Modelos Estadísticos para la Regresión y la Clasificación

Clase 6: Regresión lineal (2da parte). Regresión Ridge y regresión Lasso

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

9 de septiembre de 2024

Plan

Resumen general

La expresión general del modelo lineal es:

$$Y = \underbrace{\mathbf{x}'\beta}_{f(\mathbf{x})} + \epsilon$$

y la estimación:

$$\widehat{\mathbf{y}} = \mathbf{x}'\widehat{\beta}$$

donde $\widehat{\beta}$ es la estimación del vector β obtenida por el método de los mínimos cuadrados.

Si suponemos las hipótesis de Gauss-Markov, el modelo lineal $Y = \mathbf{x}'\beta + \epsilon$ cumple que

$$\mathbb{E}(Y) = \mathbf{x}'\beta$$

Si además de suponer las condiciones de Gauss-Markov sobre los errores, se tiene que $\epsilon_i \sim N(0, \sigma^2)$ y que $\epsilon_1, \ldots, \epsilon_n$ son independientes, entonces decimos que el modelo es normal y se tiene que:

$$Y \sim N(\mathbf{x}'\beta, \sigma^2)$$

Esto último lo podemos verificar con un test de Shapiro Wilks.

Del modelo $Y = X\beta + \epsilon$, deducimos que matricialmente:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Entonces $(X'X)\beta = X'Y \Leftrightarrow \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} n\overline{y} \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$

Por otro lado

$$(X'X)^{-1} = \frac{1}{nS_x^2} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}$$

La recta de regresión en este caso es

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

siendo los estimadores:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$$

La recta de regresión se expresa también como

$$y - \overline{y} = \widehat{\beta}_1(x - \overline{x})$$

y por lo tanto para todo $i = 1, \ldots, n$ se tiene que $\widehat{y_i} - \overline{y} = \widehat{\beta}_1(x_i - \overline{x})$ y por lo tanto $\sum(\widehat{y_i} - \overline{y})^2 = \widehat{\beta}_1^2 S_x^2$.

Distribución de los estimadores

Suponemos que la matriz de diseño X es de rango máximo $d+1$ y que los residuos cumplen con las condiciones de Gauss-Markov. Entonces:

1. $\mathbb{E}(\widehat{\beta}) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}X'X\beta = \beta$
   $\Rightarrow \widehat{\beta}$ es un estimador insesgado de β.

2. $Var(Y) = Var(Y - X\beta) = Var(\epsilon) = \sigma^2 I$
   $\Rightarrow Var(\widehat{\beta}) = Var\big((X'X)^{-1}X'Y\big) = (X'X)^{-1}X'Var(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$.

Entonces:

$$\mathbb{E}(\widehat{\beta}) = \beta \qquad Var(\widehat{\beta}) = \sigma^2(X'X)^{-1}$$

Se puede probar que $\widehat{\beta}$ es el estimador de mínima varianza (el más eficiente) entre todos los estimadores lineales insesgados de β.

La varianza residual σ^2 se estima por

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SCR}{n-2}$$

Plan

Regresión lineal simple. Descomposición variación

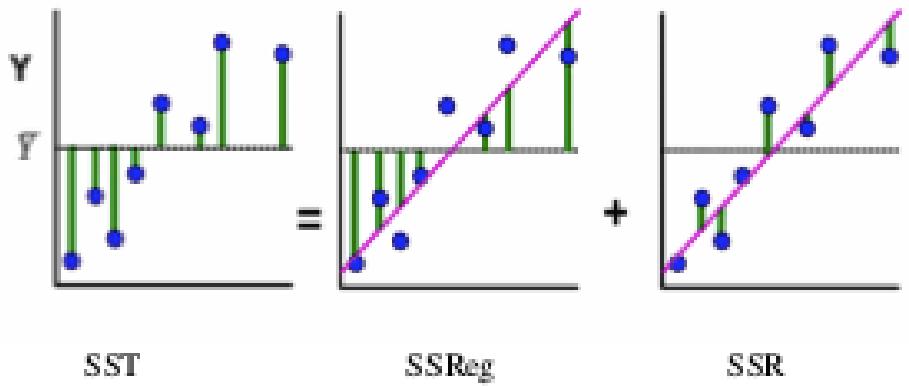
Con nuestras notaciones, si $\widehat{y}_i$ es la predicción de x_i por el modelo, se verifica lo que llamamos la *descomposición de la variación:*

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{Variación total VT o SST}} = \underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{Variación no explicada VNE o SCR}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{\text{Variación explicada VE o SSReg}}$$

En efecto:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\widehat{y}_i+\widehat{y}_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2 + 2\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y})}_{0}$$

porque $\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y}) = \sum_{i=1}^{n}(y_i-\widehat{y}_i)\widehat{y}_i - \overline{y}\sum_{i=1}^{n}(y_i-\widehat{y}_i) = \sum_{i=1}^{n} e_i\widehat{y}_i - \overline{y}\sum_{i=1}^{n} e_i = 0$ (justificar)



Entonces:

1. La variación total es $VT = \sum_{i=1}^{n}(y_i - \overline{y})^2 = S_y^2$
2. La variación no explicada por la regresión es $VNE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = SCR$
3. La variación explicada por la regresión es $VE = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \widehat{\beta}_1^2 S_x^2 = SSR$

Regresión lineal simple. Descomposición variación

De la descomposición de la variación tenemos que:

$$S_y = SCR + \underbrace{\widehat{\beta}_1^2 S_x}_{SSR}.$$

Más aún:

$$SCR = (1 - r^2) S_y$$

donde r es el coeficiente de correlación muestral entre x e y.

