efectos de cada medio sobre las ventas?
> - ÀC—mo de preciso se pueden predecir las ventas futuras?
> - ÀEs la relaci—n lineal?
> - ÀHay sinergias entre los diferentes tipos de anuncios?

```
mData=read.csv("./Datos/Advertising.csv")
regres01=lm(Sales~TV+Radio+Newspaper,data=mData)
summary(regres01)
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = mData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

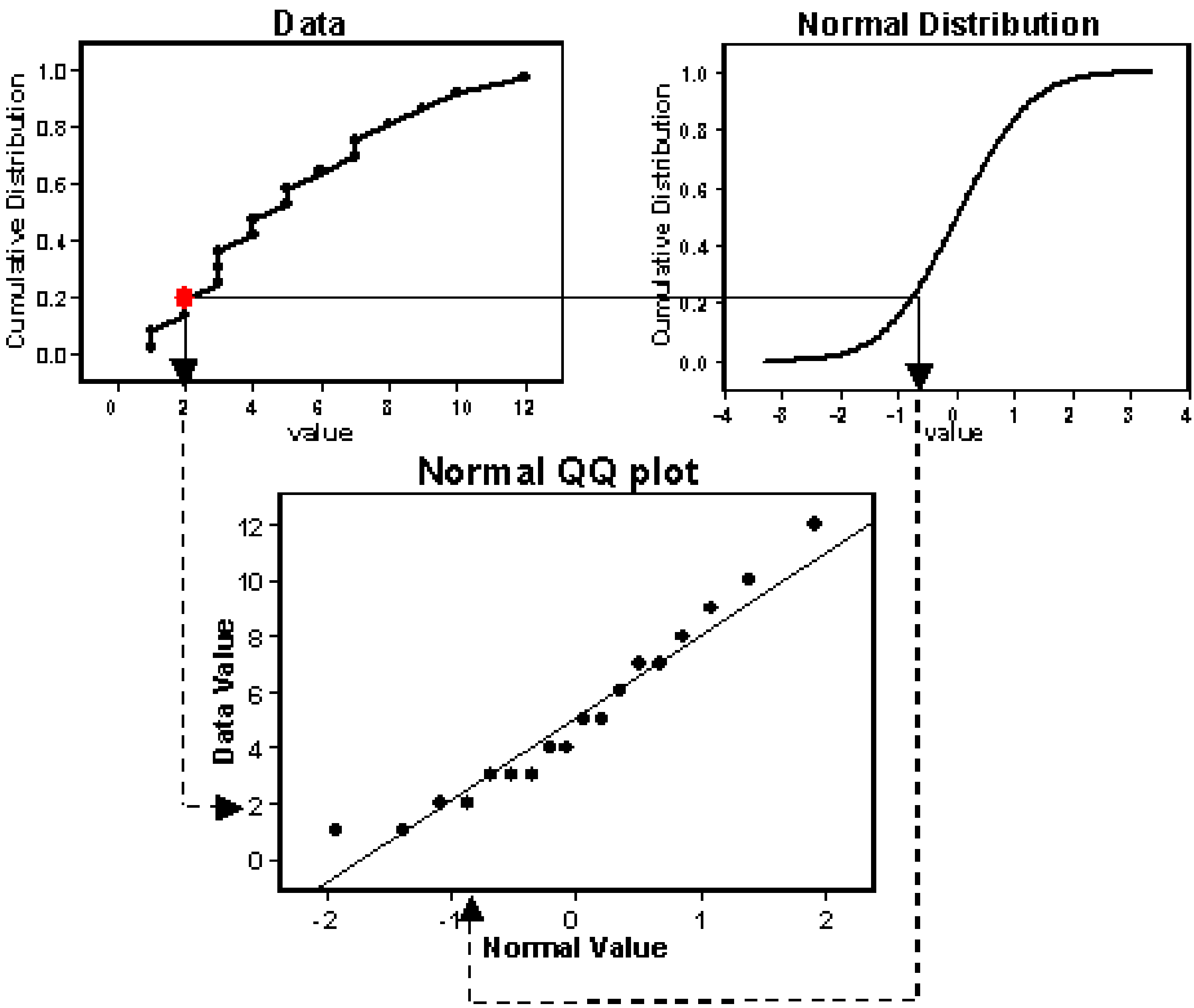
Normalidad

qqplot

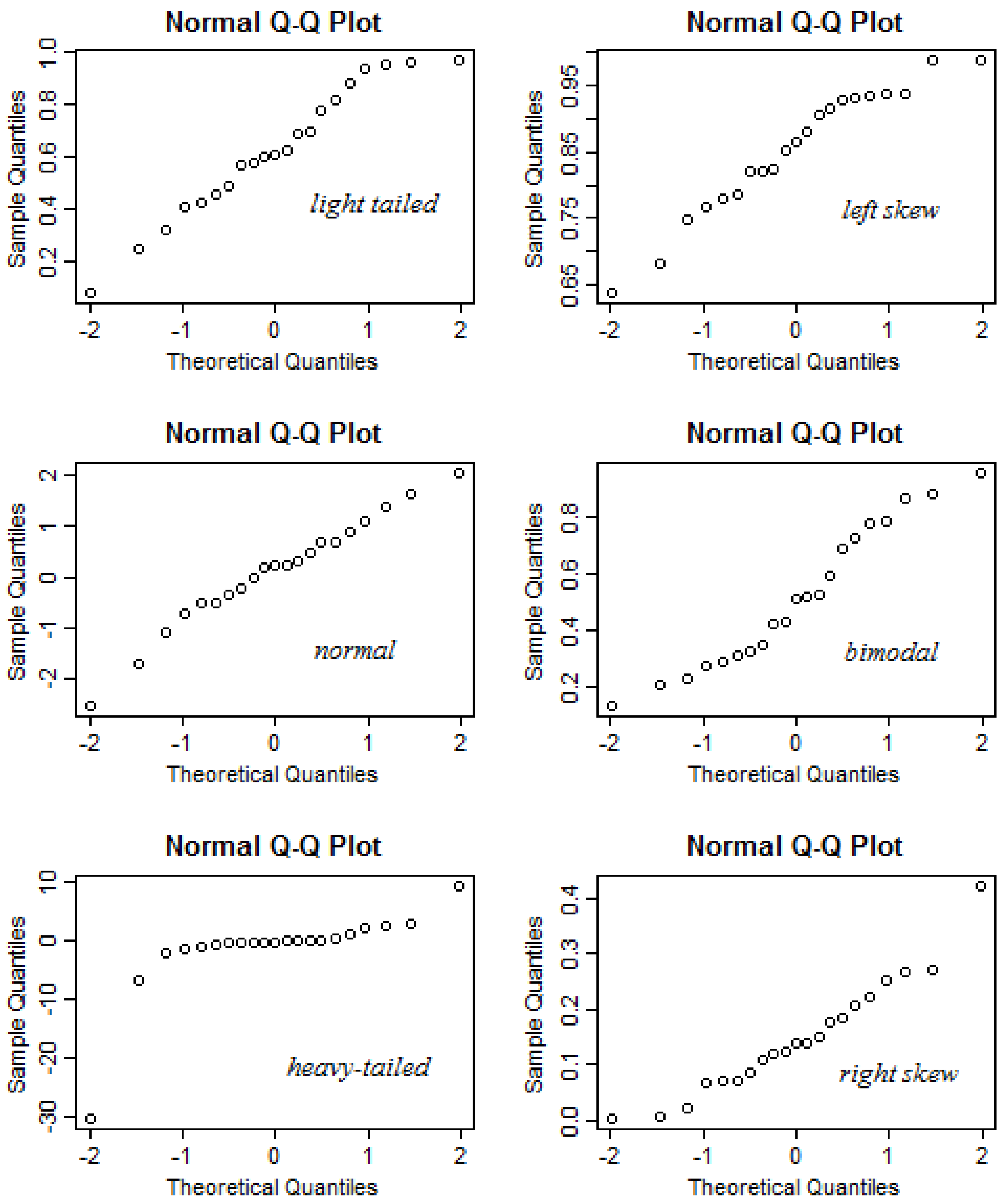
Un QQ-plots (quantile vs quantile plots) es una representaci—n gr‡fica, que sirve para comparar dos distribuciones y ver si coinciden. En realidad, para comprobar si un conjunto de datos muestrales est‡n generados por una distribuci—n te—rica como la normal , la exponencial, É

Un QQ-plot es un gr‡fico de puntos que muestra los cuantiles. Si ambos cuantiles viene de la misma distribuci—n, veremos que los puntos forman una linea recta, sino esto no ocurre entenderemos que los datos muestrales no han sido generados por la distribuci—n te—rica. Al gr‡fico de puntos se le puede a–dir un intervalo de confianza.

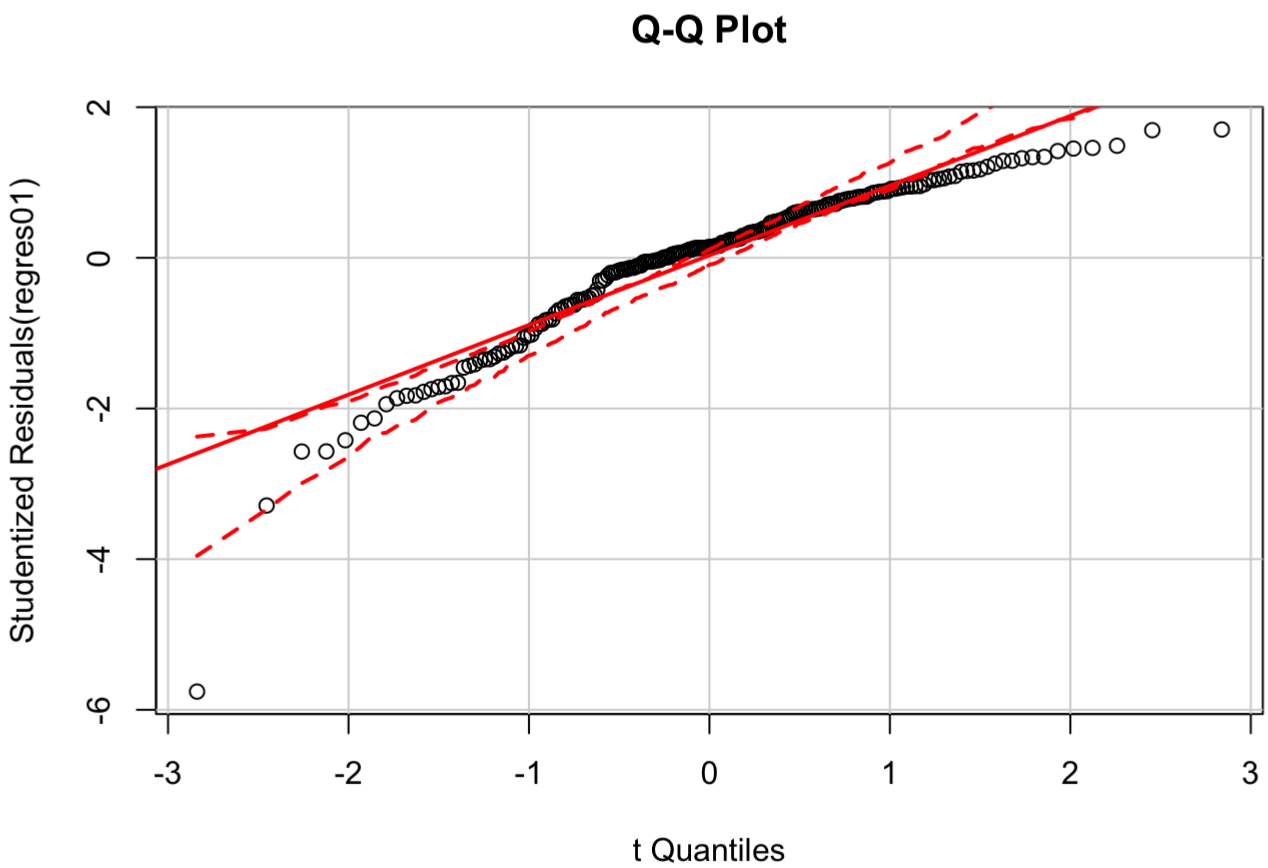
Creando un qq-plot



Interpretaci—n del qq-plot



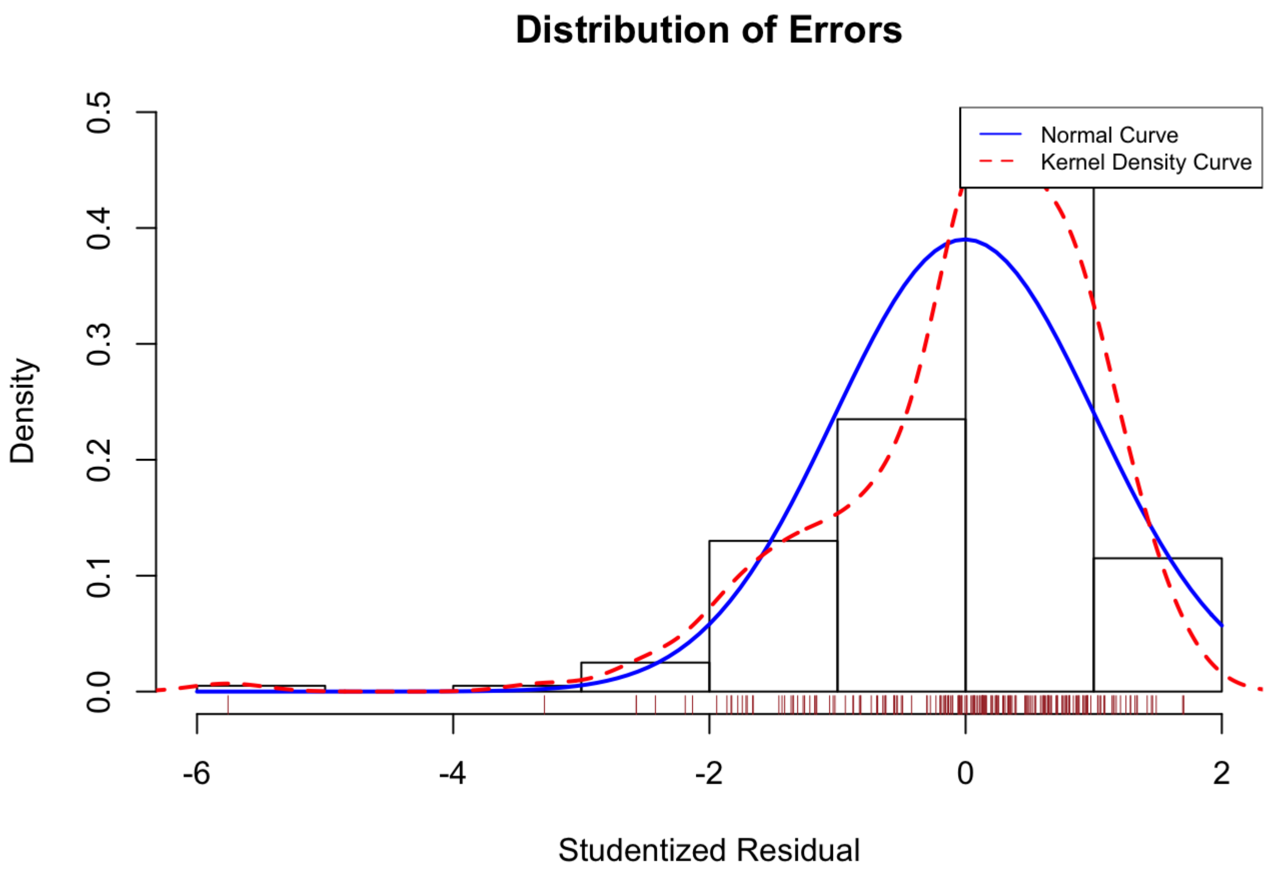

