Regresión lineal

agosto 2017

Resumen

Modelo de regresión lineal

- Asume hay una relación lineal entre las variables.
- Simple y fácilmente interpretable
- Puede aplicarse a transformaciones de las variables, ampliando su campo de aplicación
- Estudiarlos es útil para poder comprender métodos más sofisticados.

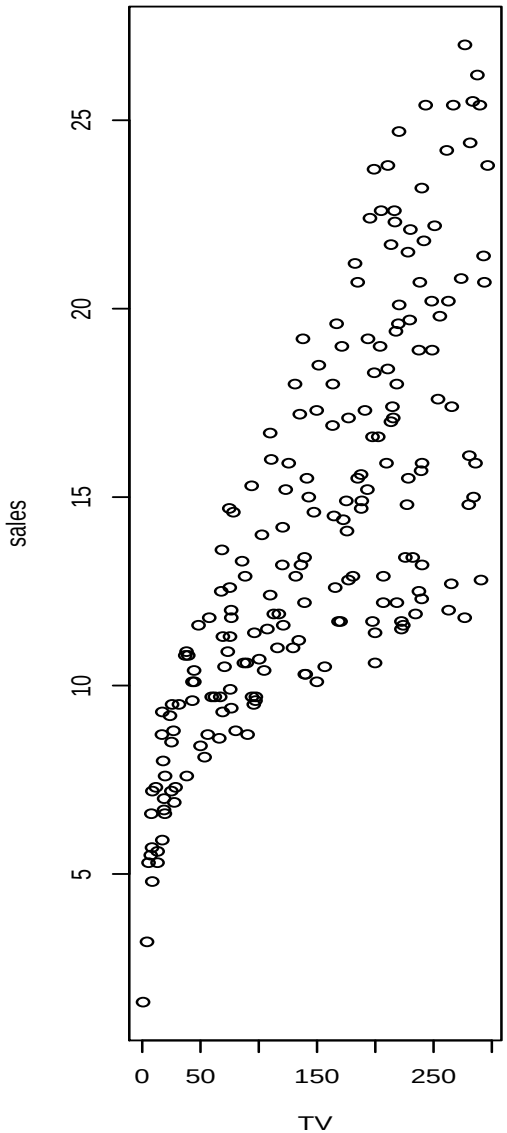
Ejemplo

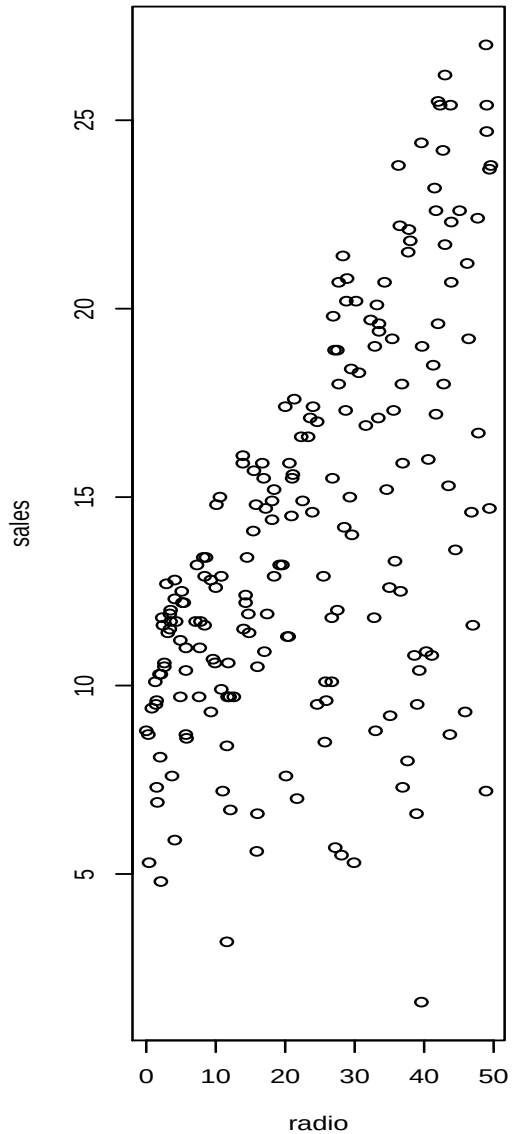
El conjunto de datos *Advertising.csv* contiene las ventas de un producto en 200 mercados, junto con el presupuesto publicitario en tres medios de cumunicación: TV, radio y periódicos.

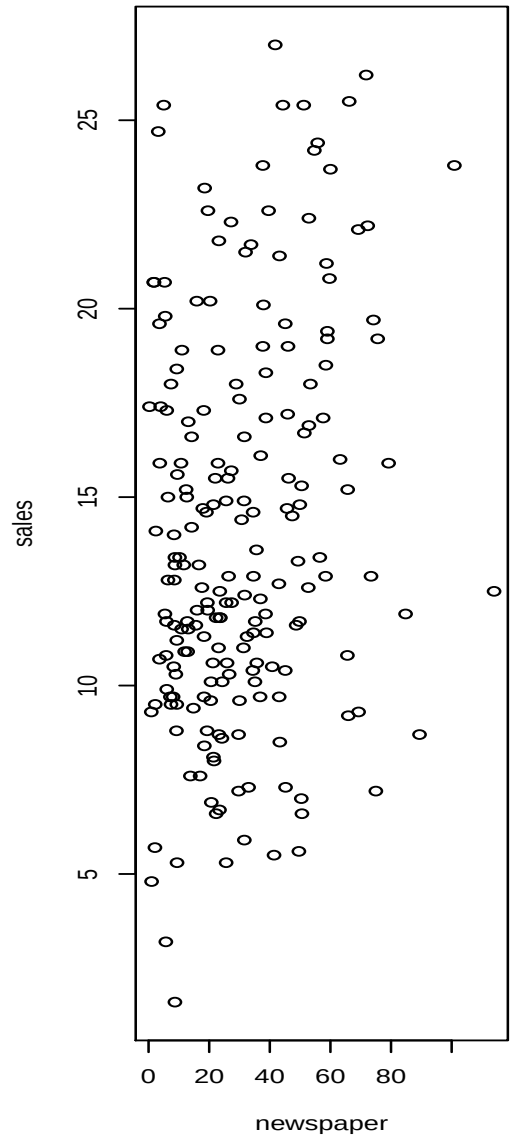
Ejemplo

El conjunto de datos *Advertising.csv* contiene las ventas de un producto en 200 mercados, junto con el presupuesto publicitario en tres medios de cumunicación: TV, radio y periódicos.
Preguntas:

- ¿Hay relación entre el presupuesto en publicidad y las ventas?
- ¿Cuán fuerte es la relación entre ambos?¿Qué medio contribuye más a las ventas?
- ¿Con qué precisión estimamos el efecto de cada medio en las ventas?
- ¿Cómo podemos predecir ventas futuras?
- ¿Es lineal la relación?
- ¿Hay interacción entre la publicidad en diversos medios?


sales
5
10
15
20
25
0
50
150
250
TV


sales
5
10
15
20
25
0
10
20
30
40
50
radio


sales
5
10
15
20
25
0
20
40
60
80
newspaper

Regresión lineal simple

El modelo

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y: respuesta o variable dependiente,
- X: variable explicativa, covariable o variable independiente,
- β_0: ordenada al origen o intercept,
- β_1: pendiente,
- ε: error aleatorio.

En nuestro ejemplo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i: ventas anuales, en miles de pesos, en el i-ésimo mercado.
- X_i: presupuesto anual en publicidad en TV en el i-ésimo mercado.
- β_0: valor medio de las ventas cuando la inversión en publicidad en TV es 0.
- β_1: indica cuánto aumentan en promedio las ventas cuándo la inversión en publicidad aumenta en 1 unidad (1000 pesos).
- ε_i: expresa todo lo que nos falta explicar de las ventas (porque la relación no es exactamente lineal y las ventas dependen de otros factores)

Estimador de mínimos cuadrados

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{argmin} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

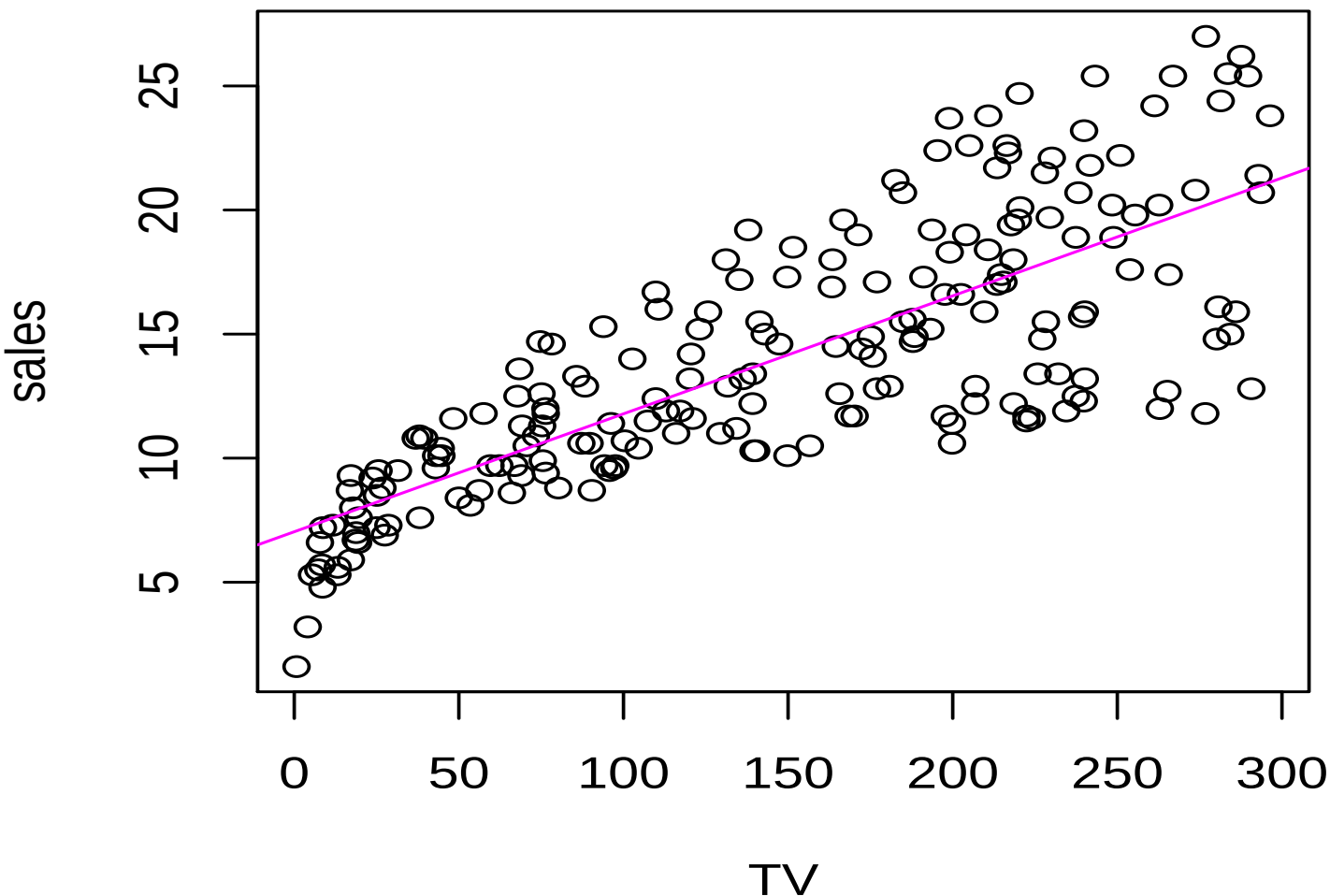
Derivando e igualando a cero se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

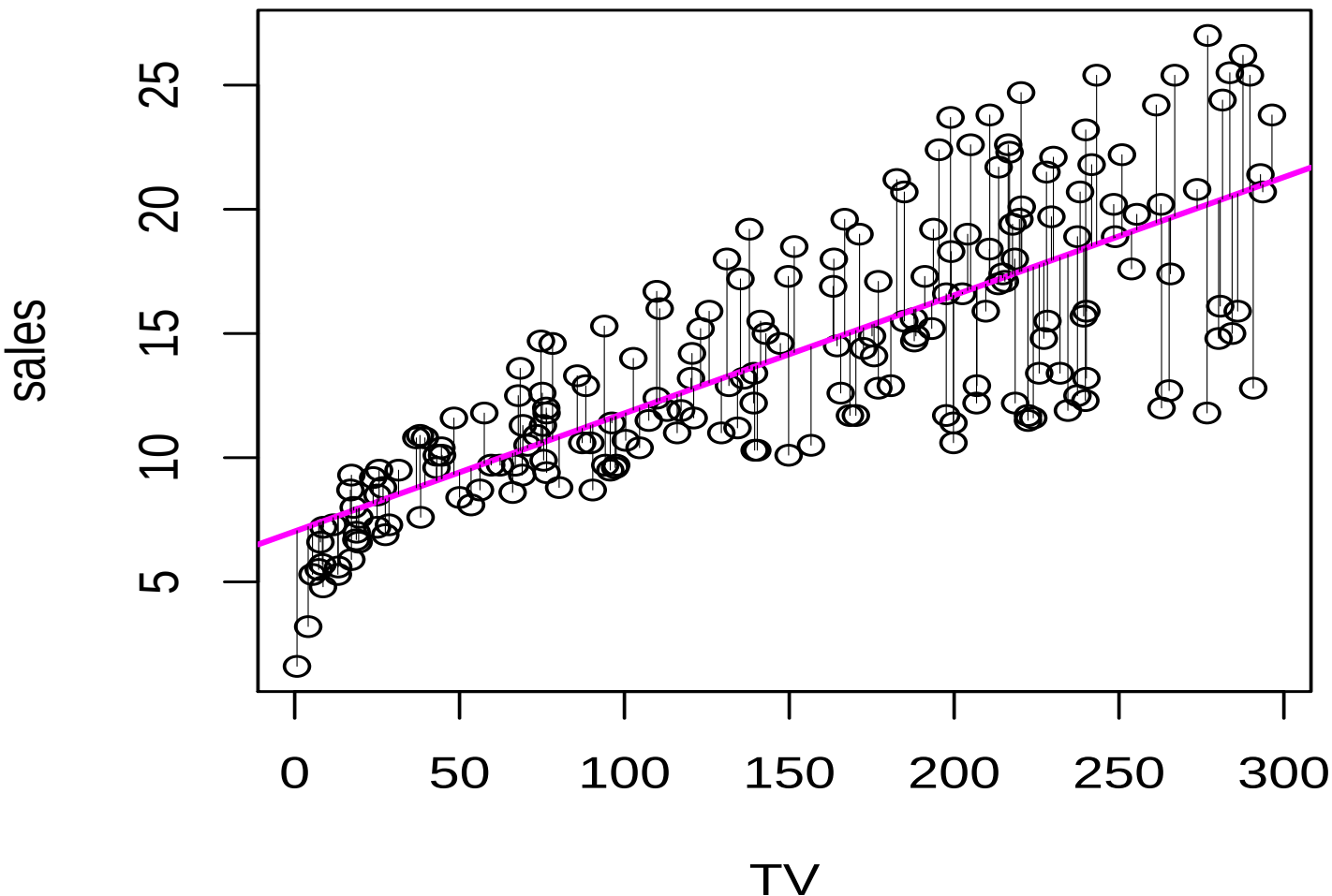
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

donde $\overline{X} = \sum_{i=1}^{n} X_i$ y $\overline{Y} = \sum_{i=1}^{n} Y_i$.