$$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = S_y^2 - \widehat{\beta}_1^2 S_x^2 = S_y^2 - \frac{S_{xy}^2}{S_x^4} S_x^2 = S_y^2 - \frac{S_{xy}^2}{S_x^2} = S_y^2 - r^2 S_y^2 = (1 - r^2) S_y^2$$

De la cuenta anterior tenemos

$$S_y^2 = SCR + \widehat{\beta}_1^2 S_x^2 = SCR + r^2 S_y^2$$

Por otro lado una estimación de σ^2 es

$$\widehat{\sigma}^2 = \frac{(1 - r^2) S_y^2}{n - 2}$$

Regresión lineal simple. Coeficiente de determinación R^2

La proporción de variabilidad explicada por el modelo es el *coeficiente de determinación* :

$$R^2 = \frac{VE}{VT} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{VT - VNE}{VT} = 1 - \frac{SCR}{S_y^2}$$

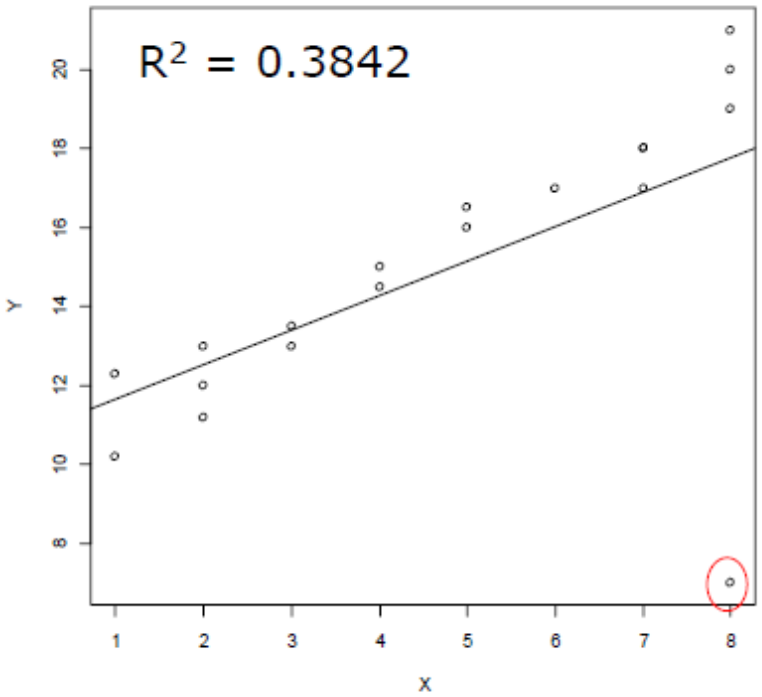
El coeficiente de determinación R^2 es una medida de la bondad del ajuste, **suponiendo que el modelo es lineal**. *En el caso de la regresión lineal simple* coincide con r^2.

- Observar que $0 \leq R^2 \leq 1$: si el valor de R^2 es cercano a 1 entonces gran parte de la variabilidad es explicada por el modelo, mientras que si está cerca de 0, una parte importante de la variabilidad no está explicada por el modelo (es probable que el modelo no sea adecuado).
- Cuidado que el R^2 no es una medida de adecuación del modelo. Es una medida de cuan significativo es el modelo una vez que establecimos que responde a un modelo lineal. Para ver si el modelo se ajusta a un modelo lineal, se usa el test Lack of Fit (LOF) cuando tenemos réplicas.
- Puede ocurrir también que la presencia de algún outlier implique que R^2 es bajo y hacernos pensar que el modelo no es bueno cuando en realidad sí lo es.
- Para corregir el peligro de sobreajuste se define el coeficiente de determinación ajustado como

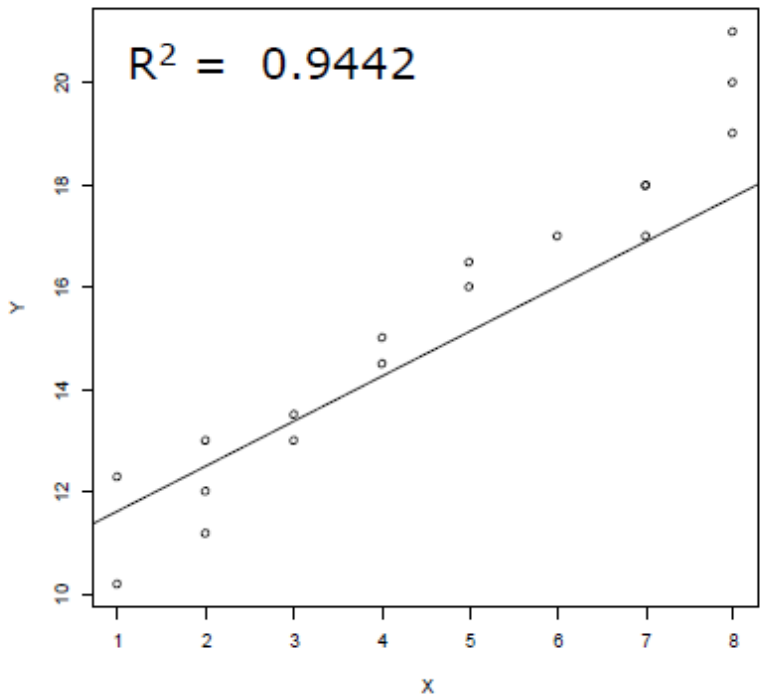
$$\overline{R}^2 = 1 - \frac{SCR/(n-2)}{S_y^2/(n-1)}$$

Si R^2 y $\overline{R}^2$ son muy distintos es que el modelo fue sobreajustado e inducirnos a mirar de más cerca las variables y/o cambiar la cantidad de términos.