```
library(car)
qqPlot(regres01, labels=row.names(mData), id.method="identify",
       simulate=TRUE, main="Q-Q Plot")
```



Histograma + densidad + normal + rug

- ÀQuŽ son los `Studentized Residual`?
- ÀLa media de los residuos y la constante?
- ÀDiferencia entre histograma y densidad?

```
residplot <- function(fit, nbreaks=10) {
  z <- rstudent(fit)
  hist(z, breaks=nbreaks, freq=FALSE,
       xlab="Studentized Residual",
       main="Distribution of Errors")
  rug(jitter(z), col="brown")
  curve(dnorm(x, mean=mean(z), sd=sd(z)),
        add=TRUE, col="blue", lwd=2)
  lines(density(z)$x, density(z)$y,
        col="red", lwd=2, lty=2)
  legend("topright",
         legend = c( "Normal Curve", "Kernel Density Curve"),
         lty=1:2, col=c("blue","red"), cex=.7)
}

residplot(regres01)
```



Jarque Bera

El contraste de normalidad de Jarque-Bera consiste en

$$\text{JB} = \frac{n-k+1}{6} \quad S^2 + \frac{1}{4}(C-3)^2$$

Donde n es el nœmero de observaciones; S es el coeficiente de asimetr'a muestral; C es la curtosis muestral y k es el nœmero de regresores:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad ^{3/2}},$$

$$C = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad ^{2}},$$

El estad'stico se distribuye como una chi-cuadrado de 2 grados de libertad, siendo *la hip—tesis nula* que la distribuci—n es normal (coeficiente de asimetr'a y curtosis es cero)

```
vResid=resid(regres01)
library(fBasics)
```

```
## Loading required package: timeDate
```

```
## Loading required package: timeSeries
```

```
##
```

```
## Rmetrics Package fBasics
```

```
## Analysing Markets and calculating Basic Statistics
```

```
## Copyright (C) 2005-2014 Rmetrics Association Zurich
```

```
## Educational Software for Financial Engineering and Computational Science
```

```
## Rmetrics is free software and comes with ABSOLUTELY NO WARRANTY.
```

```
## https://www.rmetrics.org --- Mail to: info@rmetrics.org
```

```
##
## Attaching package: 'fBasics'
```

```
## The following object is masked from 'package:car':
##
##     densityPlot
```

```
jbTest(vResid)
```

```
##
## Title:
##  Jarque - Bera Normality Test
##
## Test Results:
##   PARAMETER:
##     Sample Size: 200
##   STATISTIC:
##     LM: 151.241
##     ALM: 161.997
##   P VALUE:
##     Asymptotic: < 2.2e-16
##
## Description:
##  Mon Nov 28 01:25:41 2016 by user:
```

Shapiro-Wilk

El test de Shapiro-Wilk permite comprobar si una muestra ha sido generada por un distribuci—n normal.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

donde

$x_{(i)}$ es el nœmero que ocupa la i-Žsima posici—n en la muestra;

$\bar{x} = \; x_1 + \quad + x_n \; / n$ es la media muestral;

a_i se obtiene de:

$$(a_1, \ldots, a_n) = \frac{m^{\mathrm{T}} V^{-1}}{(m^{\mathrm{T}} V^{-1} V^{-1} m)^{1/2}},$$

donde

$m = (m_1, \ldots, m_n)^{\mathrm{T}}$,

$m_1, \ldots, m_n$ son los valores medios del estad'stico ordenado, de variables aleatorias independientes e idŽnticamente distribuidas, muestreadas de distribuciones normales. V es la matriz de covarianzas de ese estad'stico de orden.