Estimador de mínimos cuadrados



Estimador de mínimos cuadrados



Estimador de mínimos cuadrados

```
> setwd("C:/Users/Marina/Dropbox/lanueva2")
> advdata<-read.csv("advertising.csv",header=TRUE)
> attach(advdata)
> lm(sales~TV)

Call:
lm(formula = sales ~ TV)

Coefficients:
(Intercept)           TV
    7.03259      0.04754
```

```
> lm(sales~TV)

Call:
lm(formula = sales ~ TV)

Coefficients:
(Intercept)           TV
    7.03259      0.04754
```

Supuestos

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Supuestos sobre ϵ_i

1. Normalidad
2. Independencia
3. Media cero
4. Varianza constante

Es decir, asumimos que

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ independientes } .$$

Propiedades del estimador de mínimos cuadrados

Bajo los supuestos anteriores, el estimador de mínimos cuadrados es

- insesgado
- consistente
- asintóticamente normal

Además, bajo los supuestos anteriores, el estimador de mínimos cuadrados coincide con el estimador de máxima verosimilitud.

Propiedades del estimador de mínimos cuadrados

Bajo los supuestos anteriores, el estimador de mínimos cuadrados es

- insesgado
- consistente
- asintóticamente normal

Además, bajo los supuestos anteriores, el estimador de mínimos cuadrados coincide con el estimador de máxima verosimilitud.

Para que los tests e intervalos de confianza que da el R sean válidos, es necesario que se cumplan los supuestos.

Propiedades del estimador de mínimos cuadrados

Bajo los supuestos anteriores, el estimador de mínimos cuadrados es

- insesgado
- consistente
- asintóticamente normal

Además, bajo los supuestos anteriores, el estimador de mínimos cuadrados coincide con el estimador de máxima verosimilitud.

Para que los tests e intervalos de confianza que da el R sean válidos, es necesario que se cumplan los supuestos.

Si no se cumple el supuesto de normalidad pero la muestra es grande, los intervalos y tests valen en forma aproximada.

Distribución de los estimadores de mínimos cuadrados cuando los errores son normales

Se puede probar que $\hat{\beta}_0$ y $\hat{\beta}_1$ son normales con

$$E(\hat{\beta}_0) = \beta_0 \qquad SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2} \right]$$

y

$$E(\hat{\beta}_1) = \beta_1 \qquad SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

Como σ^2 es deconocido se estima por:

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2}{n-2}$$

$$\hat{SE}(\hat{\beta}_0) = s^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2} \right] \text{ y } \hat{SE}(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

Se puede probar que

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{SE}(\hat{\beta}_0)} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

Intervalos de confianza

Intervalo de confianza para β_0

$$\left[\hat{\beta}_0 - t_{\alpha/2,n-2} SE(\hat{\beta}_0), \widehat{\beta}_0 + t_{\alpha/2,n-2} SE(\hat{\beta}_0)\right]$$

Intervalo de confianza para β_1

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-2} SE(\hat{\beta}_1), \widehat{\beta}_1 + t_{\alpha/2,n-2} SE(\hat{\beta}_1)\right]$$

En nuestro ejemplo

```
> ajusteTV<-lm(sales~TV)
> confint(ajusteTV)
                 2.5 %     97.5 %
(Intercept) 6.12971927 7.93546783
TV          0.04223072 0.05284256
```

Si la inversión en publicidad en TV se redujera a 0, las ventas anuales estarían entre 6129 y 7935 unidades anuales.
Si se aumenta en 1000 la inversión en publicidad en TV, se espera que las ventas aumenten entre 43 y 45 unidades anuales.

En nuestro ejemplo

```
> ajusteTV<-lm(sales~TV)
> confint(ajusteTV)
                 2.5 %     97.5 %
(Intercept) 6.12971927 7.93546783
TV          0.04223072 0.05284256
```

Si la inversión en publicidad en TV se redujera a 0, las ventas anuales estarían entre 6129 y 7935 unidades anuales.
Si se aumenta en 1000 la inversión en publicidad en TV, se espera que las ventas aumenten entre 43 y 45 unidades anuales.