Regresión lineal simple. Coeficiente de determinación R^2



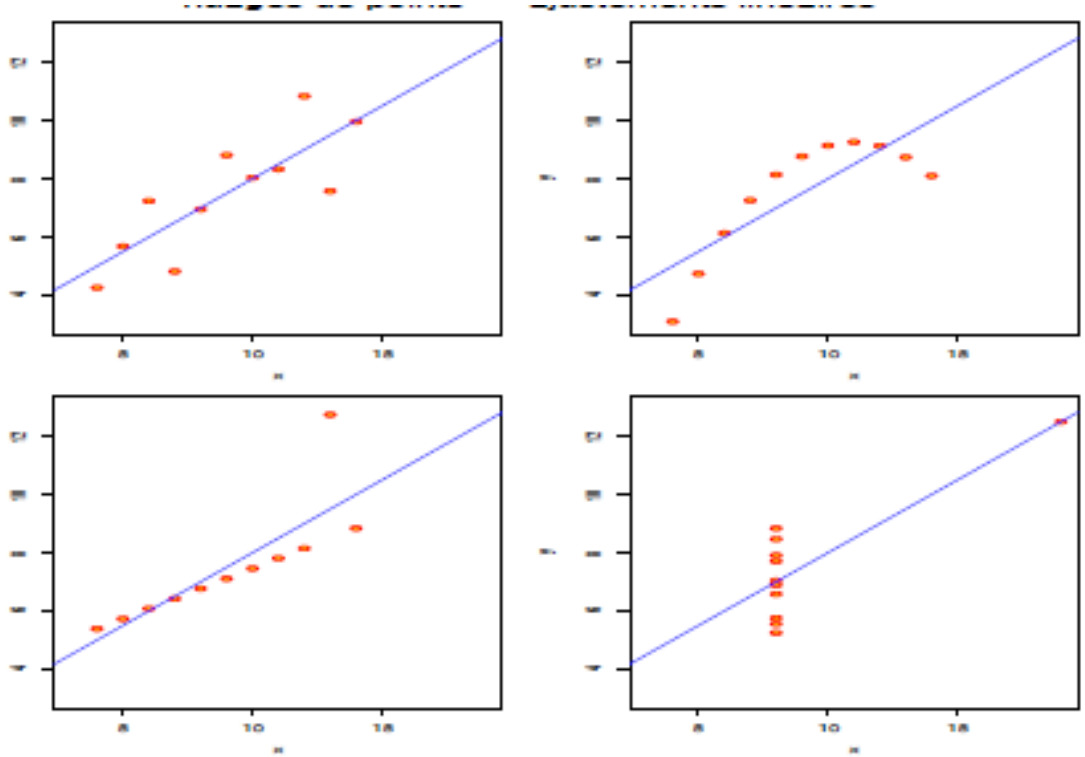
Comportamiento del R^2 con y sin un dato «outlier» en la variable Y.



Regresión lineal simple. Coeficiente de determinación R^2

x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

$$\bar{x} = 9\,;\ \bar{y} = 7.50,$$
$$S_x^2 = 10\,;\ S_y^2 = 3.75$$
$$r = 0.816.$$



Ojo que la recta dibujada es la "verdadera".

Plan

Bajo la hipótesis de normalidad de los residuos, los estimadores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ de β_0 y β_1 tienen distribución

$$\widehat{\beta_1} \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_X^2}\right) \quad \text{y} \quad \widehat{\beta_0} \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{x}^2}{S_X^2}\right)$$

donde $S_X^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$ y estimamos la varianza por $\widehat{\sigma^2} = \frac{SCR}{n-2}$.

Se prueba que:

- $\frac{n-2}{\sigma^2} SCR \sim \chi^2_{n-2}$
- $\frac{\widehat{\beta_0} - \beta_0}{s.e(\widehat{\beta_0})} \sim t_{n-2}$ donde $s.e(\widehat{\beta_0})^2 = var(\widehat{\beta_0}) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_x^2}\right) = \frac{SCR}{n-2}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_x^2}\right)$
- $\frac{\widehat{\beta_1} - \beta_1}{s.e(\widehat{\beta_1})} \sim t_{n-2}$ donde $s.e(\widehat{\beta_1})^2 = var(\widehat{\beta_1}) = \widehat{\sigma}^2 \frac{1}{S_x^2} = \frac{SCR}{n-2}\frac{1}{S_x^2}$

donde $s.e(\widehat{\beta_0})$ y $s.e(\widehat{\beta_1})$ son los estimadores de los desvios estándares de $\widehat{\beta_0}$ y $\widehat{\beta_1}$.

Esto permite construir intervalos de confianza y de testear la nulidad de los parámetros.

Tabla de significancia modelo

En la regresión lineal simple, se quiere testear si hay relación de linealidad entre Y y X. El test es:

$$\begin{cases} H_0 : \text{No hay relación lineal entre } X \text{ e } Y \ (\beta_1 = 0) \\ H_1 : \text{Hay relación lineal} \end{cases}$$

Source	grados libertad	Sum. Squares	Mean Square	F
Modelo	1	$SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	MSR=SSR/1	MSR/MSE
Error	$n-2$	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	MSE= SCR/$(n-2)$	
Total	$n-1$	$SST = S_y = \sum_{i=1}^{n} (y_i - \overline{y})^2$		

El estadístico $F = MSR/MSE$ con el que se testea la hipótesis nula $\beta_1 = 0$ contra la hipótesis $\beta_1 \neq 0$ tiene distribución F con 1 y $n-2$ grados de libertad.

Un valor de MSE pequeño indica que el model ajusta bien ($\widehat{y}_i \approx y_i$), en cambio un valor grande de MSE indica que el modelo no sería razonable.

Se rechaza H_0 si $F > F_\alpha(1, n-2)$.