```
shapiro.test(vResid)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  vResid
## W = 0.91767, p-value = 3.939e-09
```

Linealidad

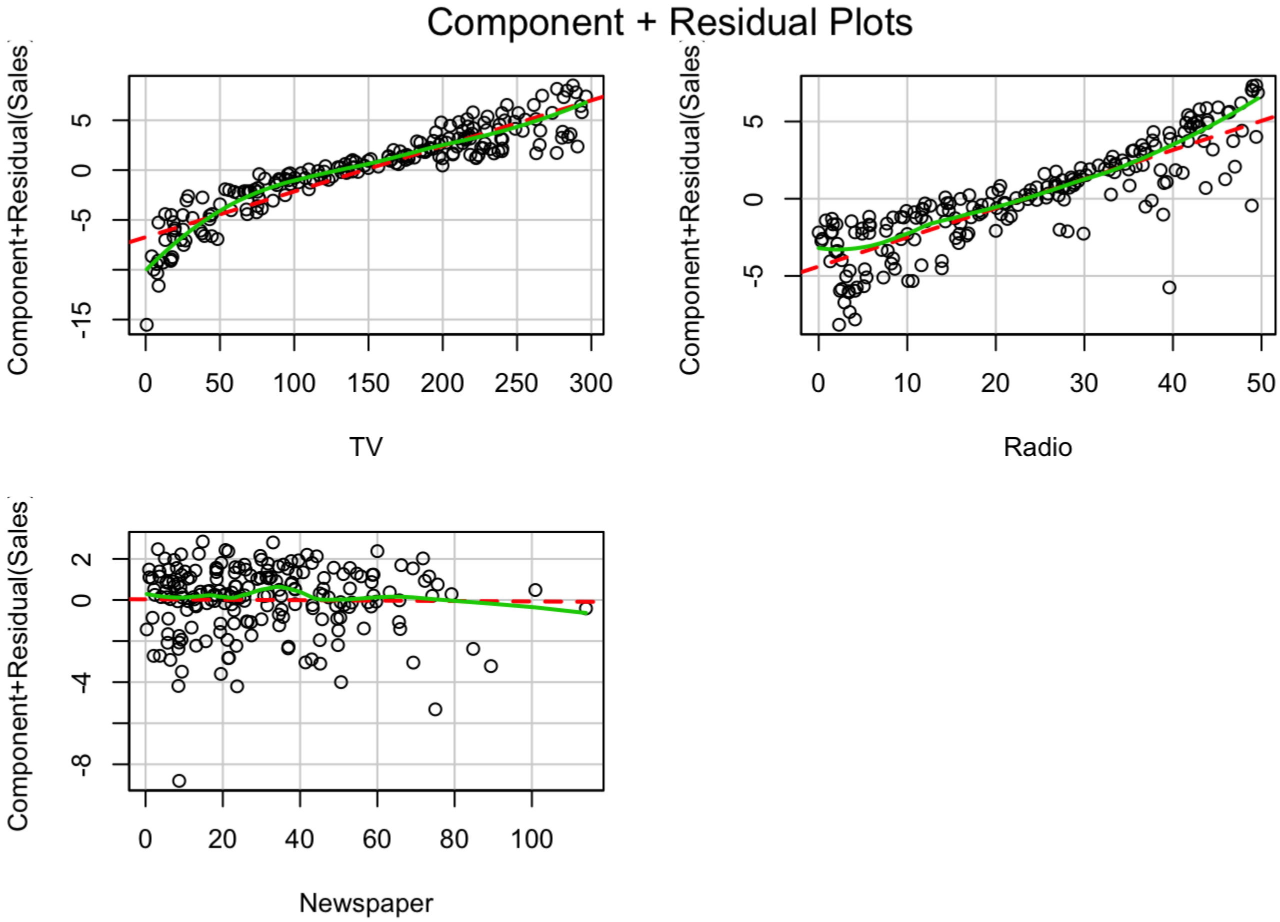
Componentes o Gr‡ficos de residuos parciales

Se grafican los valores ajustados con respecto a los predictores, si no hay problemas de linealidad se obtiene un recta sobre las que se representan los puntos.

Los gr‡ficos se incluyen una estimaci—n suavizada (verde) que deber'a aproximarse a la linea recta (roja).

$$Residuos + \hat{\beta}_i X_i \text{ versus } X_i$$

```
crPlots(regres01)
```



Varianza Constante. Homocedasticidad

Varianza no constante (ncvtest) (Breusch-Pagan)

La hip—tesis nula es que la varianza es constante.

El test BreuschÐPagan test esta basado sobre modelos del tipo $\sigma_i^2 = h(z_i' \gamma)$ para las varianzas de las observaciones, donde $z_i = (1, z_{2i}, \ldots, z_{pi})$ explican la varaibilidad de la varianza (la varianza no constante).

La hip—tesis nula de ausencia de heterocedasticidad es equivalente a $(p - 1)$ restricciones de igualdad a cero de los par‡metros:

$$\gamma_2 = \quad = \gamma_p = 0.$$

El test es equivalente a un proceso en 3 etapas:

- *Etapa 1*: Aplicar MCO al modelo

$$y = X\beta + \varepsilon.$$

  y calcular los residuos.

- *Etapa 2*: Realizar la regresi—n auxiliar:

$$e_i^2 = \gamma_1 + \gamma_2 z_{2i} + \quad + \gamma_p z_{pi} + \eta_i.$$

  Siendo, z sustituida por los regresores x. I

- *Etapa 3*: El estad'stico de contraste es el R^2 de la regresi—n auxiliar multiplicado por el tama–o muestral n:

$$\text{LM} = nR^2.$$

El estad'stico se distribuye asint—ticamente como una χ^2 bajo la hip—tesis nula de homocedasticidad.

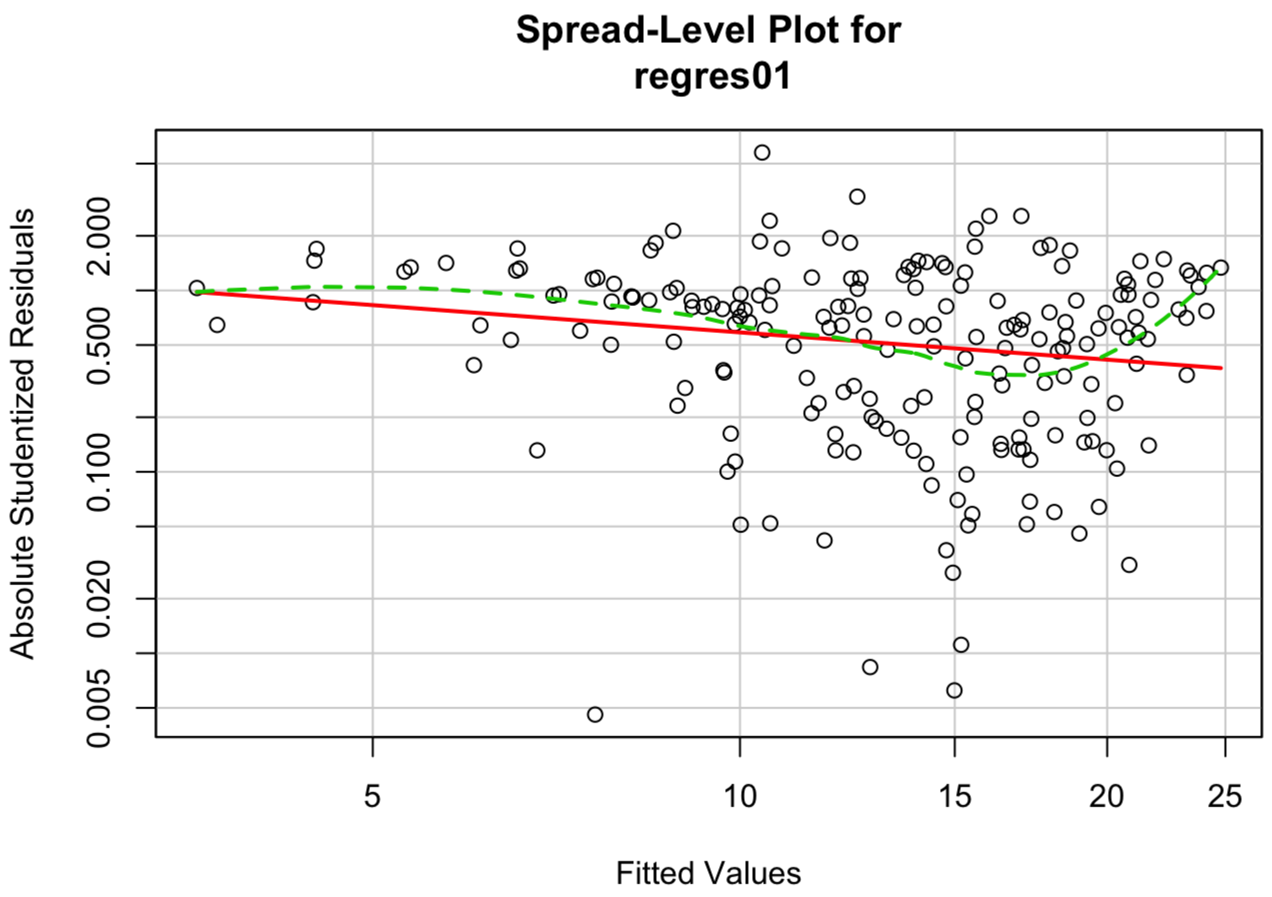
$p-1$

TambiŽn se puede representar los residuos estandarizados absolutos versus los valores ajustados, se superpone la mejor linea recta que ajusta los datos. (Si se cumple la hip—tesis de homocedasticidad se espera que la linea sea horizontal)