Estos intervalos también se pueden calcular mirando el `summary`

```
> summary(ajusteTV)

Call:
lm(formula = sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,        Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Intervalo de confianza para β_0:

$$[7.032594 - qt(1-\alpha/2, 198) * 0.457843, 7.032594 + qt(1-\alpha/2, 198) * 0.457843]$$

Intervalo de confianza para β_1:

$$[0.047537 - qt(1-\alpha/2, 198) * 0.002691, 0.047537 + qt(1-\alpha/2, 198) * 0.002691]$$

Tests de hipótesis

Test para la pendiente

$$H_0 : \beta_1 = 0 \quad \text{vs } H_1 : \beta_1 \neq 0$$

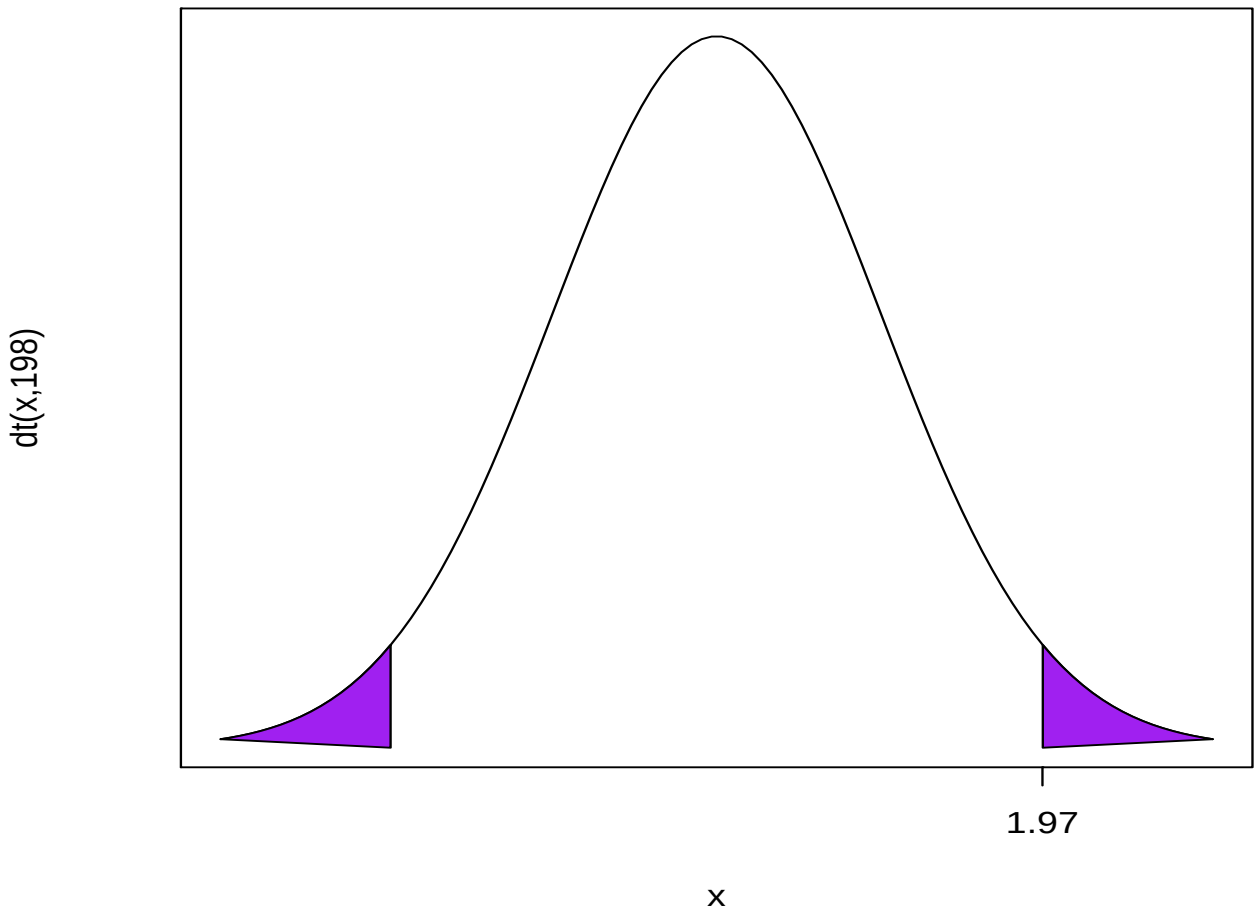
Estadístico del test

$$T = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \sim t_{n-2} \text{ bajo } H_0$$

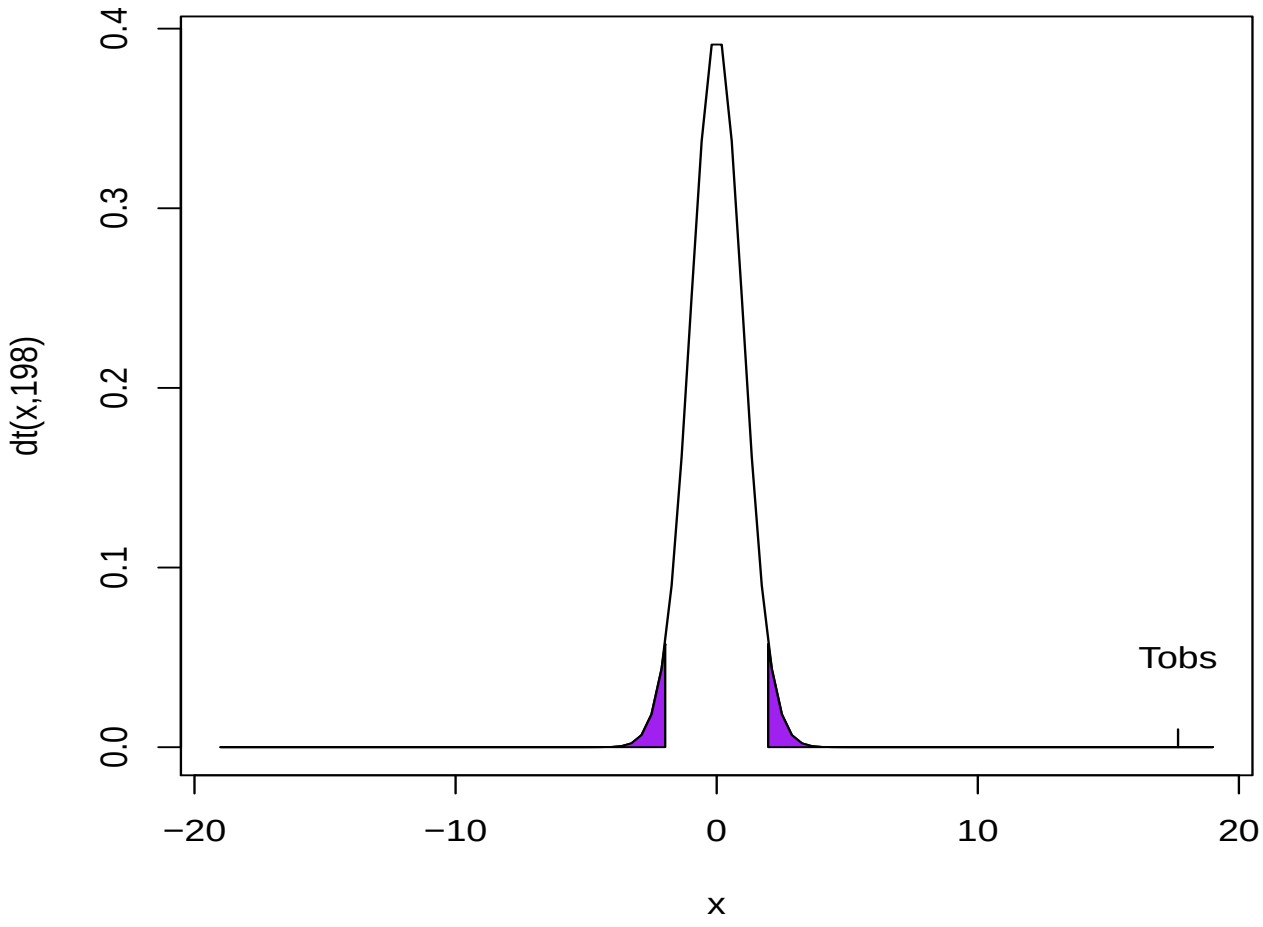
Rechazo H_0 si $|T| > t_{n-2\frac{\alpha}{2}}$

p-valor: $2P(|T| \geq |T_{\text{obs}}|)$
En nuestro caso, $Tobs = 17.67$ y p-valor$< 2e-16$

Región de rechazo de nivel 0.05



Región de rechazo de nivel 0.05 y T observado



Estos datos dan fuerte evidencia de que la inversión en publicidad en TV influye en las ventas.

Estos datos dan fuerte evidencia de que la inversión en publicidad en TV influye en las ventas.

Si la inversión en publicidad en TV no influyera en las ventas, la probabilidad de observar un valor de T tan extremo como este, o más, sería menor a 10^{-16}.

Estos datos dan fuerte evidencia de que la inversión en publicidad en TV influye en las ventas.

Si la inversión en publicidad en TV no influyera en las ventas, la probabilidad de observar un valor de T tan extremo como este, o más, sería menor a 10^{-16}.

En otras palabras,

Si la inversión en publicidad en TV no influyera en las ventas, la probabilidad de que el valor estimado de la pendiente este tan lejos de cero como el que se observa con estos datos,o más, sería menor a 10^{-16}.

Evaluando la bondad del ajuste.

Luego de rechazar la hipótesis nula de que la pendiente es cero nos preguntamos ¿Cuán bien se ajusta el modelo a los datos?
Recursos para responder a esta pregunta:

- R^2
- Error estandar residual

Error estandar residual

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \quad \text{Suma de cuadrados residual}$$

$$s = \sqrt{\frac{RSS}{n-2}} \quad \text{Error estandar residual}$$

También se lo nota RSE

En nuestro ejemplo $s = 3.26$.

Esto significa que las ventas se alejan de la recta de regresión aprox en 3.26 ($3260) unidades en promedio.

Las predicciones de futuras ventas basadas en la publicidad en TV van a diferir de la realidad en $3260 en pomedio.

Estadístico R^2

Es una proporción.

Más fácil de interpretar que el *RSE*

$$R^2 = \frac{TSS - RSS}{TSS} \text{ donde } TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

TSS (suma de cuadrados total) mide la variabilidad total de y.
RSS mide la variabilidad de y que no es explicada por el modelo de regesión.
$TSS - RSS$ indica la cantidad de variabilidad de y que es explicada por el modelo de regresión.
R^2 mide la proporción de la variabilidad de y que es explicada por el modelo de regresión.

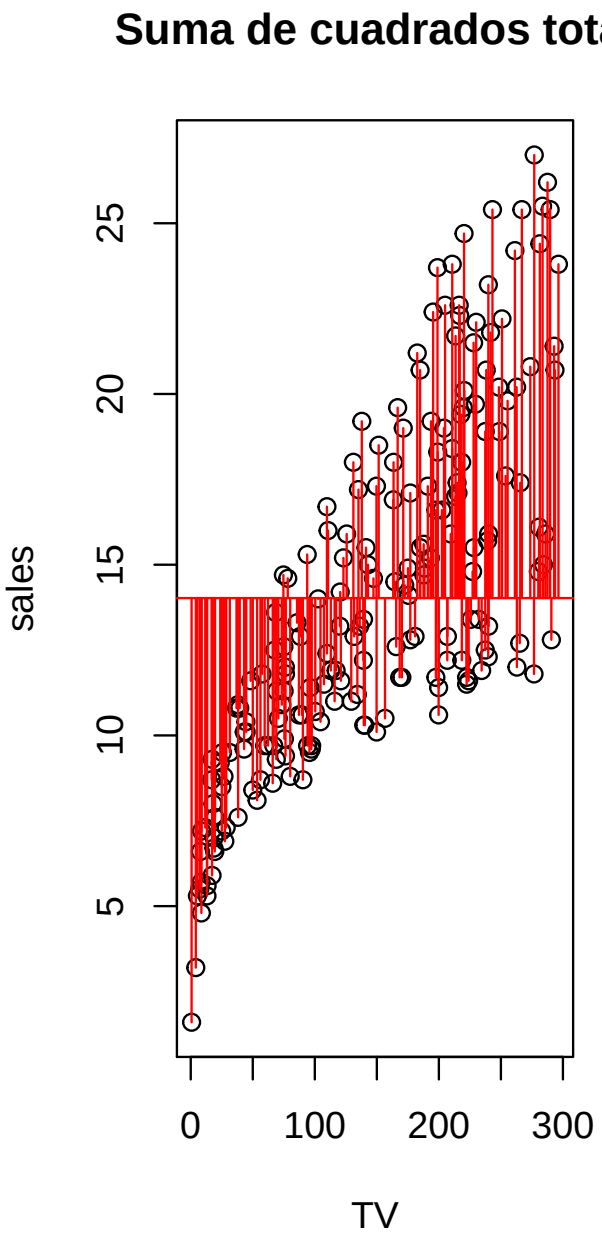
Siempre vale que $0 \leq R^2 \leq 1$.

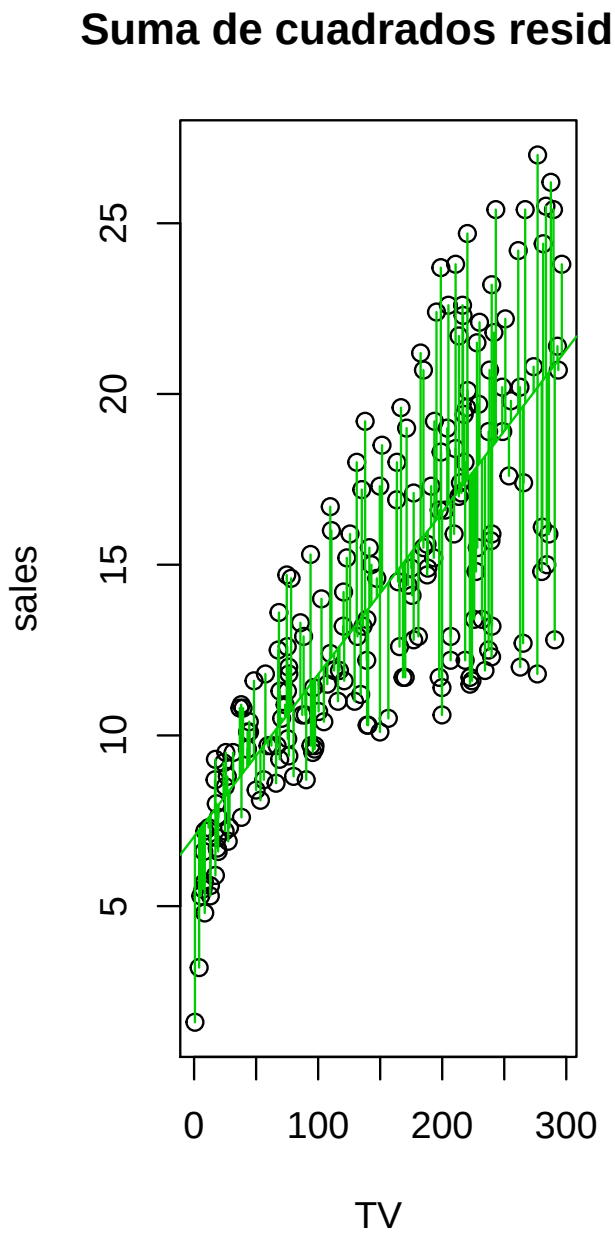
En nuestro ejemplo $R^2 = 0.61$.

Aproximadamente el 61% de la variabilidad total en las ventas puede explicarse por la inversión en publicidad en TV.

Visualización en R

```
> par(mfrow=c(1,2))
> plot(TV,sales,main="Suma de cuadrados total")
> abline(mean(sales),0,col=2)
> segments(TV, sales, TV, mean(sales),col=2)
> plot(TV,sales,main="Suma de cuadrados residual")
> abline(lm(sales~TV)$coefficients,col=3)
> segments(TV, sales, TV, predict(lm(sales~TV)),col=3)
```


Suma de cuadrados total
sales
TV
0
100
200
300
5
10
15
20
25


Suma de cuadrados residual
sales
TV
0
100
200
300
5
10
15
20
25

Tenemos 2 variables más que podrían ser útiles para predecir las ventas: la inversión en publicidad en radio y periódicos.

```
> radiolm<-lm(sales~radio)
> summary(radiolm)

Call:
lm(formula = sales ~ radio)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7305  -2.1324   0.7707   2.7775   8.1810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,        Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
> newslm<-lm(sales~newspaper)
> summary(newslm)

Call:
lm(formula = sales ~ newspaper)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2272  -3.3873  -0.8392   3.5059  12.7751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,        Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Modelo de regresión lineal múltiple

En nuestro ejemplo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

- Y_i: ventas anuales, en miles de pesos, en el i-ésimo mercado.
- X_{1i}: presupuesto anual en publicidad en TV en el i-ésimo mercado.
- X_{2i}: presupuesto anual en publicidad en radio en el i-ésimo mercado.
- X_{3i}: presupuesto anual en publicidad en diarios en el i-ésimo mercado.
- ε_i: expresa todo lo que nos falta explicar de las ventas

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

- β_0: valor medio de las ventas cuando la inversión en publicidad en TV es 0.
- β_1: indica cuánto aumentan en promedio las ventas cuándo la inversión en publicidad en TV aumenta 1000 pesos y se mantiene fija la inversión en radio y diarios,
- β_2: indica cuánto aumentan en promedio las ventas cuándo la inversión en publicidad en radio aumenta en 1000 pesos y se mantiene fija la inversión en TV y diarios .
- β_3: indica cuánto aumentan en promedio las ventas cuándo la inversión en publicidad en diarios aumenta en 1000 pesos y se mantiene fija la inversión en radio y TV .

En general
$\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ vectores aleatorios independientes

- Y_i: respuestas
- $\mathbf{X}_i$: covariables o variables explicativas $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})'$

$$Y_i \quad = \quad \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \ ,$$

En general
$\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ vectores aleatorios independientes

- Y_i: respuestas
- $\mathbf{X}_i$: covariables o variables explicativas $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})'$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \ , \\ Y_i &= \beta_0 + \beta_1 \, log(Z_i) + \cdots + \beta_p W_i^2 + \epsilon_i \ , \end{aligned}$$

En general
$\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ vectores aleatorios independientes

- Y_i: respuestas
- $\mathbf{X}_i$: covariables o variables explicativas $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})'$

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \ , \\
Y_i &= \beta_0 + \beta_1 \, log(Z_i) + \cdots + \beta_p W_i^2 + \epsilon_i \ , \\
Y_i &= \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i \rightarrow \text{ Lineal en } \boldsymbol{\beta}
\end{aligned}
$$

En general
$\{(\mathbf{X}_i, Y_i) : 1 \leq i \leq n\}$ vectores aleatorios independientes

- Y_i: respuestas
- $\mathbf{X}_i$: covariables o variables explicativas $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})'$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i \ , \\ Y_i &= \beta_0 + \beta_1 \, log(Z_i) + \cdots + \beta_p W_i^2 + \epsilon_i \ , \end{aligned}$$

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \epsilon_i \rightarrow \boldsymbol{\beta} \text{ Parámetro a estimar}$$

Estimador de Mínimos Cuadrados

$$\widehat{\boldsymbol{\beta}} \quad = \quad \underset{\mathbf{b}}{argmin} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \mathbf{b})^2$$

Estimador de Mínimos Cuadrados

$$\widehat{\boldsymbol{\beta}} = \underset{\mathbf{b}}{argmin} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \mathbf{b})^2$$

Derivando e igualando 0, tenemos que $\widehat{\boldsymbol{\beta}}$ es solución del sistema

$$\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \mathbf{b}) \mathbf{X}_i = 0$$

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{Y} \quad \text{Ecuaciones Normales}$$

donde $\mathbf{X}$ es la matriz con filas $\mathbf{X}_i$ e $\mathbf{Y} = (Y_1, \ldots, Y_2)$.
Por lo tanto, si $\mathbf{X}^T \mathbf{X}$ es inversible,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y}$$

Supuestos y propiedades.

Igual que en regresión simple.

Supuestos sobre ϵ_i

1. Normalidad
2. Independencia
3. Media cero
4. Varianza constante

Es decir, asumimos que

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ independientes } .$$

Valen las mismas propiedades que en el modelo de regresión simple.

```
> advertisinglm<-lm(sales~TV+radio+newspaper)
> summary(advertisinglm)

Call:
lm(formula = sales ~ TV + radio + newspaper)

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
```

¿Porqué la regresión lineal múltiple sugiere que no hay relación entre las ventas y la publicidad en diarios y la regresión simple dice lo contrario?

```
> cor(advdata)
```

	TV	radio	newspaper	sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
radio	0.05480866	1.00000000	0.35410375	0.5762226
newspaper	0.05664787	0.35410375	1.00000000	0.2282990
sales	0.78222442	0.57622257	0.22829903	1.0000000

La correlación entre las variables `radio` y `newspaper` es 0.35.

Se tiende a inventir más en diarios en los mercados en los que se invierte más en radio.

Algunas preguntas importantes

1. ¿Es significativa la relación entre las variables exlicativas y la respuesta?
2. ¿Conviene incluir a todas las variables en el modelo, o sólo algunas son útiles?
3. ¿Cuán bien ajusta el modelo a esta conjunto de datos ?
4. Dado un valor de las covariables, ¿cómo predecimos la respuesta y cuán precisa es la predicción?