Supongamos el modelo $Y = \beta_0 + \beta_1 X + \epsilon$

Prueba de hipótesis sobre la pendiente
Con hipótesis de normalidad sobre los residuos se testea:

$$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$$

cuyo estadśitico es $T_1 = \frac{\widehat{\beta}_1 - b_1}{s.e(\widehat{\beta}_1)}$.

Región critica: $\left| \frac{\widehat{\beta}_1 - b_1}{s.e(\widehat{\beta}_1)} \right| > t_{n-2}(\alpha/2)$.

Observación: En el caso $b_1 = 0$, con un $p-$valor pequeño podemos inferir que existe una relación entre Y y X. O sea, un resultado significativo que rechace H_0 puede implicar que el modelo lineal sea adecuado, pero podría ser que no lo sea igual (no confundir significación de la regresión con causalidad). Por otro lado, es equivalente al test F, pues

$$T_1 = \frac{\widehat{\beta}_1}{s.e(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{\sqrt{\frac{SCR}{(n-2)S_x^2}}} = \frac{\widehat{\beta}_1 \sqrt{S_x^2}}{\sqrt{\frac{SCR}{(n-2)}}} = \sqrt{\frac{SSR}{MSE}} = \sqrt{F}$$

Intervalo de confianza al $100(1-\alpha)\,\%$ **para** β_1**:**

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

Prueba de hipótesis sobre el intercepto
Con hipótesis de normalidad:

$$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$$

Región critica: $\left| \frac{\widehat{\beta}_0 - b_0}{s.e(\widehat{\beta}_0)} \right| > t_{n-2}(\alpha/2)$.

Intervalo de confianza al $100(1-\alpha)\,\%$ **para** β_0**:**

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

Intervalo de confianza al $100(1-\alpha)\,\%$ **para** σ^2**:**
Se prueba que un estimador para σ^2 es $\widehat{\sigma}^2 = \frac{SCR}{n-d-1}$. Como $SCR/\sigma^2 \sim \chi^2_{n-2}$, se tiene que:

$$\left[\frac{SCR}{\chi^2_{n-2}(\alpha/2)}, \frac{SCR}{\chi^2_{n-2}(1-\alpha/2)}\right]$$

Regresión lineal múltiple

En este test de hipótesis, bajo hipotesis normalidad, nos preguntamos si los coeficientes de la regresión lineal son nulos o no. Es el análogo al test t de la regresión lineal simple.

En regresión lineal múltiple:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0 \\ H_1 : \text{al menos un } \beta_j \text{ es no nulo} \end{cases}$$

Source	grados libertad	Sum. Squares	Mean Square	F
Modelo	d	$SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	MSR=SSR/p	MSR/MSE
Error	$n-d-1$	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	MSE= SCR/$(n-d-1)$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$		

Plan

Intervalo de confianza para la recta de regresión

Intervalo de confianza para respuesta media:
Se trata de un intervalo de confianza para $\mathbb{E}(Y|X = x_0)$, la respuesta media al valor x_0.

Dado un valor determinado x_0 de la variable independiente, como el error tiene una distribución $\mathcal{N}(0, \sigma^2)$, la variable $Y = \beta_0 + \beta_1 x_0 + \epsilon$ tiene distribución $\mathcal{N}(\mu, \sigma^2)$ donde la media μ de Y es $\mu = \beta_0 + \beta_1 x_0$

Para un x_0 dado, consideramos el pronóstico $\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$. El valor $\widehat{y_0}$ es un estimador de $\mu = \mathbb{E}(Y|x_0))$.
Un intervalo de confianza al nivel $1 - \alpha$ para la respuesta media $\mu = \beta_0 + \beta_1 x_0 = \mathbb{E}(Y|x_0)$ es

$$\left[\widehat{y_0} - t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x^2} \right)}, \widehat{y_0} + t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x^2} \right)} \right]$$

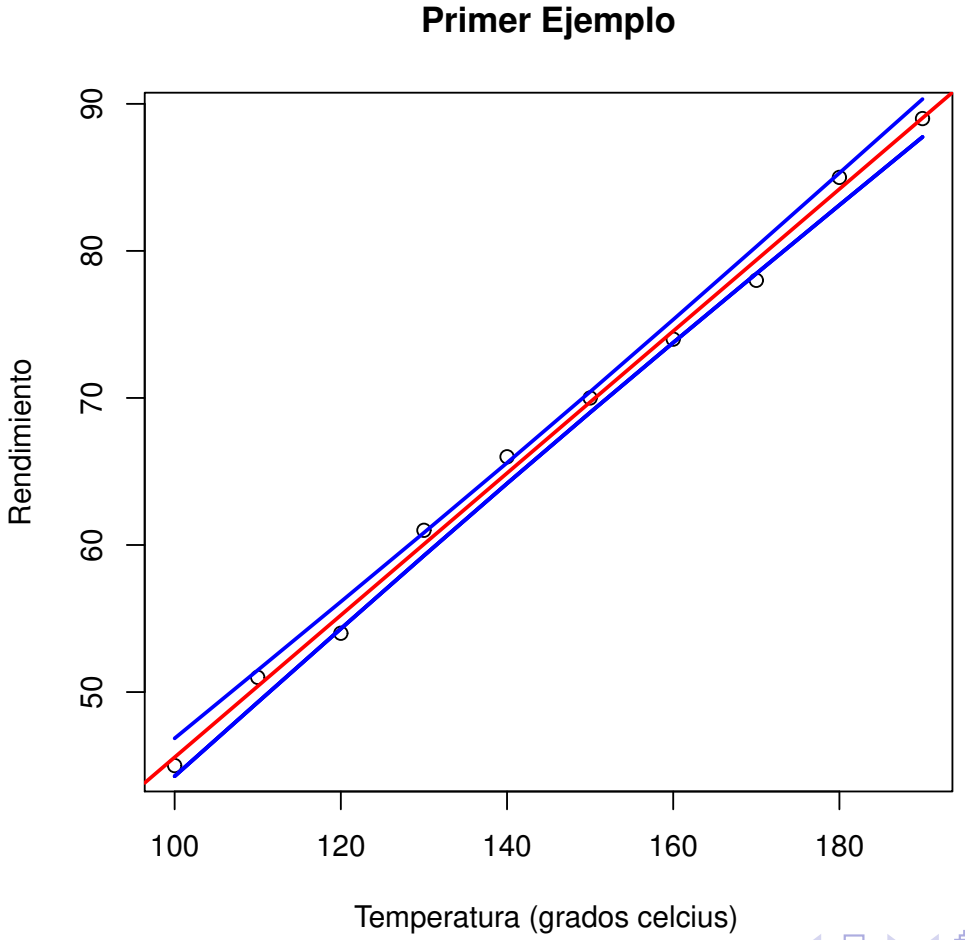
El ancho del intervalo depende de x_0, es mínimo cuando $x_0 = \overline{x}$ y crece cuando $|x_0 - \overline{x}|$ crece.