```
ncvTest(regres01)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.355982    Df = 1     p = 0.02065131
```

```
spreadLevelPlot(regres01)
```




```
##
## Suggested power transformation:  1.499852
```

Validaci—n Global

Podemos contrastar todas las hip—tesis del modelo mediante el test de Pe–a, EA and Slate, EH (2006). ÒGlobal validation of linear model assumptions,Ó J.Amer. Statist. Assoc., 101(473):341-354.

```
library(gvlma)
gvmodel <- gvlma(regres01)
summary(gvmodel)
```

```
##
```

```
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper, data = mData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## Radio        0.188530   0.008611  21.893   <2e-16 ***
## Newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = regres01)
##
##                        Value   p-value                   Decision
## Global Stat         197.3907 0.000e+00 Assumptions NOT satisfied!
## Skewness             58.7289 1.810e-14 Assumptions NOT satisfied!
## Kurtosis             92.5125 0.000e+00 Assumptions NOT satisfied!
## Link Function        45.9814 1.194e-11 Assumptions NOT satisfied!
## Heteroscedasticity    0.1678 6.820e-01    Assumptions acceptable.
```

Multicolinealidad

Es la existencia de alta correlaci—n entre los predictores puede producir problemas de imprecisi—n de los estimadores (las varianzas de los estimadores son mayores de lo que deber'an ser). As', los intervalos de confianza son muy anchos, hay dificultad para interpretar los coeficientes y se tiende a no rechazar las hip—tesis nula de significaci—n.

El caso m‡s extremo es cuando dos regresores son combinaci—n lineal, en este caso no se puede obtener el estimador.

En algunas ocasiones el a–adir nuevos datos, el valor de los estimadores cambian bastante, entonces seguramente es causado por la multicolinealidad.

Detecci—n de la multicolinealidad

MŽtodo del Factor de Inflaci—n de la Varianza

Para detectar la multicolinealidad se utiliza el `Factor de inflación de varianza (VIF)`.

Se define como:

$$FIV(\beta_j) = \frac{1}{1 - R_j^2} \qquad Tolerancia(\beta_j) = 1 - R_j^2$$

Donde R_j^2 es el coeficiente de determinaci—n de la regresi—n del regresor j con respecto a todos los demas regresores.

Para cualquier regresor la ra'z del VIF indica cuantas veces es la varianza del estimador es mayor que la que se obtendr'a si no hubiera correlaci—n entre los regresores. Cuando $\sqrt{VIF} > 2$ se considera que hay problemas de multicolinealidad.

```
vif(regres01)
```

```
##        TV     Radio Newspaper
##  1.004611  1.144952  1.145187
```

```
sqrt(vif(regres01)) > 2 # problem?
```

```
##        TV     Radio Newspaper
##     FALSE     FALSE     FALSE
```

Observaciones an—malas

```
(1) Atípicos:
Una observación es atípica si el residuo asociado es grande.
(2) Extrema o Apalancada:
Una observación es extrema (o potencialmente influyente o apalancada) si se encuentra aprec
iablemente alejada del resto de observaciones de la muestra.
(3) Influyente:
Una observación es influyente si la presencia de dicha observación en la muestra altera sig
nificativamente algún aspecto de la estimación del modelo.
```

Identificamos los valores at'picos mediante un Bonferroni p-values. En este caso solo contrasta el mayor de los residuos, si se rechaza que sea at'pico se concluye que no hay at'picos.

TambiŽn se puede realizar un gr‡fico de los residuos estandarizados con $\pm 2\sigma$.

```
# Assessing outliers

outlierTest(regres01)
```

```
##      rstudent unadjusted p-value Bonferonni p
## 131 -5.757983          3.267e-08    6.534e-06
```

Para determinar valores extremos, se calcula el `hat statistic`, siendo la media p/n donde p es el nœmero de par‡metros estimados y n el tama–o muestral. Las observaciones con un valor (2 o 3 veces la media) `hat` alto se consideran extremas.

Dado el estimador:

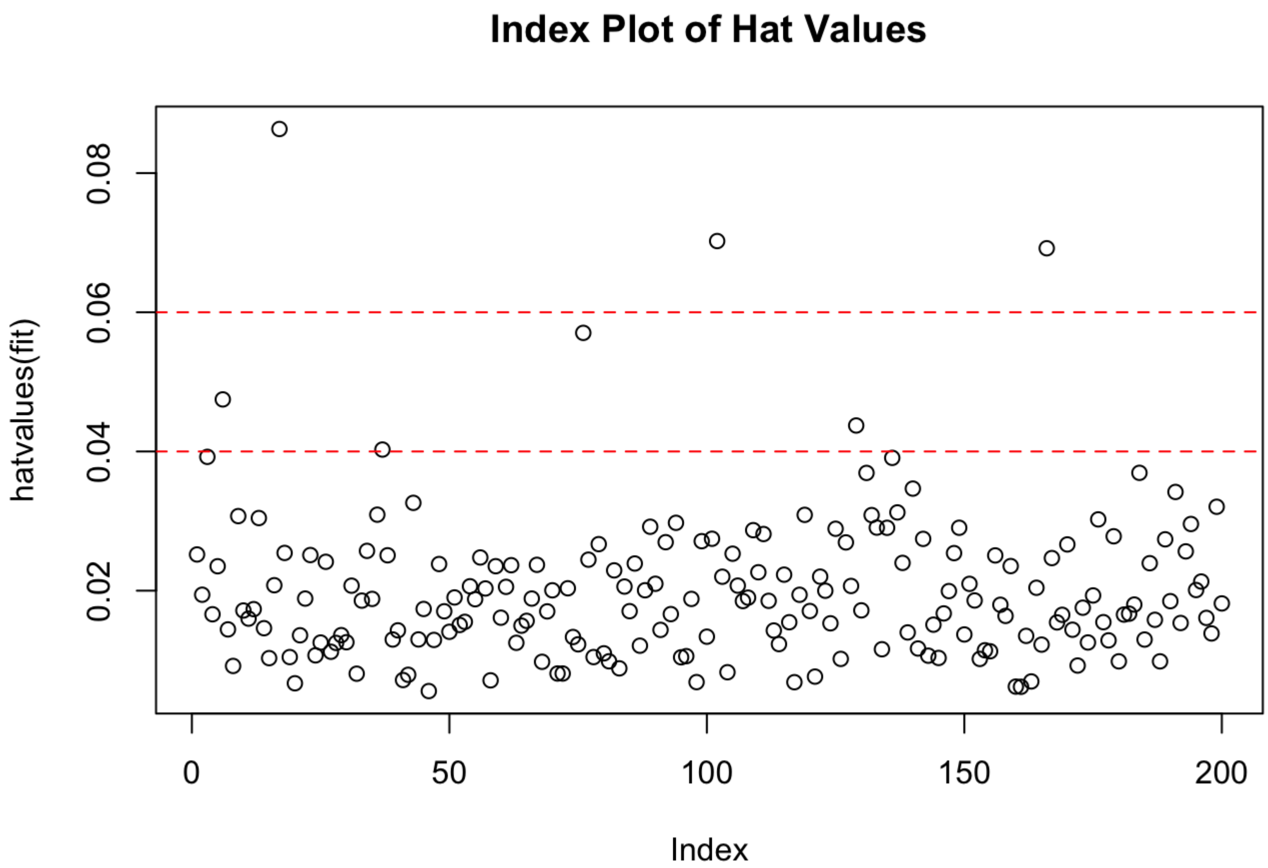
$$\hat{\boldsymbol{\beta}} = \ \mathbf{X}^{\mathsf{T}}\mathbf{X}\ ^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Los valores ajustados son: \hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol \beta} = \mathbf{X} \left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathsf{T}} \mathbf{y}

Entonces la matriz de proyecci—n (hat matrix) es: \mathbf{H} = \mathbf{P} \equiv \mathbf{X} \left(\mathbf{X}^{\mathsf{T}}

\mathbf{X} \right)^{-1} \mathbf{X}^{\mathsf{T}}