1- ¿Es significativa la relación entre las variables explicativas y la respuesta? Test F

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad \text{vs} \qquad H_1 : \exists i / \beta_i \neq 0$$

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

Recuerdo: $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ y $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

F tiene una distribución F de Fisher-Snedecor

$$F \sim F_{p,n-p-1}$$

Rechazo H_0 si $F > f_{p,n-p-1,1-\alpha}$

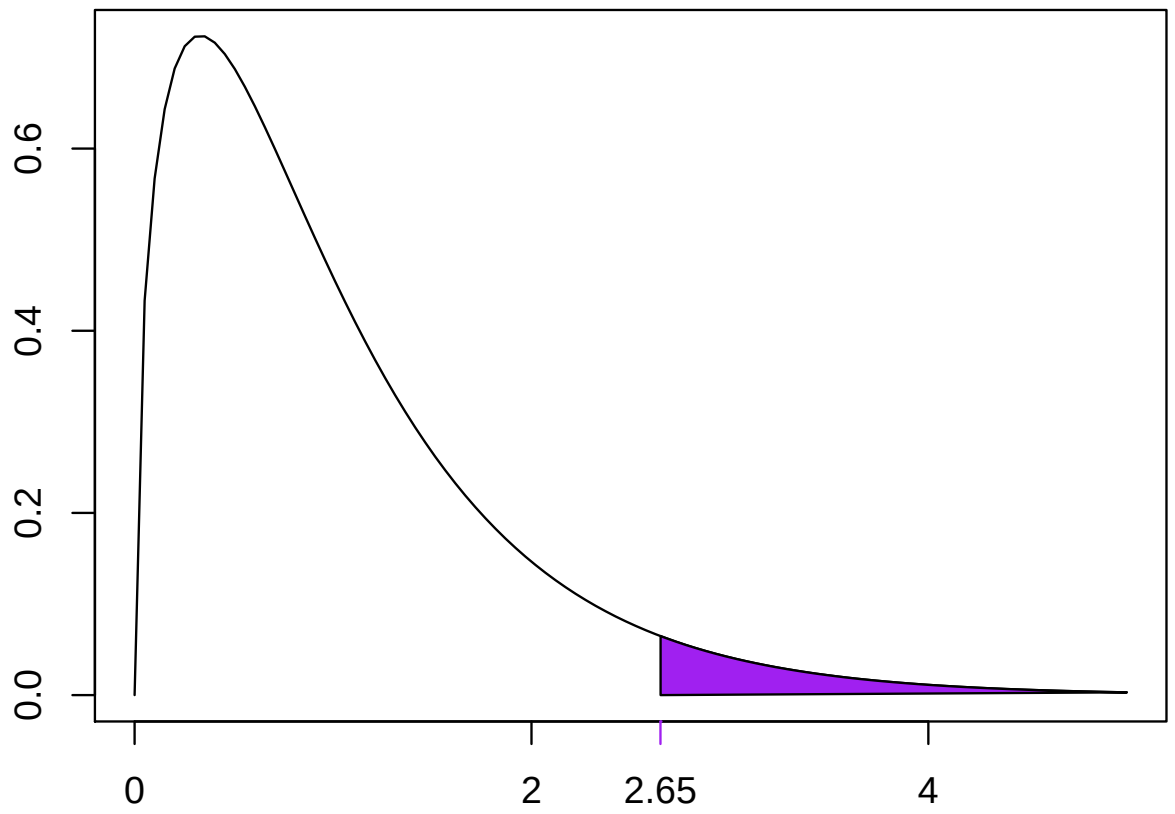
Región de rechazo de nivel 0.05 del test F

```
> par(mfrow=c(1,1))
> x<-seq(0,5,length=100)
> y<-df(x,df1=3,df2=196)
> plot(x,y,type="l")
> alpha<-0.05
> falpha<-qf(1-alpha,df1=3,df2=196)
> falpha
```

```
[1] 2.650677
```

```
> polygon(c(x[x>falpha],falpha,falpha ),
+         c(y[x>falpha],0,df(falpha,df1=3,df2=196)),
+         col="purple")
```

Región de rechazo de nivel 0.05 del test F



p-valor del test F para los datos de `advertising`

```
> fobs<-570.3
> pvalor<-1-pf(fobs,df1=3,df2=196)
> pvalor
[1] 0
```

Conclusión del test: Al menos uno de los β_i es distinto de cero. Es decir, la publicidad en al menos uno de los tres medios influye en las ventas.

¿Qué son los p-valores que aparecen asociados a cada covariable en la salida de R?

Nos dicen si cada covariable está relacionada con la respuesta, después de ajustar el modelo con las otras covariables.

En el ejemplo, indican que `TV` y `radio` están relacionadas con las ventas pero no hay evidencia de que `newspaper` esté asociada con las ventas en presencia de las otras dos. Es decir, no hay evidencia de que predecir as ventas usando las tres covariables sea mejor que predecir las ventas usando sólo `TV` y `radio`.

2. ¿Conviene incluir a todas las variables en el modelo, o sólo algunas son útiles?

Métodos de selección de variables

1. **Selección del mejor subconjunto** Se ajustan todos los modelos posibles y se comparan los ajustes.
2. **Selección Forward** se empieza con la variable que mejor ajusta y se van agregando de a una la que más reduce el RSE.
3. **Selección Backward** se comienza con el modelo completo, se va sacando la variable que menos aporta al modelo.
4. **Selección Mixta** Es una combinación de los dos.

1-Selección del mejor subconjunto

- Se ajusta un modelo de regresión lineal a cada subconjunto de variables.
- Luego comparamos los ajustes de todos los modelos e identificamos cuál es mejor.
- En nuestro ejemplo habría que ajustar 8 modelos.

```
lm(sales~1)
lm(sales~newspaper)
lm(sales~radio)
lm(sales~TV)
lm(sales~newspaper+radio)
lm(sales~radio+TV)
lm(sales~TV+newspaper)
lm(sales~radio+TV+newspaper)
```

En un conjunto de datos con p covariables, la cantidad de modelos que hay que ajustar es 2^p.

Por motivos computacionales la selección del mejor subconjunto no puede ser aplicada cuando p es grande.

Además hay problemas estadísticos: Si p es grande, puede ocurrir que encontremos modelos que se ajusten muy bien al conjunto de datos que se usó para estimar pero describan muy mal un nuevo conjunto de datos. Esto se llama sobreajuste u *overfitting.*

Selección forward

Se comienza con un modelo sin ninguna covariable, sólo el intercept.

Se agregan las covariables, una a una, hasta que todas están en el modelo.

En cada paso, se agrega la covariable que provoca la mayor mejora en el modelo.

Selección backward

Se comienza con el modelo completo, con todas las covariables.

Se retiran las covariables, una a una, hasta que queda sólo el intercept.

En cada paso, se retira la covariable que menos aporta al modelo.

Selección mixta

Las variables se agregan al modelo secuencialmente, como en el método forward.

Sin embargo, después de agregar una covariable, el método también puede retirar una covariable que ya no aporte mejoras al modelo

```
> library(leaps)
> advforward<-regsubsets(sales~TV+
radio+newspaper,data=advdata,method="forward")
> summary(advforward)$which
  (Intercept)   TV radio newspaper
1        TRUE TRUE FALSE     FALSE
2        TRUE TRUE  TRUE     FALSE
3        TRUE TRUE  TRUE      TRUE
> summary(advforward)$cp
[1] 544.081354   2.031228   4.000000
> summary(advforward)$bic
[1] -178.6890 -439.0879 -433.8214
> summary(advforward)$adjr2
[1] 0.6099148 0.8961505 0.8956373
```

```
> advbackward<-regsubsets(sales~TV+radio+
newspaper,data=advdata,method="backward")
> summary(advbackward)$which
  (Intercept)   TV radio newspaper
1        TRUE TRUE FALSE      FALSE
2        TRUE TRUE  TRUE      FALSE
3        TRUE TRUE  TRUE       TRUE
> summary(advbackward)$cp
[1] 544.081354   2.031228   4.000000
> summary(advbackward)$bic
[1] -178.6890 -439.0879 -433.8214
> summary(advbackward)$adjr2
[1] 0.6099148 0.8961505 0.8956373
```

```
> advmixto<-regsubsets(sales~TV+radio+
newspaper,data=advdata,method="seqrep")
> summary(advmixto)$which
  (Intercept)   TV radio newspaper
1        TRUE TRUE FALSE      FALSE
2        TRUE TRUE  TRUE      FALSE
3        TRUE TRUE  TRUE       TRUE
> summary(advmixto)$cp
[1] 544.081354   2.031228   4.000000
> summary(advmixto)$bic
[1] -178.6890 -439.0879 -433.8214
> summary(advmixto)$adjr2
[1] 0.6099148 0.8961505 0.8956373
```

¿Cómo elijo cuál es el mejor modelo?

¿Puedo comparar los RSS y R^2 de los diferentes modelos?

¿Cómo elijo cuál es el mejor modelo?

¿Puedo comparar los RSS y R^2 de los diferentes modelos?

Sólo si los modelos tienen la misma cantidad de variables

¿Cómo elijo cuál es el mejor modelo?

¿Puedo comparar los RSS y R^2 de los diferentes modelos?

Sólo si los modelos tienen la misma cantidad de variables

- RSS siempre disminuye cuando agrego variables y R^2 siempre aumenta
- RSS pequeño y R^2 grande indican que el modelo ajusta bien a los datos con los que fue entrenado, es decir a los datos que se usaron para estimar los parámetros.
- Pero lo más importante es obtener un modelo que haga buenas predicciones con nuevos datos.

- $C_p = RSS/s^2 + 2d - n$
- $AIC = (RSS/s^2 + 1 + 2d)/n$
- $BIC = (RSS + s^2 + \log(n)2ds^2)/n$
- R^2ajustado$= 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

RSS se calcula con el modelo con d predictoras, mientras que s^2 se calcula con todas las disponibles.

Se puede probar que C_p es un estimador insesgado del error de testeo.
En regresión lineal C_p y AIC elijen el mismo modelo.
BIC penaliza más severamente a los modelos on muchas variables.
Cp, AIC y BIC tienen justificaciones teóricas que se basan en argumentos asintóticos (n grande).
R^2ajustado no tiene una motivación teórica tan clara.

3. ¿Cuán bien ajusta el modelo a esta conjunto de datos ? R^2 y RSE.

Se calculan y se itepretan de la misma manera que en regresión simple

En los datos de `Advertising`:

- Modelo con las tres variables explicativas: $R^2 = 0.8972$.
- Modelo con solo `TV` y `radio`: $R^2 = 0.89719$.
- Modelo con solo `TV`: $R^2 = 0.61$

R^2 siempre aumenta cuando agrego variables explicativas.

En los datos de `Advertising`:

- Modelo con las tres variables explicativas: $RSE = 1.686$.
- Modelo con solo `TV` y `radio`: $RSE = 1.681$.
- Modelo con solo `TV`: $RSE = 3.26$

Análisis de residuos

i-ésimo residuo: $r_i = y_i - \hat{y}_i$. Es un estimador de ε_i.

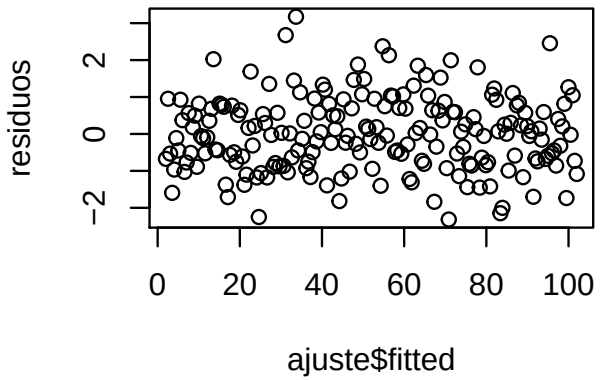
Se puede probar que, si se cumplen los supuestos, la correlación entre los residuos y los valores ajustados es 0.

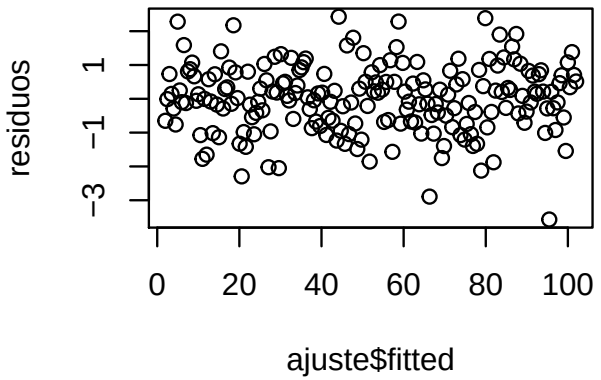
Si el modelo se cumple, al graficar los residuos versus los valores ajustados deberíamos ver una nube de puntos sin estructura alrededor de la recta $y = 0$.