Intervalo de confianza para la recta de regresión

:

Intervalo de confianza para respuesta media $IC_{1-\alpha}\big(\mathbb{E}(Y|X=x_0)\big)$:

x_0	100	110	120	130	140	150	160	170	180	190
y	45	51	54	61	66	70	74	78	85	89
$\widehat{y}_0$	45.56	50.39	55.22	60.05	64.88	69.72	74.55	79.38	84.21	89.04
límites	$\pm$1.30	$\pm$1.10	$\pm$0.93	$\pm$0.79	$\pm$0.71	$\pm$0.71	$\pm$0.79	$\pm$0.93	$\pm$1.10	$\pm$1.30



Intervalo de confianza para la predicción

El intervalo definido anterioremente es adecuado para el valor esperado de la respuesta, pero ahora queremos un intervalo de predicción para una respuesta individual concreta.

Intervalo de confianza para la predicción: $IC_{1-\alpha}(y_0)$
Sea y_0 el verdadero valor (desconocido por lo tanto) de Y cuando la variable independiente es x_0. $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ es estimador puntual de un nuevo valor de la respuesta $Y_0 = Y|x_0$. Si consideramos un intervalo de confianza para esta futura observación Y_0, el intervalo de confianza para la respuesta media en $x = x_0$ no es apropiado ya que es un intervalo sobre la media de Y_0 (un parámetro), y no sobre futuras observaciones de la distribución. La variable $Y_0 - \widehat{y}_0 \sim \mathcal{N}\left(0, \text{Var}(Y_0 - \widehat{y}_0)\right)$ donde

$$\text{Var}(Y_0 - \widehat{y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_x^2} \right)$$

pues Y_0, una futura observación, es independiente de $\widehat{y}_0$.

Un intervalo de confianza al nivel $1 - \alpha$ para y_0 es

$$\left[\widehat{y}_0 - t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x^2} \right)}, \widehat{y}_0 + t_{\alpha/2, n-2} \sqrt{\underbrace{\frac{SCR}{n-2}}_{MSE} \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{S_x^2} \right)} \right]$$

Consideraciones finales

- Los radios de los dos intervalos crecen cuando x_0 se aleja de $\overline{x}$.
- Los intervalos de predicción para una nueva observación son más amplios que los intervalos de confianza para los parámetros desconocidos. El tamaño del intervalo de confianza para un parámetro depende de la incertidumbre de la estimación que hacemos a partir de una muestra. Mientras que el tamaño del intervalo de predicción para una nueva observación tiene dos fuentes de incertidumbre: una debida a la estimación de los parámetros desconocidos y la otra es propia de la aleatoriedad que suponemos (es una variable aleatoria!).
- No conviene usar el intervalo de confianza de la recta de regresión para hacer previsiones porque es un intervalo de confianza para la verdadera respuesta media en el punto x_0, o sea un paramétro de la población, y no una nueva observación (un nuevo valor para la variable aleatoria Y).