```
#  Identifying high leverage points
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main="Index Plot of Hat Values")
  abline(h=c(2,3)*p/n, col="red", lty=2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(regres01)
```




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

Hay dos mŽtodos par identificar observaciones influyentes:

- La distancia de Cook (D-estad'stico)

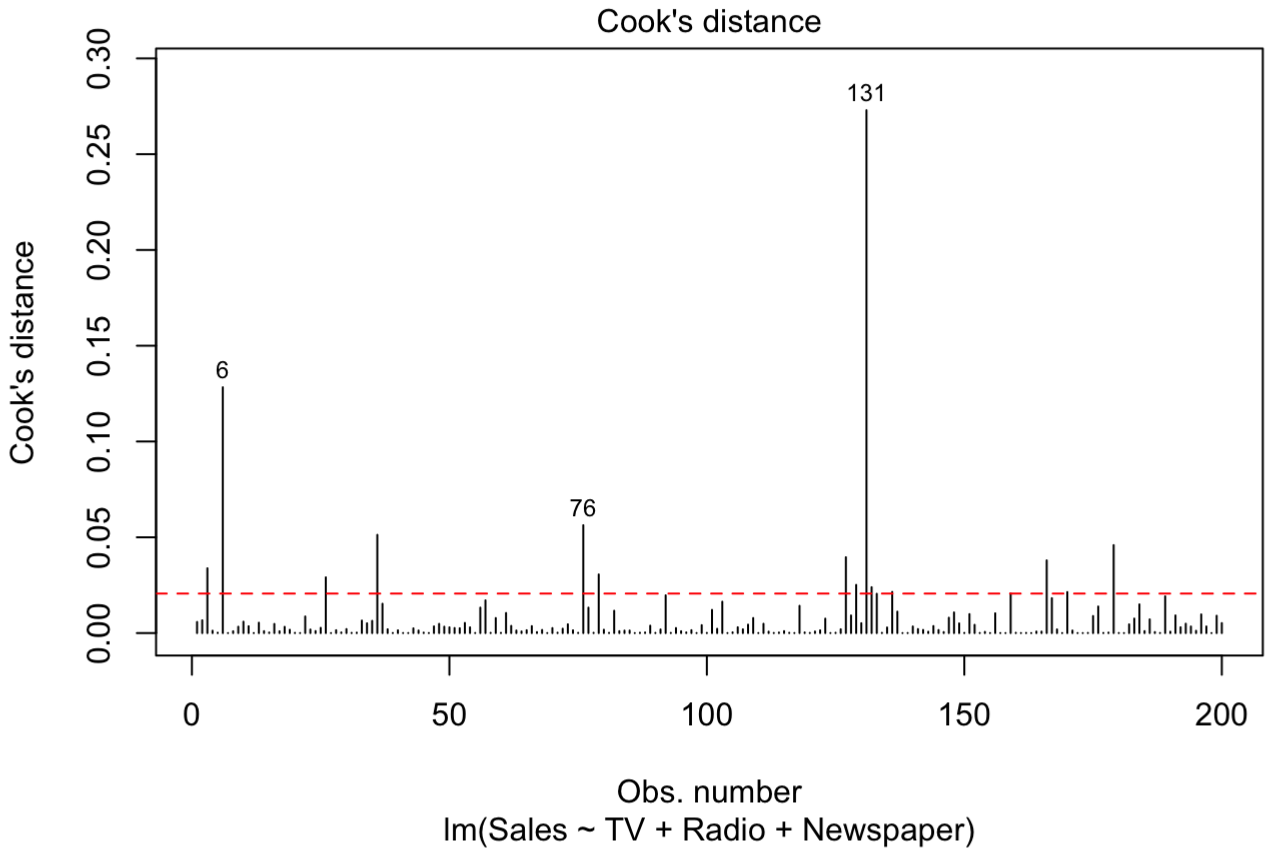
Valores de D mayores que \frac{4}{(n-k-1)} indican que son variables influyentes.

Ayuda a identificar los valores influyentes pero no indica como afectan al modelo.

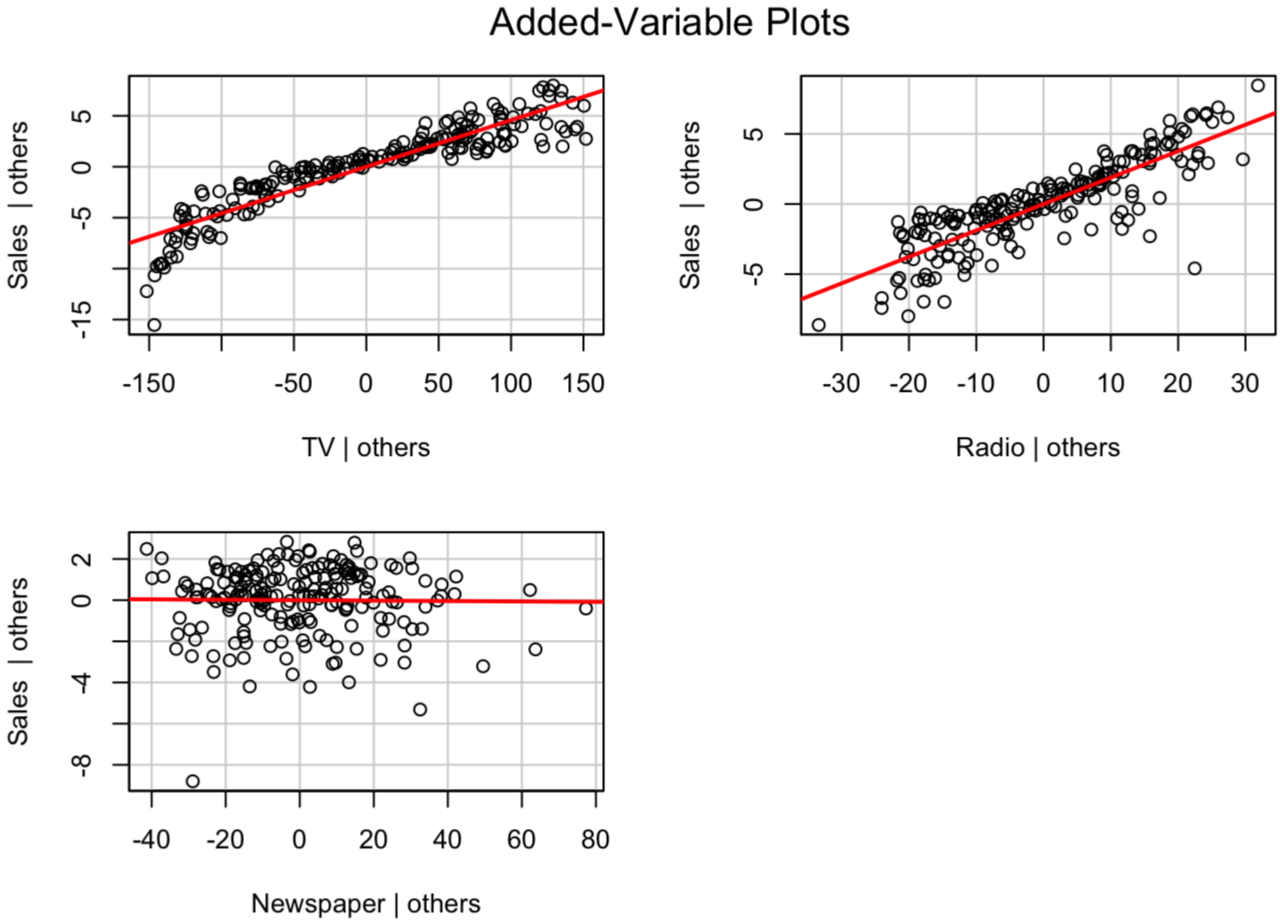
- Gr‡ficos `added variable`

Para cada regresor X_k, se grafican los residuos de la regresi—n de y sobre todos los regresores menos X_k y los residuos de la regresi—n de X_k con los otros regresores.

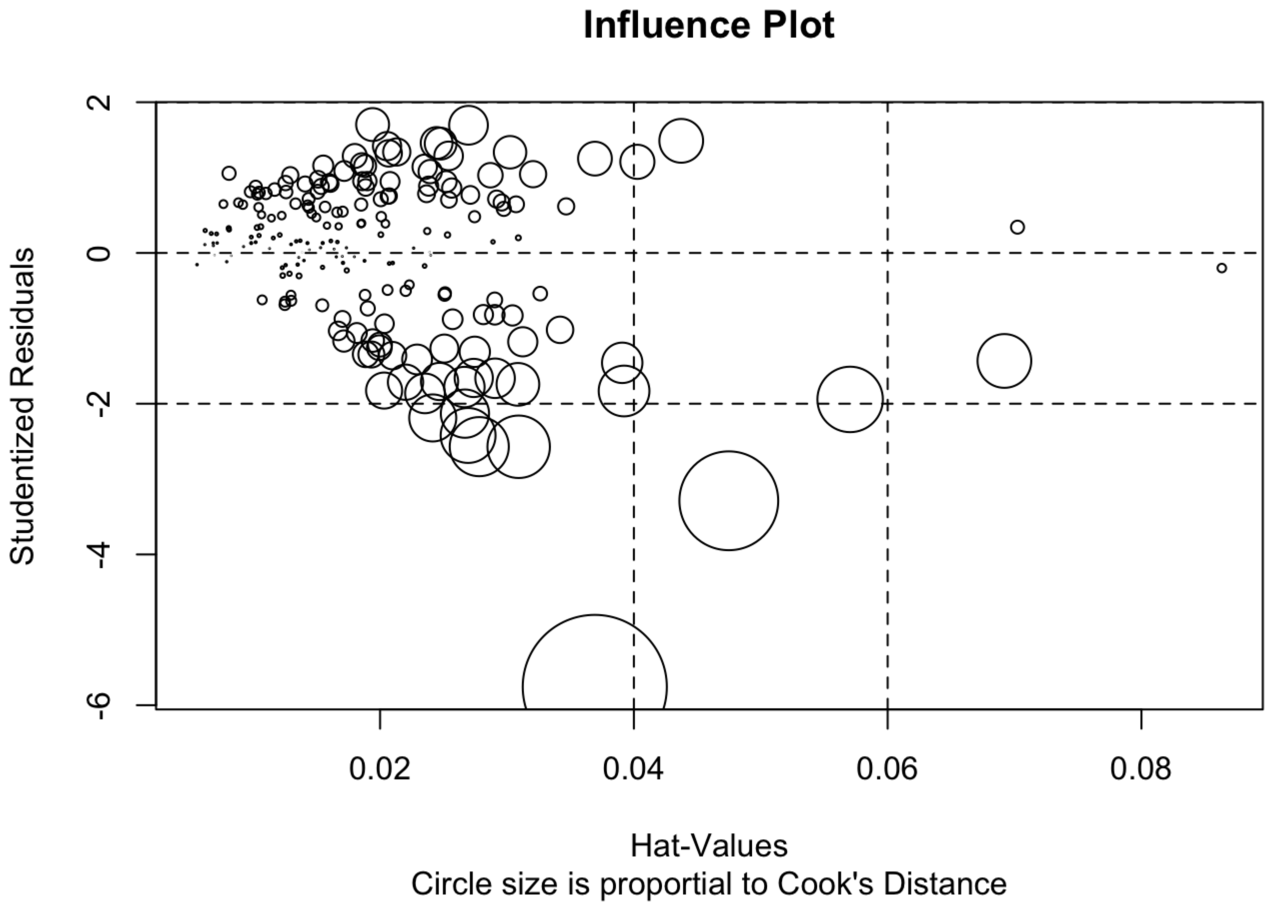
```
# Identifying influential observations