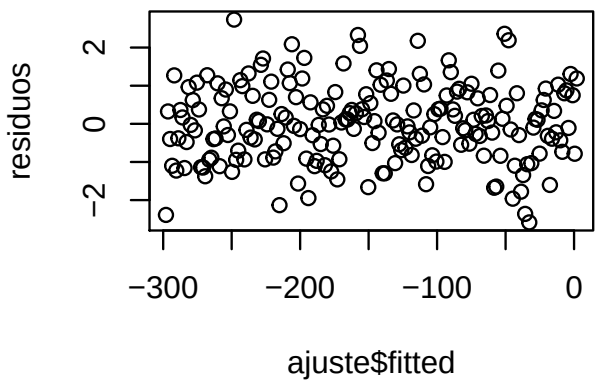
Gráficos de residuos estandarizados cuando el ajuste es correcto

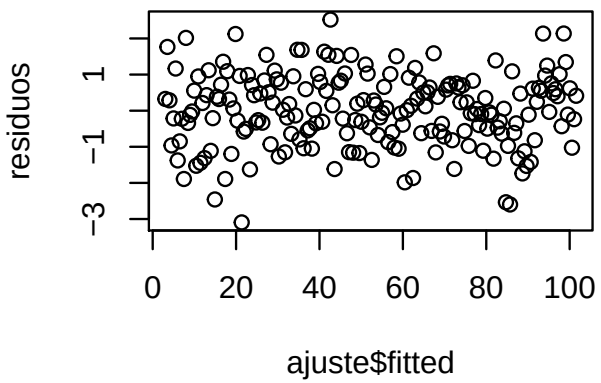
```
> par(mfrow=c(2,2))
> x<-seq(0,100,length=200)
> y<-2+x+rnorm(200,0,1)
> ajuste<-lm(y~x)
> plot(ajuste$fitted,rstandard(ajuste),ylab="residuos")
```

Gráficos de residuos estandarizados (ajuste correcto)



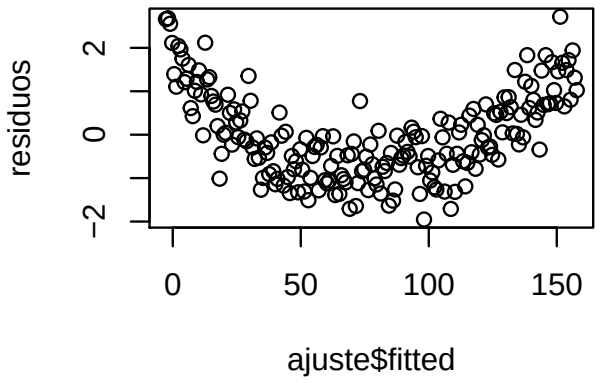




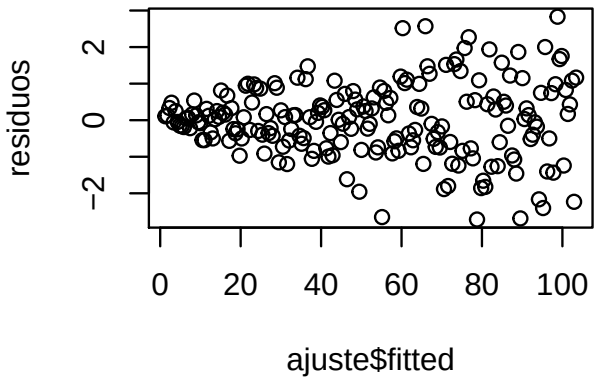


Gráficos de residuos (ajustes con problemas)