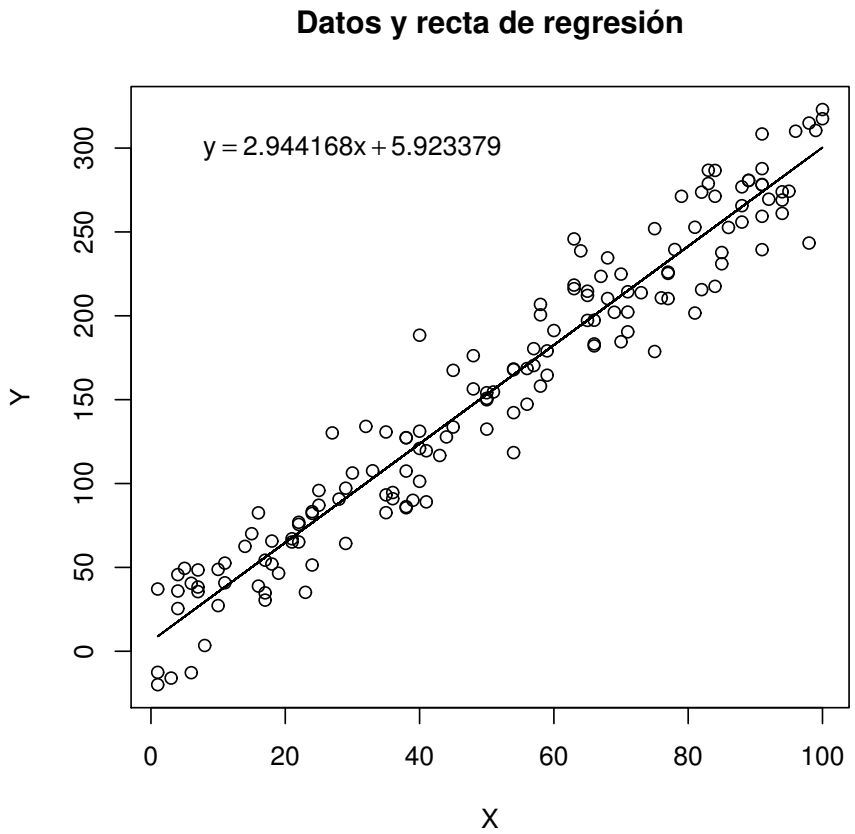
Plan

Ejemplo simulado

Simulemos 150 datos que provienen del modelo

$$Y = 2 + 3X + \epsilon \quad \epsilon \sim N(0, 50)$$

```
>x=1:100
>X=sample(x,150,replace=T)
>Y=2+3*X+rnorm(150,0,50)
>modelo=lm(Y~X)
> modelo

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)            X
      5.923        2.944
```



```
> anova(modelo)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
X           1 1097949 1097949  2269.6 < 2.2e-16 ***
Residuals 148   71598     484
```

```
> summary(modelo)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-51.064 -16.705   0.299  14.702  64.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9234     3.6476   1.624    0.107
X             2.9442     0.0618  47.640   <2e-16 ***
---

Residual standard error: 21.99 on 148 degrees of freedom
Multiple R-squared:  0.9388,    Adjusted R-squared:  0.9384
F-statistic:  2270 on 1 and 148 DF,  p-value: < 2.2e-16
```

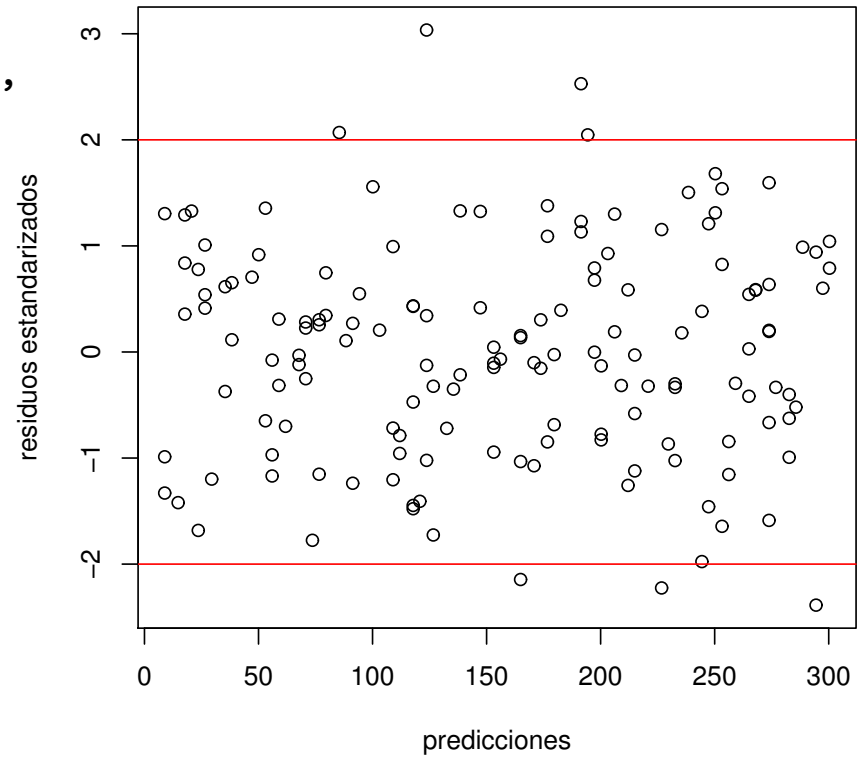
Para ver los residuos y verificar supuesto de normalidad y de iid:

```
> modelo$res
```