# Cooks Distance D
# identify D values > 4/(n-k-1)
cutoff <- 4/(nrow(mData)-length(regres01$coefficients)-2)
plot(regres01, which=4, cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```




```
# Added variable plots
# add id.method="identify" to interactively identify points
avPlots(regres01, ask=FALSE, id.method="identify")
```




```
# Influence Plot
influencePlot(regres01, id.method="identify", main="Influence Plot",
              sub="Circle size is proportial to Cook's Distance" )
```



Selecci—n de Variables

Comparando modelos

Se pueden comparar modelos `anidados` mediante an‡lisis de varianza (anova). Dos modelos ser‡n anidados si uno de los modelos contiene todo las variables del otro.

El modelo 1 est‡ anidado dentro del modelo 2. Al ser el valor p alto, se considera que el modelo 1 no se puede rechazar.

```
regres02=lm(Sales~TV+Radio,data=mData)
anova(regres02, regres01)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ TV + Radio
## Model 2: Sales ~ TV + Radio + Newspaper
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    197 556.91
## 2    196 556.83  1  0.088717 0.0312 0.8599
```

Los criterios que se utilizan para seleccionar modelos son:

- R^2: no es un buen criterio al aumentar cuando se aumenta el nœmero de regresores.
- R^2 ajustado: mejor. Penaliza la inclusi—n de regresores.
- MallowÕs Cp.

$C_p=\frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ Donde

- RSS es la `residual sum of squares`.
- d es en nœmero de predictores.
- y $\hat{\sigma}^2$ se refiere a la estimaci—n de la varianza de los residuos.

Se selecciona el modelo con menor C_p

- AkaikeÕs Information Criterion(AIC)

$AIC=\frac{1}{n\sigma^2}(RSS + 2d\hat{\sigma}^2)$ Se selecciona el modelo con menor AIC

- SchwarzÕs BIC

$BIC=\frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2)$

Se selecciona el modelo con menor BIC

```
AIC(regres01,regres02)
```

```
##          df      AIC
## regres01  5 782.3622
## regres02  4 780.3941
```

```
BIC(regres01,regres02)
```

```
##          df      BIC
## regres01  5 798.8538
## regres02  4 793.5874
```

MŽtodos de Selecci—n

- Selecci—n Best Subset
- Selecci—n Stepwise
  - Forward Stepwise
  - Backward Stepwise
  - Mixto

Best Subset

Consiste en estimar todas las regresiones posibles con las combinaciones de los p regresores.

Esto es, primero estimamos los p modelos con las variables individuales, todos los modelos ${p \choose 2} =\frac{p(p-1)}{2}$con dos variables, luego los de tres variables y as' hasta que est‡n todas.

El algoritmo ser'a:

> 1.- Consideremos $\mathcal{M}_0$ el modelo nulo, que no contiene regresores. Este modelo predice la variable con su media.
>
> 2.- Para $k = 1, 2,\dots,p$
>
> a. Se estiman ${p \choose k}$ modelos que contienen exactamente k predictores.
> b. Se elige el mejor de estos ${p \choose k}$ modelos y se le llama $\mathcal{M}_k$. Aqui el mejor modelo se selecciona con el menor RSS o su equivalente mayor R^2.
>
> 3.- Se selecciona un œnico modelo de los $\mathcal{M}_0, \dots \mathcal{M}_p$, utilizando

> C_p, AIC, BIC \; — \;\bar R^2

```
library (leaps)
regfit.full=regsubsets(Sales~.-X,mData )
reg.summary=summary(regfit.full)
reg.summary
```

```
## Subset selection object
## Call: regsubsets.formula(Sales ~ . - X, mData)
## 3 Variables  (and intercept)
##           Forced in Forced out
## TV            FALSE      FALSE
## Radio         FALSE      FALSE
## Newspaper     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          TV  Radio Newspaper
## 1  ( 1 ) "*" " "   " "
## 2  ( 1 ) "*" "*"   " "
## 3  ( 1 ) "*" "*"   "*"
```

```
reg.summary$rss
```

```
## [1] 2102.5306  556.9140  556.8253
```

```
reg.summary$cp
```

```
## [1] 544.081354   2.031228   4.000000
```

```
reg.summary$aic
```

```
## NULL
```

```
reg.summary$bic
```

```
## [1] -178.6890 -439.0879 -433.8214
```

Forward Stepwise

Empieza con un modelo que no incluye ningœn regresor y se van a–adiendo regresores de uno en uno.

En cada etapa la variable que m‡s mejora adicional aporta al modelo es incluida.

El algoritmo ser'a:

> 1.- Consideremos \mathcal{M}_0 el modelo nulo, que no contiene regresores.
>
> 2.- Para k = 0, 2,\dots,p-1
>
> a. Se estiman todos losp-k modelos que aumentan de predictores a \mathcal{M}_k con un

> solo predictor adicional.
>
> b. Se elige el mejor de estos $p-k$ modelos y se le llama $\mathcal{M}_k+1$. Aqu' el mejor modelo se selecciona con el menor RSS o su mayor R^2.
>
> 3.- Se selecciona un œnico modelo de los $\mathcal{M}_0, \dots \mathcal{M}_p$, utilizando $C_p, AIC, BIC \; — \;\bar R^2$

Mientras que la selecci—n Best necesita estimar 2^p modelos, la selecci—n Forward Stepwise solo necesita estimar $1+ modelos.

Con 20 variables en un caso ser'an 1.048.576 modelos y en el otro solo 211.

```
library(MASS)

regfit.fwd=regsubsets(Sales~.-X,mData,method ="forward")
summary (regfit.fwd )
```

```
## Subset selection object
## Call: regsubsets.formula(Sales ~ . - X, mData, method = "forward")
## 3 Variables  (and intercept)
##           Forced in Forced out
## TV            FALSE      FALSE
## Radio         FALSE      FALSE
## Newspaper     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: forward
##          TV  Radio Newspaper
## 1  ( 1 ) "*" " "   " "
## 2  ( 1 ) "*" "*"   " "
## 3  ( 1 ) "*" "*"   "*"
```

Backward Stepwise

Empieza con un modelo que incluye todos los regresores y se van eliminando regresores de uno en uno.

En cada etapa la variable que menos mejora adicional aporta al modelo es excluida.

El algoritmo ser'a:

> 1.- Consideremos $\mathcal{M}_p$ el modelo completo, que contiene todos los regresores.
>
> 2.- Para $k = p,p-1,\dots,1$
>
> a. Se estiman todos los k modelos que contienen todos menos un predictor en $\mathcal{M}_k$, par un toal de $k-1$ predictores.
>
> b. Se elige el mejor de estos k modelos y se le llama $\mathcal{M}_k-1$. Aqu' el mejor modelo se selecciona con el menor RSS o su mayor R^2.
>
> 3.- Se selecciona un œnico modelo de los $\mathcal{M}_0, \dots \mathcal{M}_p$, utilizando $C_p, AIC, BIC \; — \;\bar R^2$

Tanto los Forward como los Backward no garantiza la selecci—n del modelo Best.