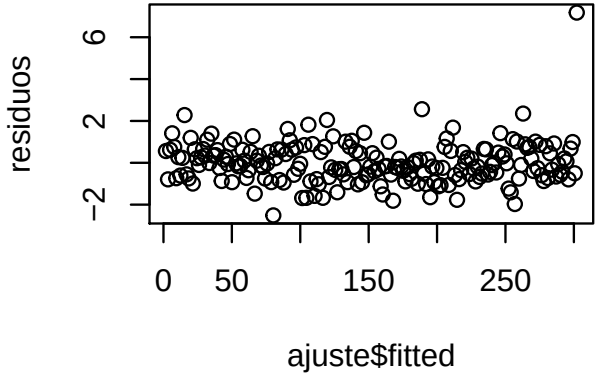
Falta de linealidad



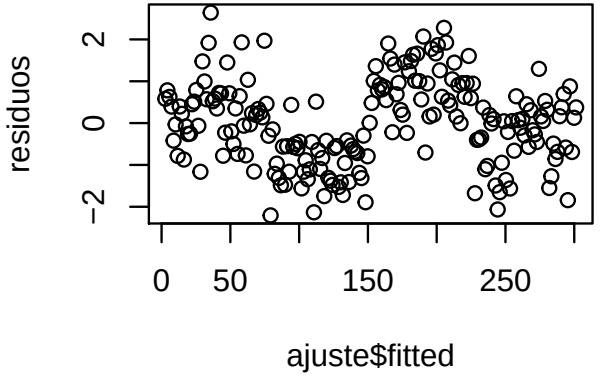
Heteroscedasticidad



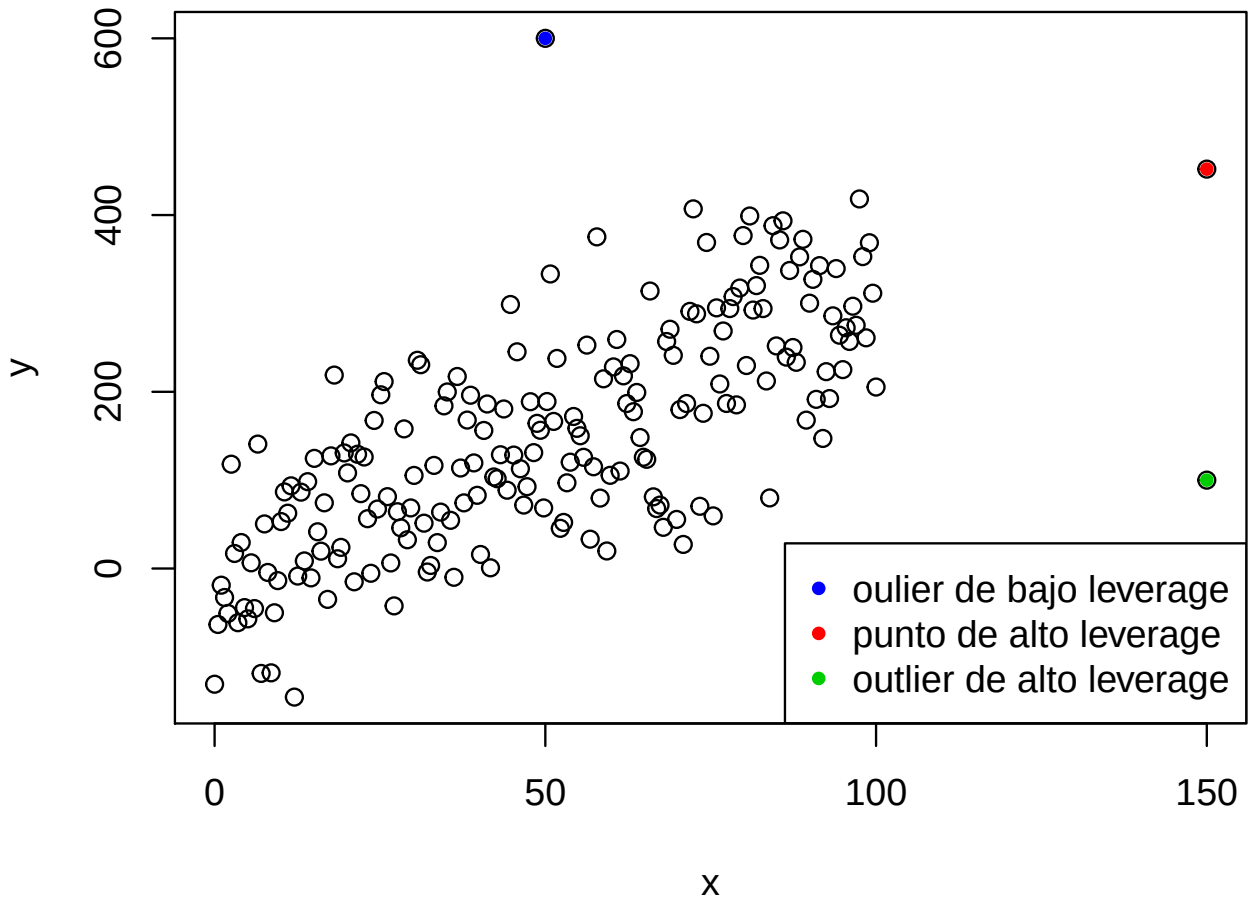
Presencia de outliers



Errores Correlacionados

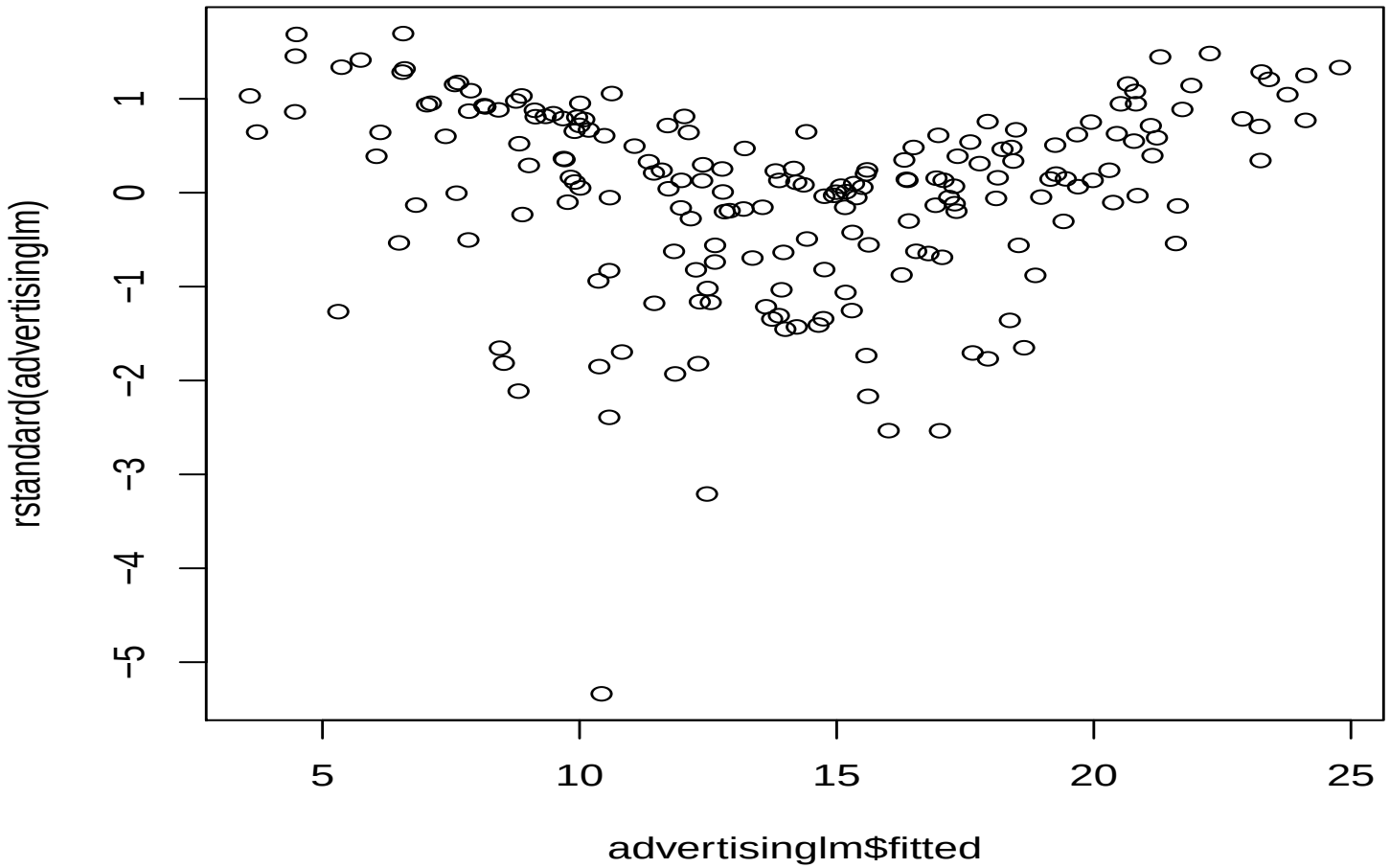


Outliers y punto de alto leverage en regresión simple



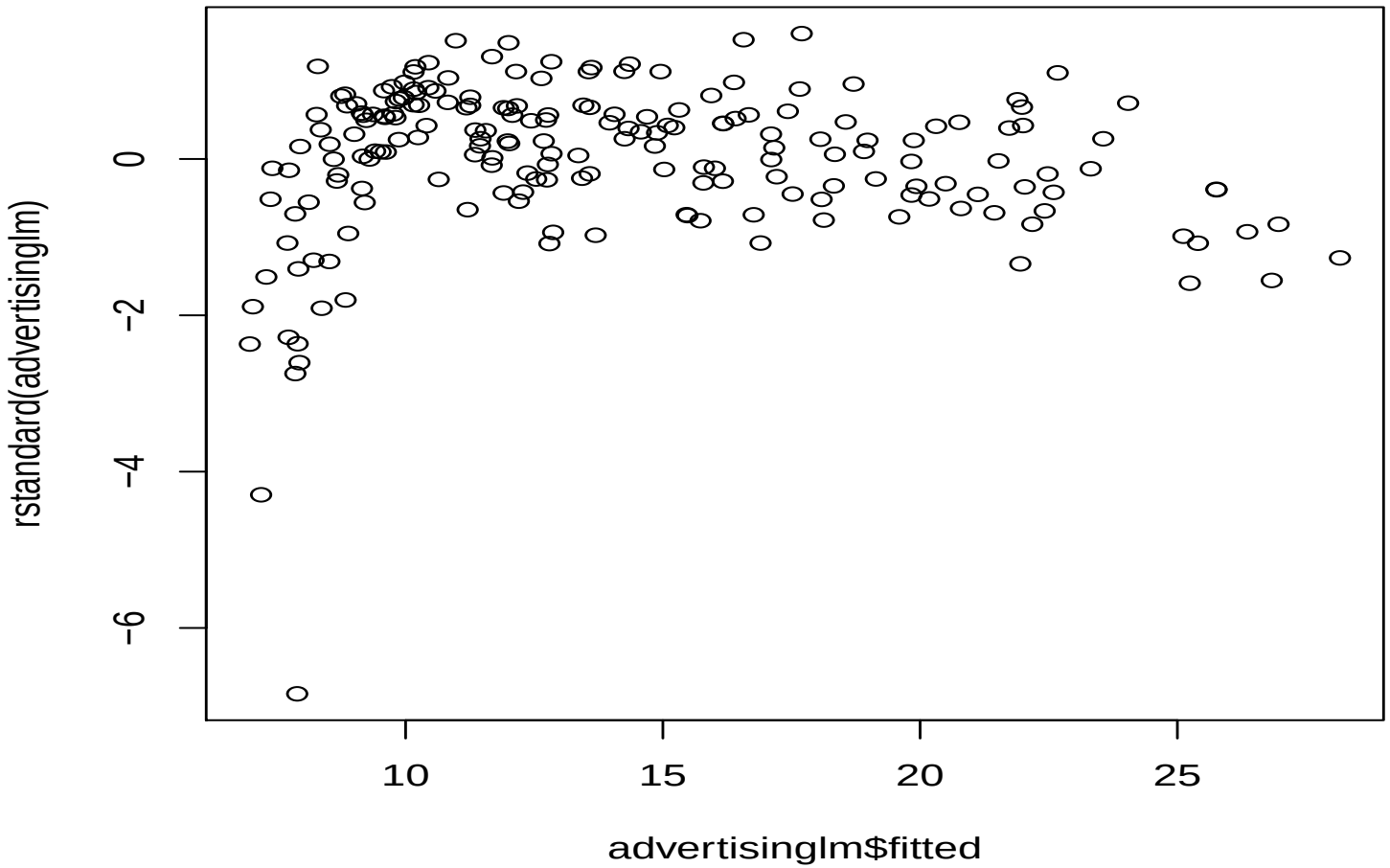
Aplicación a los datos de publicidad

```
> advertisinglm<-lm(sales~TV+radio+newspaper,data=advdata)
> plot(advertisinglm$fitted,rstandard(advertisinglm))
```


rstandard(advertisinglm)
1
0
−1
−2
−3
−4
−5
5
10
15
20
25
advertisinglm$fitted

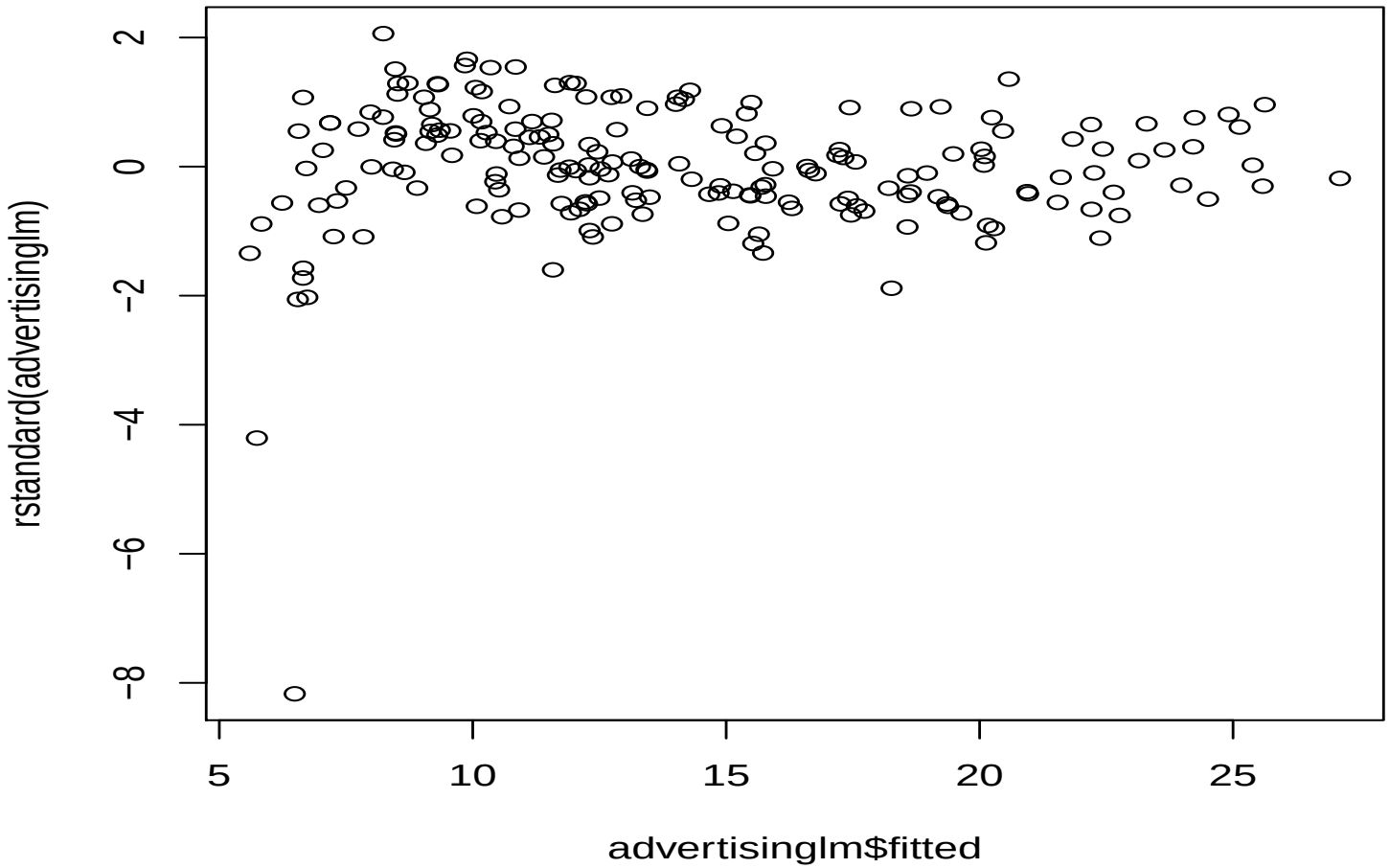
Aplicación a los datos de publicidad

```
> advertisinglm<-lm(sales~TV*radio+newspaper,data=advdata)
> plot(advertisinglm$fitted,rstandard(advertisinglm))
```


rstandard(advertisinglm)
0
−2
−4
−6
10
15
20
25
advertisinglm$fitted

Aplicación a los datos de publicidad

```
TVcuad<-advdata$TV^2
advertisinglm<-lm(sales~TV*radio+TVcuad+newspaper,data=advc
plot(advertisinglm$fitted,rstandard(advertisinglm))
```


rstandard(advertisinglm)
2
0
−2
−4
−6
−8
5
10
15
20
25
advertisinglm$fitted

Dado un valor de las covariables, ¿cómo predecimos la respuesta y cuán precisa es la predicción?

Respuesta predicha cuando la inversión en publicidad en TV es x_1 y la inversión en publicidad en radio es x_2:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Predicciones

Ventas predichas si la inversión en publicidad en TV es de \$120000, la inversión en publicidad en radio es \$10000 y la publicidad en diarios es de \$10000.

```
> nuevox<-data.frame(TV=120,radio=10,newspaper=10)
> predict(advertisinglm,newdata=nuevox)
       1
10.30557
```

Intervalo de predicción

```
> predict(lm(sales~TV+radio),newdata = nuevox,
+         interval = "prediction")
       fit      lwr      upr
1 10.29162 6.960349 13.62289
```

Podemos afirmar con un nivel de confianza de 95% que,

- si en una ciudad, la inversión en publicidad en TV es de \$120000 y la inversión en publicidad en radio es \$10000, las ventas estarán entre 6.960349 y 13.62289 miles de unidades.

Intervalo de confianza para el valor esperado de Y

```
> predict(lm(sales~TV+radio),newdata = nuevox,
+         interval = "confidence")
       fit    lwr      upr
1 10.29162 9.9707 10.61254
```

Podemos afirmar con un nivel de confianza de 95% que,

- si en una gran cantidad de ciudades, la inversión en publicidad en TV es de \$120000 y la inversión en publicidad en radio es \$10000, las ventas promedio estarán entre 9.9707 y 10.61254 miles de unidades.

Muestra de entrenamiento

- **Muestra de entrenamiento** Es la muestra que utilizamos para hacer las estimaciones
- **Error cuadrático medio de entrenamiento**

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

En nuestro caso: Es el error cuadrático medio que se comete al predecir las ventas en los mismos mercados que se usaron para estimar los coeficientes.

$$\downarrow$$

Fácil de calcular.

Error de testeo

- **Error de testeo** Es el error medio que se comete al usar el método de aprendizaje en una muestra con nuevas observaciones, que no fueron usadas en la etapa de entrenamiento o aprendizaje.
  En nuestro caso: Es el error cuadrático medio que se comete al predecir las ventas en nuevos mercados, no incluídos en la muestra de entrenamiento.

  $\downarrow$

  Conviene estimarlo de antemano.
- El error de entrenamiento suele subestimar el error de testeo.

Estimando el error de testeo

Se divide la muestra en 2 partes: muestra de entrenamiento y muestra de testeo.

Se ajusta el modelo usando la muestra de entrenamiento y el modelo ajustado se usa para predecir las respuestas de la muestra de testeo

El error cuadrático medio calculado con las observaciones de la muestra de testeo es un estimador del error de testeo.

Aplicación a datos de publicidad

```
> set.seed(10)
> train<-sample(200,150)
> test<-(-train)
> advlmtrain<-lm(sales~TV+radio+newspaper,data=advdata[train,])
> nuevox<-data.frame(cbind(TV,radio,newspaper))[test,]
> pred1<-predict(advlmtrain,newdata=nuevox)
> mean((pred1-sales[test])^2)
[1] 4.997066
> advlmtrain<-lm(sales~TV+radio,data=advdata[train,])
> nuevox<-data.frame(cbind(TV,radio,newspaper))[test,]
> pred1<-predict(advlmtrain,newdata=nuevox)
> mean((pred1-sales[test])^2)
[1] 4.970777
```

Desventajas de este enfoque

- Este estimador del error de testeo puede ser muy variable, dependiendo de cuáles observaciones están en la muestra de testeo y cuáles en la muestra de entrenamiento.

- Cómo los métodos estadísticos en general funcionan mejor cuando hay muchas observaciones, este enfoque suele sobreestimar el error de testeo.

Validación cruzada *Leave one out*

Una sola observación se usa como muestra de testeo y todas las demás como muestra de entrenamiento. Supongamos que sacamos $(\mathbf{x}_1, y_1)$

$$ECM_1 = (y_1 - \hat{y}_1)^2$$

es un estimador "aproximadamente insesgado" del error de testeo
Este procedimiento se repite n veces, dejando afuera una observación cada vez. Obtenemos

$$ECM_1, ECM_2, \ldots, ECM_n.$$

El estimador $LOOCV$ del error de testeo es el promedio de estos :

$$ECM_k = \frac{1}{n} \sum_{i=1}^{n} ECM_i.$$

Aplicación a los datos de publicidad

```
> advglm<-glm(sales~TV+radio+newspaper)
> cverr1<-cv.glm(advdata,advglm)
> cverr1$delta
[1] 2.946900 2.946486
> advglm<-glm(sales~TV+radio)
> cverr1<-cv.glm(advdata,advglm)
> cverr1$delta
[1] 2.910676 2.910357
```

Error de validación cruzada

Estima el error de testeo.

Cómo calcularlo:

- Se divide la muestra en 10 submuestras de aproximadamente igual cantidad de ciudades.
- Dejando afuera una de las submuestras, se estiman los parámetros usando los datos de las muestras restantes.
- Con los estimadores obtenidos, se predice la variable de respuesta de las observaciones que quedaron afuera. Se calcula el error cuadrático medio que se comete en esas observaciones.
- Se repite esto para cada una de las submuestras. Se obtienen 10 estimadores del error de testeo: $E\hat{C}M_k$, $k = 1, \ldots, 10. ECM_k = \sum_{i=1}^{10} \left(y_i - \hat{y}_i^k\right)^2$ donde $\hat{y}_i^k$ es la predicción de la i-ésimo respuesta calculada sin usar la k-ésima submuestra.
- El error de validación cruzada es el promedio de los $E\hat{C}M_k$.

Aplicación a los datos de publicidad

```
> advglm<-glm(sales~TV+radio+newspaper)
> cverr1<-cv.glm(advdata,advglm,K=10)
> cverr1$delta
[1] 2.968930 2.959015
> advglm<-glm(sales~TV+radio)
> cverr1<-cv.glm(advdata,advglm,K=10)
> cverr1$delta
[1] 2.880532 2.875480
```

- Dividir a la muestra en submuestra de entrenamiento y de testeo da un estimador del error de testeo sesgado.
- LOOCV da estimadores aproximadamente insesgados del error de testeo, pero tienen mayor varianza.
- 5-fold o 10-fold cv es conveniente porque da un buen compromiso entre sesgo y varianza.

Posibles problemas

- Falta de linealidad
- Heteroscedasticidad
- Presencia de outliers
- Errores correlacionados
- Puntos de alto leverage
- Colinealidad

Posibles soluciones

Falta de linealidad: Considerar interacciones entre las variables explicativas y/o transformaciones de las variables explicativas.

Heteroscedasticidad: Hacer mínimos cuadrados pesados o transformar la variable de respuesta.

Presencia de outliers o puntos de alto leverage: Si el outlier fue producto de un error de medición o de registro puede eliminarse la observacion. Ante la duda, usar estimadores robustos.