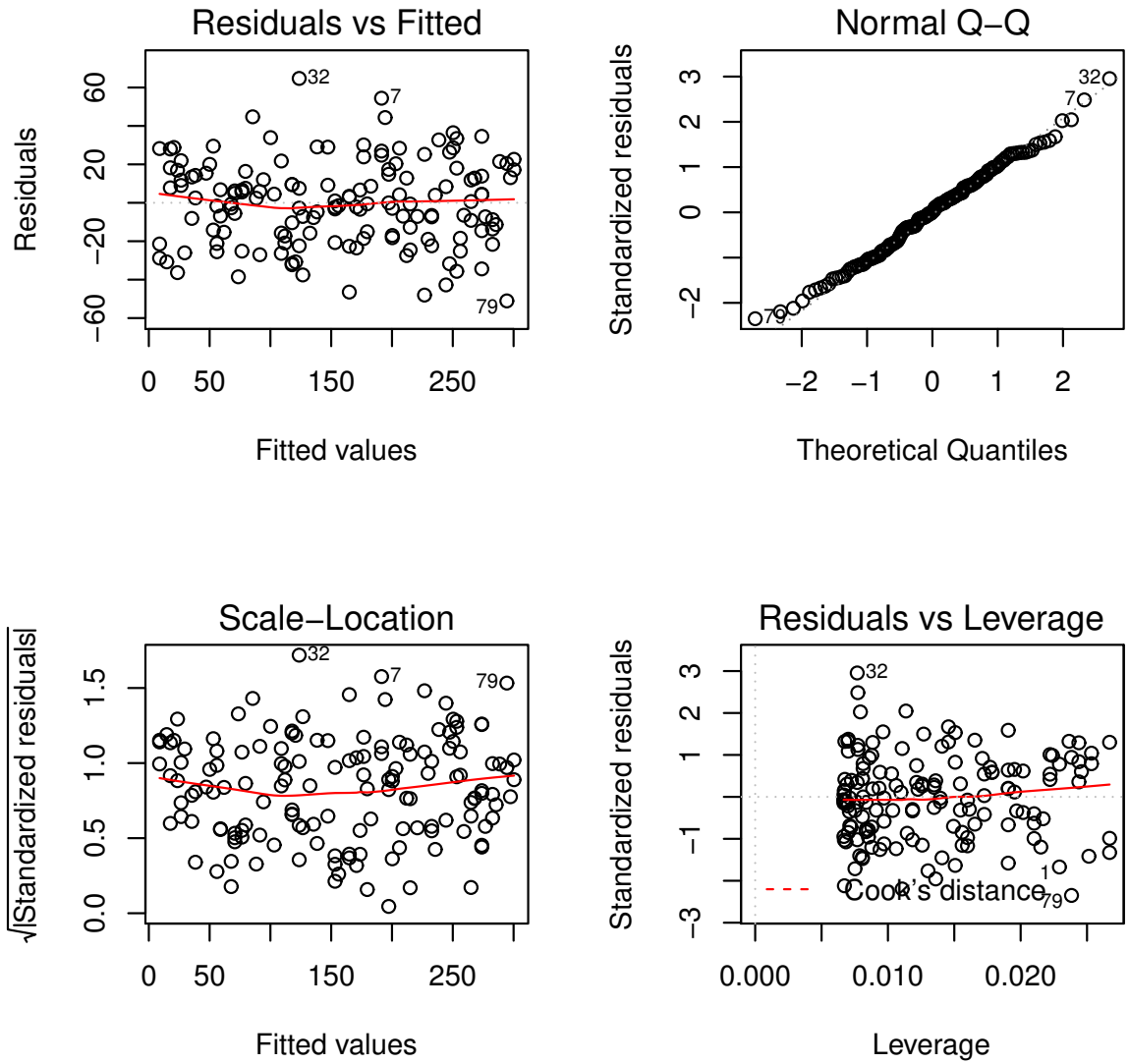
```
> rstudent(modelo)
```

Si un punto tiene residuo studentizado ($e_i/s.e(e_i)$) mayor que 2 en valor absoluto entonces el punto es sospechoso.

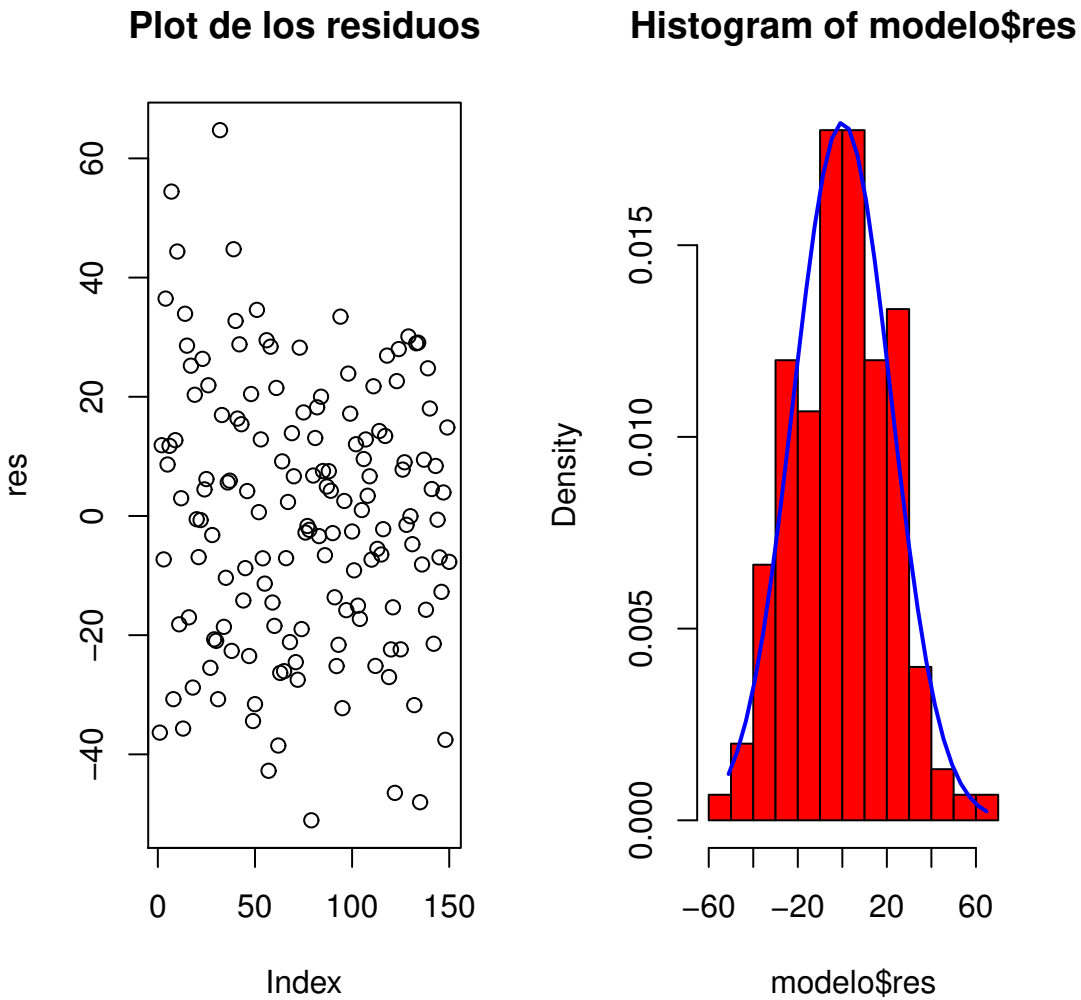
```
> plot(modelo$fitted,rstudent(modelo),
xlab="predicciones",
ylab="residuos estandarizados")

> abline(h=2,col="red")
> abline(h=-2,col="red")
```



```
> par(mfrow=c(2,2))
> plot(modelo)
```



```
>res=resid(modelo)
>par(mfrow=c(1,2))
>plot(res,main=paste("Plot de los residuos"))
>hist(modelo$res,breaks=10,col="red",proba=T)
> xfit=seq(min(res),max(res),length=31)
> yfit=dnorm(xfit,mean=mean(res),sd=sd(res))
> lines(xfit,yfit,col="blue",lwd=2)
```