Se pueden utilizar modelos mixtos.

```
library(MASS)

stepAIC(regres01, direction="backward")
```

```
## Start:  AIC=212.79
## Sales ~ TV + Radio + Newspaper
##
##             Df Sum of Sq    RSS    AIC
## - Newspaper  1      0.09  556.9 210.82
## <none>                    556.8 212.79
## - Radio      1   1361.74 1918.6 458.20
## - TV         1   3058.01 3614.8 584.90
##
## Step:  AIC=210.82
## Sales ~ TV + Radio
##
##         Df Sum of Sq    RSS    AIC
## <none>                556.9 210.82
## - Radio  1    1545.6 2102.5 474.52
## - TV     1    3061.6 3618.5 583.10
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = mData)
##
## Coefficients:
## (Intercept)           TV        Radio
##     2.92110      0.04575      0.18799
```

```
regfit.bwd=regsubsets(Sales~.-X,mData,method ="backward")
summary (regfit.bwd )
```

```
## Subset selection object
## Call: regsubsets.formula(Sales ~ . - X, mData, method = "backward")
## 3 Variables  (and intercept)
##           Forced in Forced out
## TV            FALSE      FALSE
## Radio         FALSE      FALSE
## Newspaper     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: backward
##          TV  Radio Newspaper
## 1  ( 1 ) "*" " "   " "
## 2  ( 1 ) "*" "*"   " "
## 3  ( 1 ) "*" "*"   "*"
```

```
stepAIC(regres01, direction="both")
```

```
## Start:  AIC=212.79
## Sales ~ TV + Radio + Newspaper
##
```

```
##              Df Sum of Sq    RSS    AIC
## - Newspaper  1      0.09  556.9 210.82
## <none>                    556.8 212.79
## - Radio      1   1361.74 1918.6 458.20
## - TV         1   3058.01 3614.8 584.90
##
## Step:  AIC=210.82
## Sales ~ TV + Radio
##
##              Df Sum of Sq    RSS    AIC
## <none>                    556.9 210.82
## + Newspaper  1      0.09  556.8 212.79
## - Radio      1   1545.62 2102.5 474.52
## - TV         1   3061.57 3618.5 583.10
```

```
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = mData)
##
## Coefficients:
## (Intercept)           TV        Radio
##     2.92110      0.04575      0.18799
```

Cross Validation

Es importante separa la muestra en dos partes:

- Training: estimaremos el modelo y obtendremos los valores ajustados que se comparan los valores reales. A partir de aqu' se calculan los residuos. `Training Error`
- Testing: Se predice con el modelo y se comparan las predicciones con los valores reales obteniendo los errores de predicci—n. `Test error`

Validation Set

Consiste en dividir la muestra de forma aleatoria en dos submuestras. Utilizar una para el training (se estima el modelo) y la otra para el testing (se predice el modelo).

Esta selecci—n se realiza repetidamente.

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

```
library(ISLR)
set.seed(250)
numData=nrow(mData)
train=sample(numData ,numData/2)

regres.train =lm(Sales~.-X,mData ,subset =train )
attach(mData)
mean((Sales-predict(regres.train ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 200 rows
```

```
## [1] 2.991664
```

```
set.seed(251)
regres.train2 =lm(Sales~.-X-Newspaper,mData,subset =train )
mean((Sales-predict(regres.train2 ,Auto))[-train ]^2)
```

```
## Warning: 'newdata' had 392 rows but variables found have 200 rows
```

```
## [1] 2.823357
```

Leave-One-Out Cross-Validation

El Leave-One-Out Cross-Validation (LOOCV) consiste en tomar una muestra con todos los datos menos uno. Se estima el modelo y se predice sobre el dato que se ha dejado fuera. Este proceso se repite para los n datos.

En este caso el estimador del `test error` es:

CV_{(n)}=\frac{1}{n} \sum_{i=1}^{n}MSE_i

Es el mŽtodo que menor error comete al estimar el test error. Sin embargo tiene un coste computacional alto.

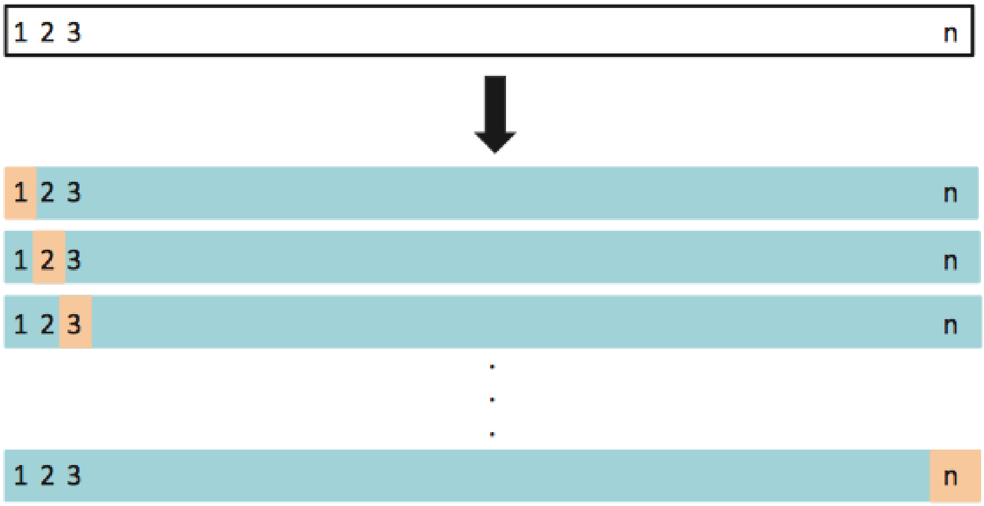


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

```
glm.fit1=glm(Sales~.-X,mData,family = gaussian())
coef(glm.fit1)
```

```
##  (Intercept)           TV        Radio    Newspaper
##  2.938889369  0.045764645  0.188530017 -0.001037493
```

```
library (boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
cv.err =cv.glm(mData,glm.fit1)
cv.err$delta
```

```
## [1] 2.946900 2.946486
```

```
glm.fit2=glm(Sales~.-X-Newspaper,mData,family = gaussian())
cv.err2 =cv.glm(mData,glm.fit2)
cv.err2$delta
```

```
## [1] 2.910676 2.910357
```

K-Fold Cross-Validation

Supone dividir la muestra en k grupos o folds, de aproximadamente igual tama–o. Cada folds es tratado como un conjunto de validaci—n, de tal forma que se estima el modelo con los datos que no est‡n el fold (los otros k-1 folds) y se predicen en el fold.

El procedimiento se repite k veces y se estima el error:

$$CV_{(k)}=\frac{1}{k} \sum_{i=1}^{k}MSE_i$$

El LOOCV es un caso especial del K-fold en el cual k=n.

Los calores habituales para la k suelen ser 5 o 10.

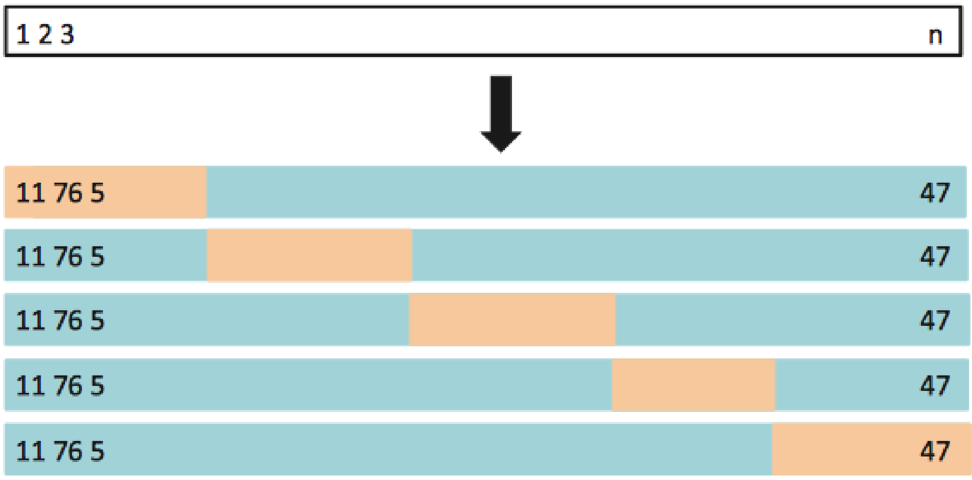


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

```
glm.fit1=glm(Sales~.-X,mData,family = gaussian())


library (boot)
cv.err =cv.glm(mData,glm.fit1,K=10)
cv.err$delta
```

```
## [1] 2.934905 2.926889
```

```
glm.fit2=glm(Sales~.-X-Newspaper,mData,family = gaussian())
cv.err2 =cv.glm(mData,glm.fit2,K=10)
cv.err2$delta
```

```
## [1] 2.891888 2.886227
```

Importancia relativa

ÀCu‡l variable es m‡s œtil o importante para predecir la variable dependiente?

Lo que se hace es calcular los estimadores estandarizados.

```
zData=data.frame(scale(mData))
zlm.fit=lm(Sales~.-X-Newspaper,zData)
coef(zlm.fit)
```

```
##   (Intercept)            TV         Radio
## -8.665545e-17  7.529042e-01  5.349569e-01
```

Un aumento en una desviaci—n t'pica de los gastos en publicidad en TV supone que las ventas se incrementan en 0.752 desviaciones t'picas.