Colinealidad: Eliminar o combinar variables explicativas o usar métodos de regularización

Ejercicio de aplicación: Datos inmobiliarios

El conjunto de datos "casas.txt" contiene datos de 400 barrios de una cierta región. Se quiere predecir el valor mediano de las propiedades de ese barrio usando las covariables :

1. `hab`: cantidad media de habitaciones por casa,
2. `pecon`: porcentage de casas de bajo nivel económico,
3. `dis`: distancia a centro urbano,
4. `anti`: antiguedad media de las casas,
5. `crim`: índice de crímenes per cápita.

Seleccionar un modelo lineal para predecir el valor mediano de las casas usando las covariables que considere con el fin de obtener el menor error de predicción posible.
Luego de seleccionado el modelo y ajustados los coeficientes se testeará con una nueva muestra.

Regresión ridge

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Regresión ridge

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

El estimador ridge se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.
- El parámetro de calibración λ sirve para controlar el impacto relativo de estos dos términos en los estimadores

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.
- El parámetro de calibración λ sirve para controlar el impacto relativo de estos dos términos en los estimadores
- Elegir un buen valor de λ es crítico. Para eso usamos validación cruzada.

Escala de las covariables

Los estimadores regularizados no son equivariantes por cambios de escala. Esto quiere decir que si se cambian las unidades de medición de las variables explicativas, las predicciones cambiarán.

Como se penaliza usando la norma de beta, los tamaõs de los coeficientes deben ser comparables.

Conviene estandarizar las variables explicativas para que estén todas en la misma escala reemplazando cada x_{ij} por

$$\tilde{x}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$$

Datos de cancer de próstata

Este conjunto de datos se recolectó con el objetivo de indentificar factores de riesgo para cancer de próstata Se quiere predecir el logaritmo of PSA (lpsa) a partir de las siguientes variables

- `lcavol` log del volumen del cancer
- `lweight` log del peso de la prostata
- `age` edad
- `lbph` log of benign prostatic hyperplasia amount lbph
- `svi` seminal vesicle invasion
- `lcp` log de penetración capsular
- `gleason` índice de Gleason
- `pgg45` percent of Gleason scores 4 or 5

Aplicación a los datos de cancer de pr óstata

```
> fitridge<-glmnet(x,y,alpha=0)
> names(fitridge)
 [1] "a0"         "beta"       "df"         "dim"        "lambda"
 [6] "dev.ratio" "nulldev"    "npasses"    "jerr"       "offset"
[11] "call"       "nobs"
```

Aplicación a los datos de cancer de pr óstata

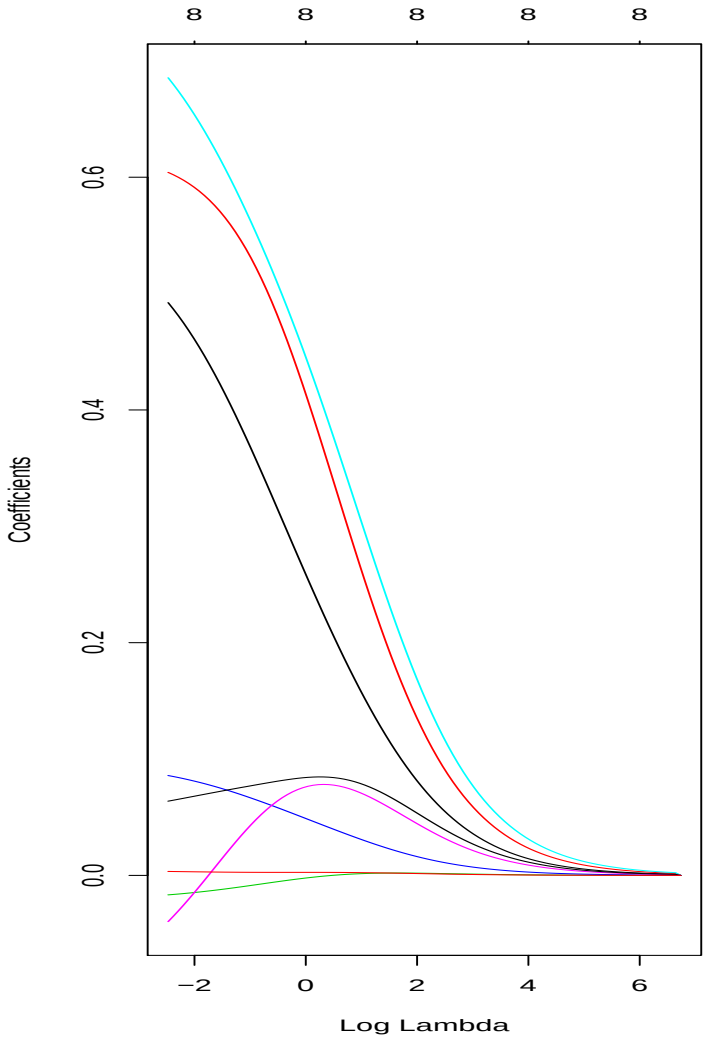
```
> fitridge
Call:  glmnet(x = x, y = y, alpha = 0)

      Df      %Dev    Lambda
 [1,]  8 3.484e-36 843.40000
 [2,]  8 5.118e-03 768.50000
 [3,]  8 5.613e-03 700.20000
 [4,]  8 6.156e-03 638.00000
 [5,]  8 6.750e-03 581.30000
 [6,]  8 7.402e-03 529.70000
 [7,]  8 8.115e-03 482.60000
 [8,]  8 8.897e-03 439.80000
 [9,]  8 9.753e-03 400.70000
[10,]  8 1.069e-02 365.10000
```

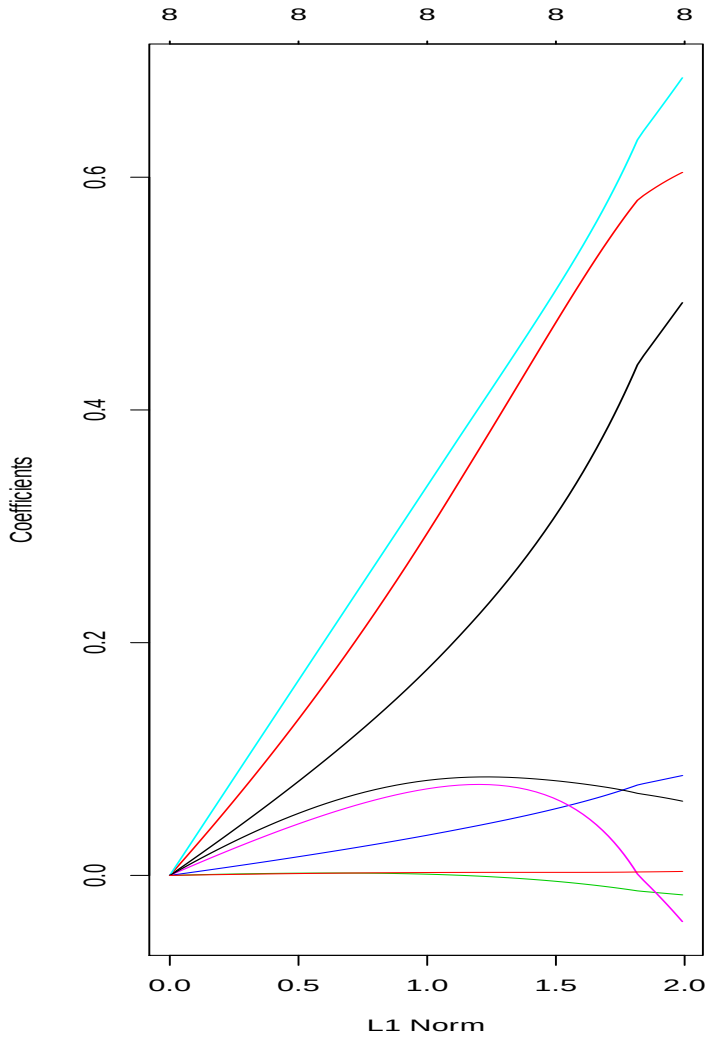
```
[11,]  8 1.171e-02 332.70000
[12,]  8 1.284e-02 303.10000
[13,]  8 1.406e-02 276.20000
[14,]  8 1.541e-02 251.60000
[15,]  8 1.687e-02 229.30000
[16,]  8 1.848e-02 208.90000
[17,]  8 2.023e-02 190.40000
[18,]  8 2.214e-02 173.50000
[19,]  8 2.423e-02 158.00000
[20,]  8 2.650e-02 144.00000
[21,]  8 2.898e-02 131.20000
[22,]  8 3.168e-02 119.60000
[23,]  8 3.463e-02 108.90000
[24,]  8 3.783e-02  99.26000
[25,]  8 4.130e-02  90.44000
[26,]  8 4.508e-02  82.40000
[27,]  8 4.917e-02  75.08000
```

Aplicación a los datos de cancer de próstata

```
> par(mfrow=c(1,2))
> plot(fitridge,xvar="lambda")
> plot(fitridge,xvar="norm")
```

Aplicación a los datos de cancer de próstata





Comentarios de la figura anterior

- En la figura de la izquierda, cada curva corresponde al estimador ridge de uno de los coeficientes, en función de λ. A medida que λ aumenta, los estimadores se acercan a cero de manera suave.
- En la figura de la izquierda, cada curva corresponde al estimador ridge de uno de los coeficientes, en función de la norma de $\hat{\beta}$.

¿Cómo elegimos λ?

Decidir qué valor de λ producirá mejores predicciones para nuevos pacientes.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

Mayor flexibilidad $\Rightarrow$ menor sesgo y mayor varianza.
Menor flexibilidad $\Rightarrow$ mayor sesgo y menor varianza.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

Mayor flexibilidad $\Rightarrow$ menor sesgo y mayor varianza.
Menor flexibilidad $\Rightarrow$ mayor sesgo y menor varianza.

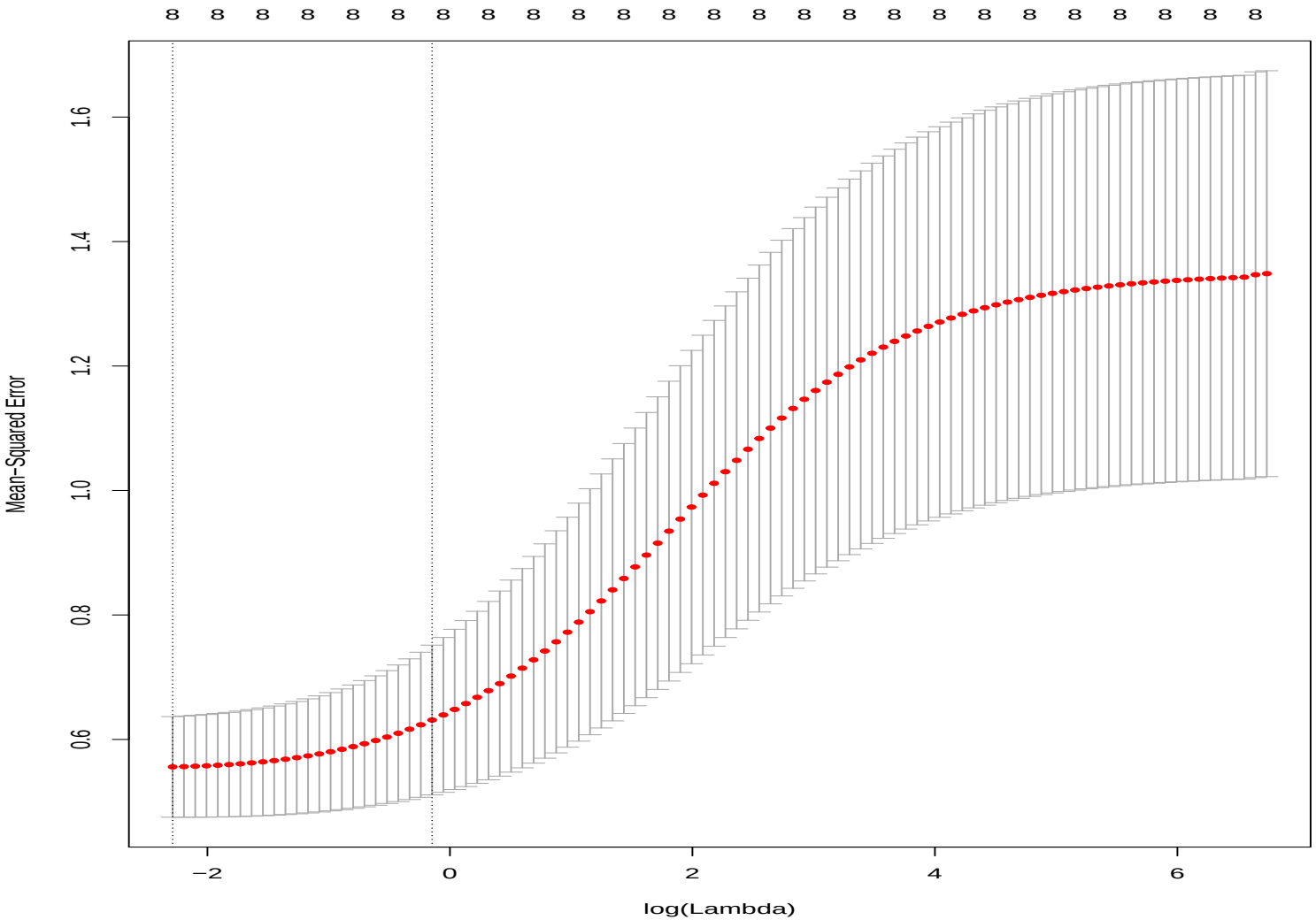
Ridge y lasso ajustan este compromiso automáticamente. Logran un equilibrio entre sesgo y varianza de manera de minimizar el error de pedicción. ¿Como? Eligiendo λ apropiadamente.