Datos Ausentes

```
library(rminer)
```

```
## Loading required package: kknn
```

```
nelems=function(d) paste(nrow(d),"x",ncol(d))
#missing Data
# missing data example
# since sale does not include missing data, lets
# synthetically create such data:
set.seed(12345) # set for replicability
mData3=mData
N=20 # randomly assign N missing values (NA) to 2nd and 3rd attributes
srow1=sample(1:nrow(mData),N) # N rows
srow2=sample(1:nrow(mData),N) # N rows
mData3[srow1,2]=NA # tv
mData3[srow2,3]=NA # radio
print("Show summary of sales attributes (with NA values):")
```

```
## [1] "Show summary of sales attributes (with NA values):"
```

```
print(summary(mData3[,1:2]))
```

```
##        X                TV
##  Min.   :  1.00   Min.   :  0.70
##  1st Qu.: 50.75   1st Qu.: 74.38
##  Median :100.50   Median :154.05
##  Mean   :100.50   Mean   :147.30
##  3rd Qu.:150.25   3rd Qu.:218.43
##  Max.   :200.00   Max.   :296.40
##                   NA's   :20
```

```
cat("mData3:",nelems(mData3),"\n")
```

```
## mData3: 200 x 5
```

```
cat("NA values:",sum(is.na(mData3)),"\n")
```

```
## NA values: 40
```

```
# 1st method: case deletion
print("-- 1st method: case deletion --")
```

```
## [1] "-- 1st method: case deletion --"
```

```
mData4=na.omit(mData3)
cat("mData4:",nelems(mData4),"\n")
```

```
## mData4: 162 x 5
```

```
cat("NA values:",sum(is.na(mData4)),"\n")
```

```
## NA values: 0
```

```
# 2nd method: average imputation for tv, mode imputation for radio:
# substitute NA values by the mean:
print("-- 2nd method: value imputation --")
```

```
## [1] "-- 2nd method: value imputation --"
```

```
print("original tv summary:")
```

```
## [1] "original tv summary:"
```

```
print(summary(mData3$TV))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.70   74.38  154.00  147.30  218.40  296.40      20
```

```
meanTV=mean(mData3$TV,na.rm=TRUE)
mData5=imputation("value",mData3,"TV",Value=meanTV)
print("mean imputation TV summary:")
```

```
## [1] "mean imputation TV summary:"
```

```
print(summary(mData5$TV))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.70   76.38  147.30  147.30  214.90  296.40
```

```
# substitute NA values by the mode (most common value of mData$radio):
print("original Radio summary:")
```

```
## [1] "original Radio summary:"
```

```
print(summary(mData3$Radio))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.30   10.48   23.75   23.60   36.52   49.60      20
```

```
mData5=imputation("value",mData5,"Radio",Value=(which.max(table(mData$Radio))))
print("mode imputation Radio summary:")
```

```
## [1] "mode imputation Radio summary:"
```

```
print(summary(mData5$Radio))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.30   11.68   20.55   23.24   35.18   49.60
```

```
# 3rd method: hot deck
# substitute NA values by the values found in most similar case (1-nearest neighbor):
print("-- 3rd method: hotdeck imputation --")
```

```
## [1] "-- 3rd method: hotdeck imputation --"
```

```
print("original Radio summary:")
```

```
## [1] "original Radio summary:"
```

```
print(summary(mData3$TV))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.70   74.38  154.00  147.30  218.40  296.40      20
```

```
mData6=imputation("hotdeck",mData3,"TV")
print("hot deck imputation age summary:")
```

```
## [1] "hot deck imputation age summary:"
```

```
print(summary(mData6$TV))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.70   74.38  154.00  147.30  218.40  296.40
```

```
# substitute NA values by the values found in most similar case:
print("original Radio summary:")
```

```
## [1] "original Radio summary:"
```

```
print(summary(mData3$Radio))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    0.30   10.48   23.75   23.60   36.52   49.60      20
```

```
mData6=imputation("hotdeck",mData6,"Radio")
print("hot deck imputation Radio summary:")
```

```
## [1] "hot deck imputation Radio summary:"
```

```
print(summary(mData6$Radio))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.300   9.975  22.400  23.050  35.450  49.600
```

```
cat("mData6:",nelems(mData6),"\n")
```

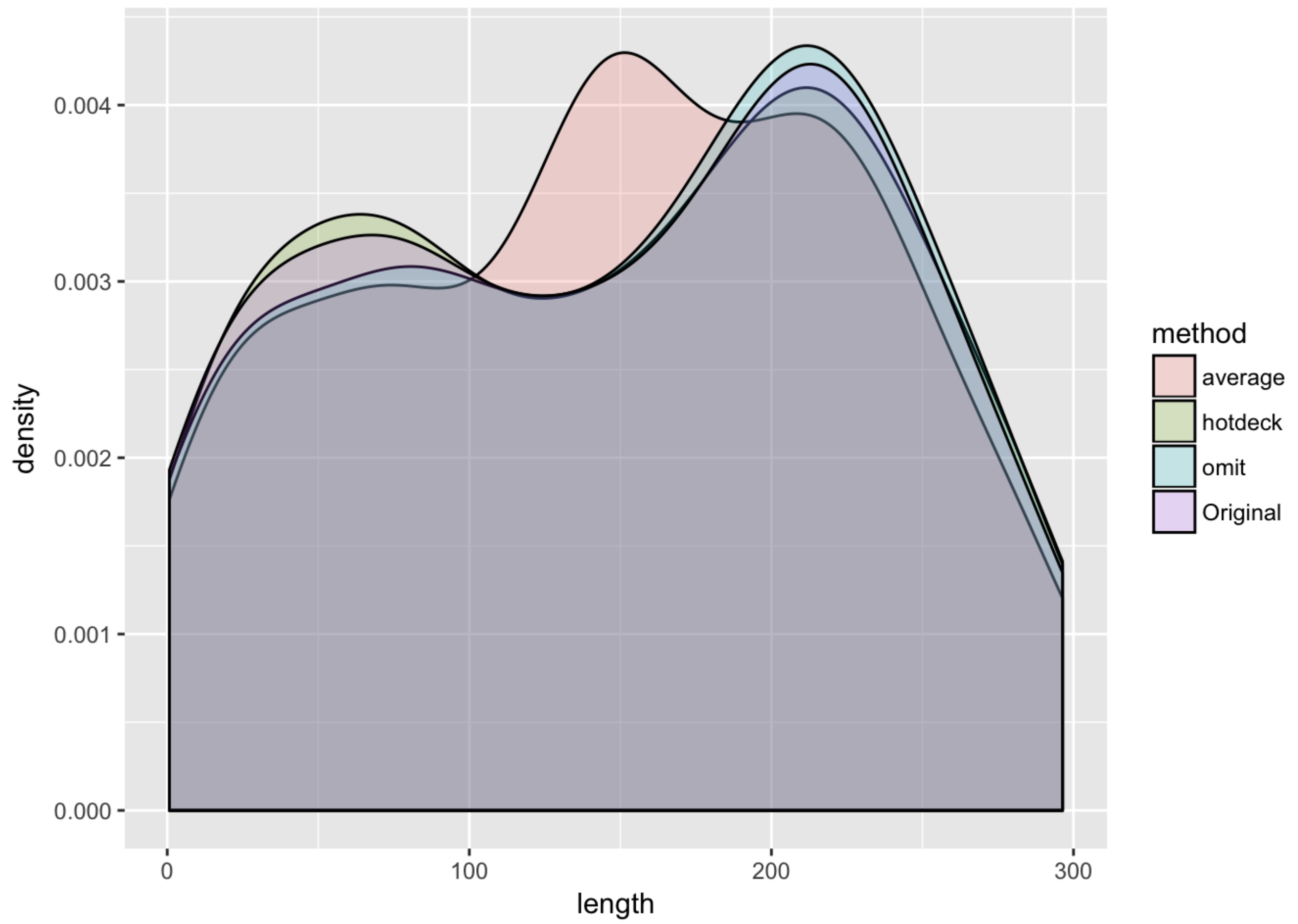
```
## mData6: 200 x 5
```

```
cat("NA values:",sum(is.na(mData6)),"\n")
```

```
## NA values: 0
```

```
# comparison of age densities (mean vs hotdeck):
library(ggplot2)
meth1=data.frame(length=mData4$TV)
meth2=data.frame(length=mData5$TV)
meth3=data.frame(length=mData6$TV)
meth0=data.frame(length=mData3$TV)
meth1$method="omit"
meth2$method="average"
meth3$method="hotdeck"
meth0$method="Original"
all=rbind(meth1,meth2,meth3,meth0)
ggplot(all,aes(length,fill=method))+geom_density(alpha = 0.2)
```

```
## Warning: Removed 20 rows containing non-finite values (stat_density).
```



Ejemplo: Advertising

> Hay importantes preguntas que queremos contestar mediante el modelo de regresi—n:

> - ÀHay una relaci—n entre el presupuesto de publicidad y las Ventas?
> - ÀC—mo de fuerte es la relaci—n entre los gastos de publicidad y las ventas?
> - ÀCu‡l de los los medios contribuye m‡s a las ventas?
> - ÀC—mo de precisos se pueden estimar los efectos de cada medio sobre las ventas?
> - ÀC—mo de preciso se pueden predecir las ventas futuras?
> - ÀEs la relaci—n lineal?
> - ÀHay sinergias entre los diferentes tipos de anuncios?