Elección del valor de λ por validación cruzada

1. Tomamos una grilla de valores de λ y calculamos el error de validación cruzada para cada λ.
2. Elegimos el valor de λ para el cual el error de validación cruzada es menor.
3. Finalmente se ajusta el modelo utilizando todas las covariables y el valor elegido de λ.

Aplicación a los datos de cancer de próstata

```
> set.seed(12)
> cv0=cv.glmnet(x,y,alpha=0,nfolds=5)
> cv0$lambda.min
[1] 0.09256606
> which(cv0$lambda==cv0$lambda.min)
[1] 98
> cv0$cvm[98]
[1] 0.5558981
> plot(cv0)
```


8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
Mean-Squared Error
0.6
0.8
1.0
1.2
1.4
1.6
−2
0
2
4
6
log(Lambda)

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

y el estimador ridge se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

y el estimador ridge se define como

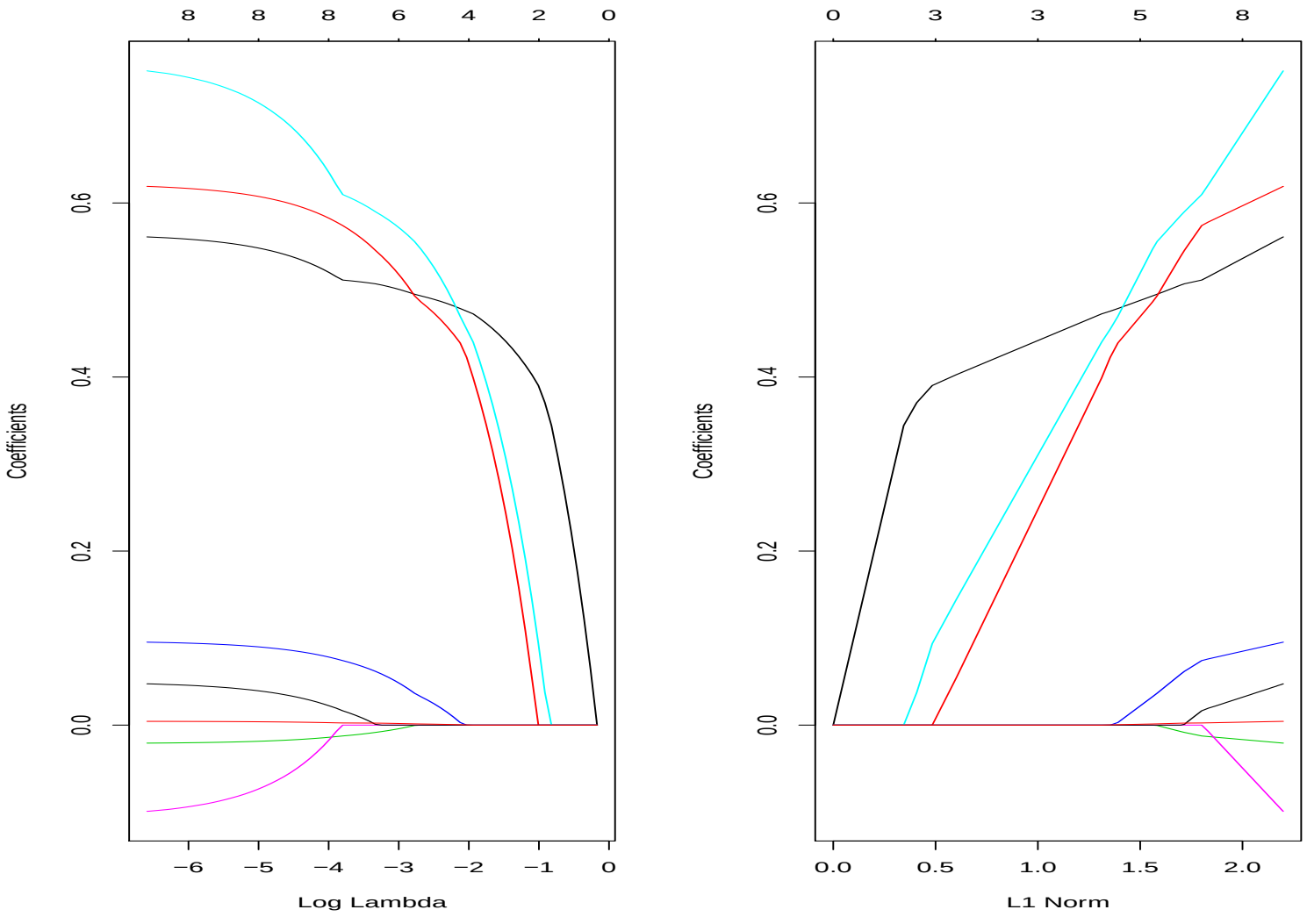
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

el estimador lasso se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

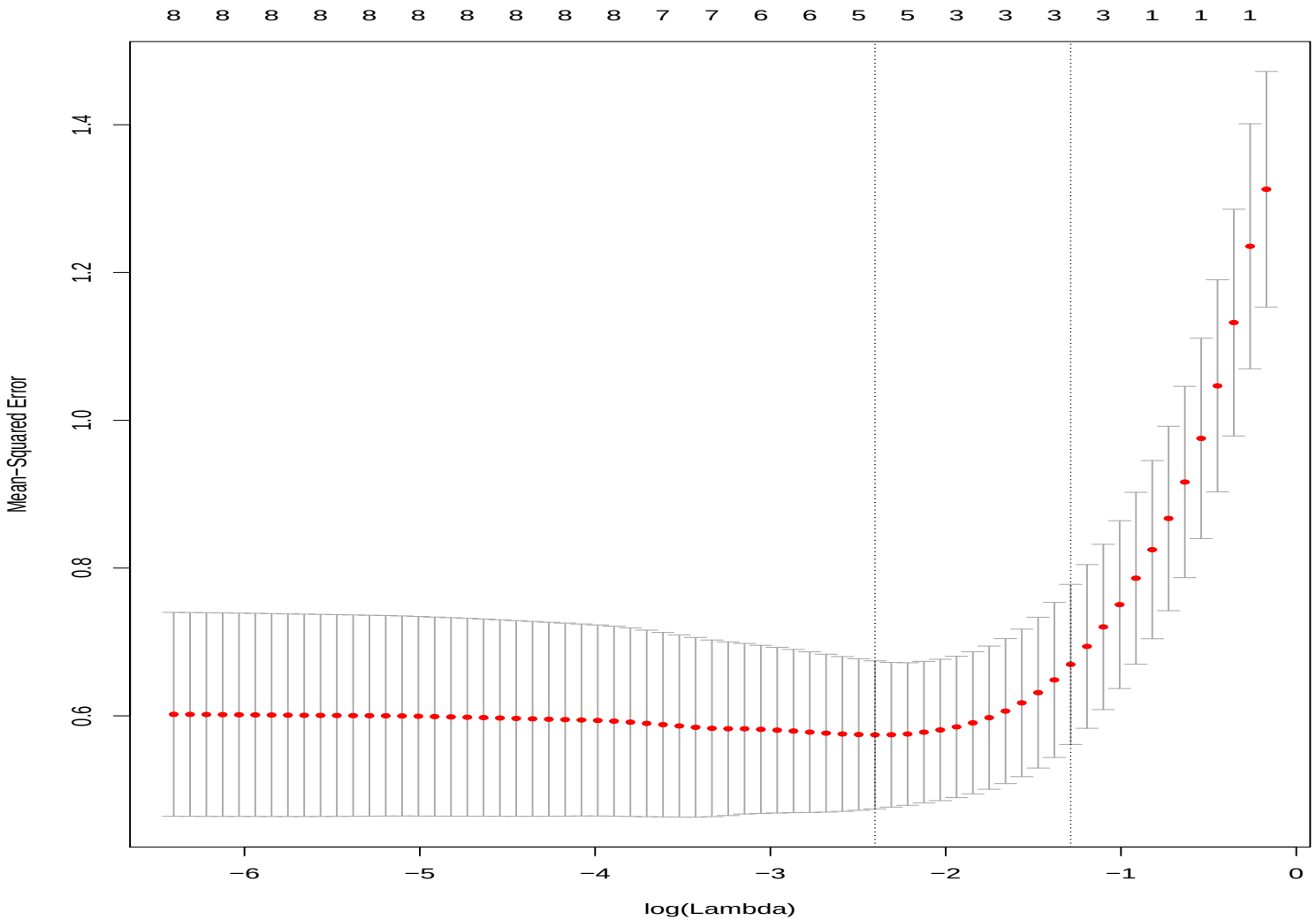
Aplicación a los datos de cancer de próstata

```
fitlasso<-glmnet(x,y,alpha=1)
par(mfrow=c(1,2))
plot(fitlasso,xvar="lambda")
plot(fitlasso,xvar="norm")
```


8 8 8 6 4 2 0
Coefficients
0.0
0.2
0.4
0.6
−6 −5 −4 −3 −2 −1 0
Log Lambda
0 3 3 5 8
Coefficients
0.0
0.2
0.4
0.6
0.0 0.5 1.0 1.5 2.0
L1 Norm

Elección de λ por validación cruzada

```
> cv1=cv.glmnet(x,y,alpha=1,nfolds=5)
> which(cv1$lambda==cv1$lambda.min)
[1] 25
>   cv1$cvm[25]
[1] 0.5743025
> plot(cv1)
```


8 8 8 8 8 8 8 8 8 8 7 7 6 6 5 5 3 3 3 3 1 1 1
Mean-Squared Error
0.6
0.8
1.0
1.2
1.4
−6
−5
−4
−3
−2
−1
0
log(Lambda)

Coeficientes estimados

```
9 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept) -0.0062505892
lcavol       0.4869613127
lweight      0.4659677422
age          .
lbph         0.0198718550
svi          0.5137160983
lcp          .
gleason      .
pgg45        0.0009455418
```

Propiedad de selección de variables del lasso.

El lasso selecciona 5 covariables.
Decimos que el lasso da modelos ralos, es decir donde sólo algunos coeficientes son no nulos.

Propiedad de selección de variables del lasso.

El lasso selecciona 5 covariables.
Decimos que el lasso da modelos ralos, es decir donde sólo algunos coeficientes son no nulos.

Muchas veces se necesita identificar un subconjunto de covariables lo más pequeño posible con el cual poder predecir correctamente (por ejemplo, para usar en diagnósticos futuros).

Geometría de ridge y lasso

Puede probarse que

$$\hat{\boldsymbol{\beta}}^R = \arg\max_{\boldsymbol{\beta}} L(\beta_0, \beta_1, \ldots, \beta_p)$$

sujeto a

$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

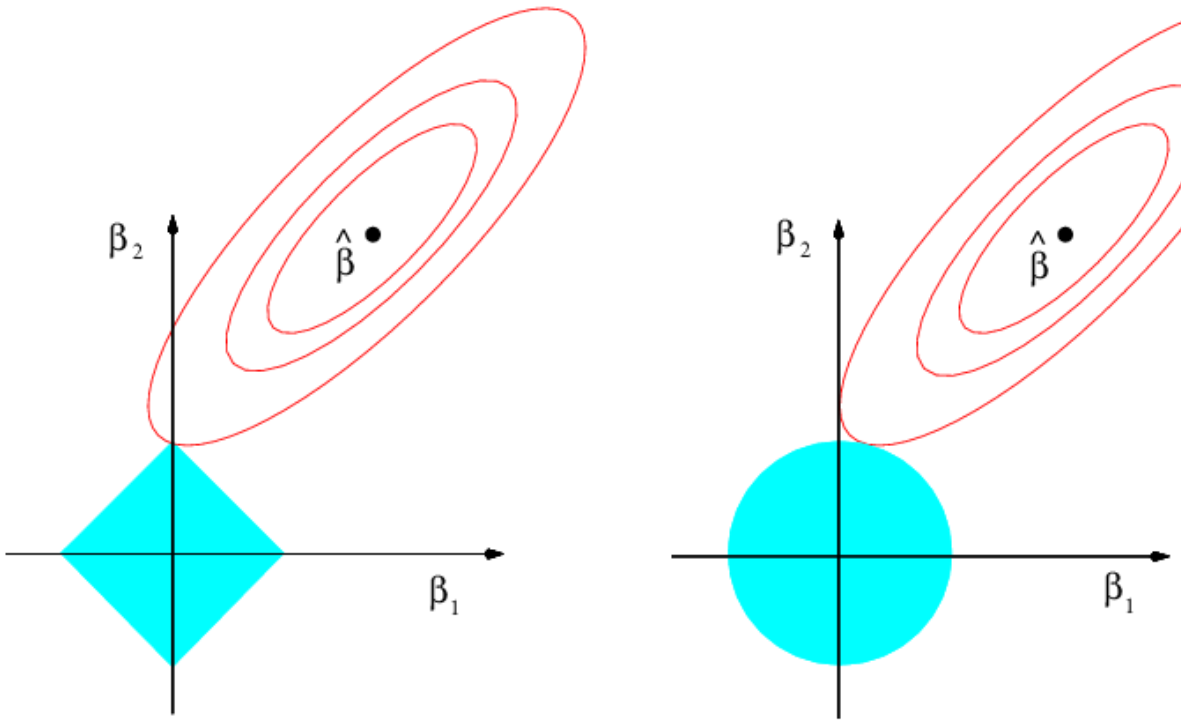
y

$$\hat{\boldsymbol{\beta}}^L = \arg\max_{\boldsymbol{\beta}} L(\beta_0, \beta_1, \ldots, \beta_p)$$

sujeto a

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Geometría de ridge y lasso



Comparación entre ridge y lasso

- Ninguno de los dos métodos supera al otro en todo contexto.
- En general si solo una pequeña proporción de las covariables se relaciona con la respuesta funcionará mejor lasso.
- Si una gran proporción de las covariables se relaciona con la respuesta funcionará mejor ridge.