También se puede aplicar el test de Shapiro Wilks

```
> shapiro.test(res)

Shapiro-Wilk normality test

data: res

W = 0.9789, p-value = 0.7811
```

Acepto H0: variable normal

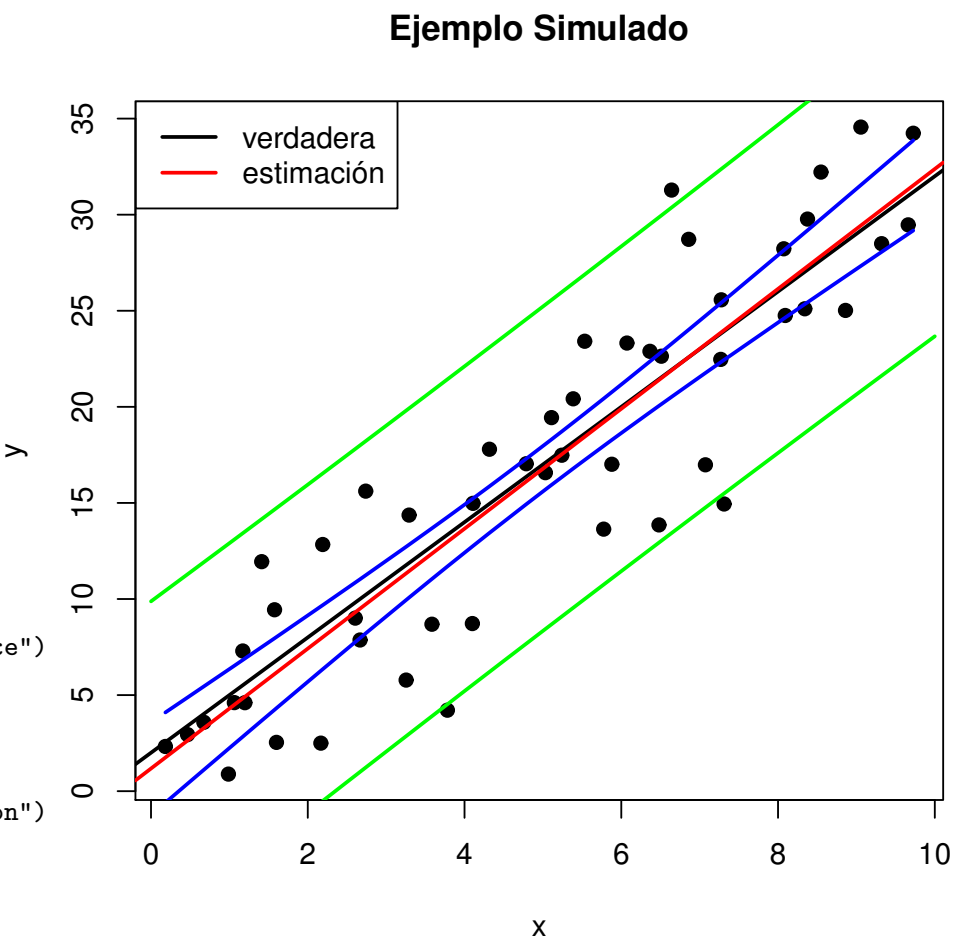
Interpretación y conclusión sobre el modelo

1. los residuos parecerían ser gaussianos e indenticamente distribuidos.
2. El modelo tiene una buena performance explicativa $R^2 = 0{,}9388$ (cerca de 1) y el error residual (residual standard error, RSE), $\widehat{\sigma} = \sqrt{\frac{SCR}{n-2}}$, es bajo $(21, 99)$ por lo que augura buenas predicciones.
3. los errores estandares de $\widehat{\beta}_0$ (3.64) y $\widehat{\beta}_1$ (0.06) son pequeños: esto indica una cierta estabilidad del modelo.
4. El termino constante no es significativamente distinto de cero (podríamos prescindir de él).
5. El coeficiente en X, β_1, es significativamente distinto de cero.
   Otra manera de verlo: el $F = 2269{,}6$. Hay fuerte evidencia de que $\beta_1 \neq 0$.

Intervalo de confianza para la recta de regresión e intervalo de confianza de una predicción

Ojo que acá cambia la simulación:

```
n <- 50
x <- sort(10 * runif(n))
y <- 2 + 3 * x + rnorm(n, sd = 5)
fit <- lm(y ~ x)
plot(x, y, pch = 19)  # datos
abline(2, 3, lwd = 2) # verdadera
abline(coef(fit), lwd = 2, col = 'red') # estimaci\'on
legend("topleft", c("verdadera", "estimaci\'on"),
lty = 1, lwd = 2, col = c(1, 2))

new=data.frame(x=seq(0, 10, .5))
pred=predict(fit, interval="confidence")
pred2=predict(fit,newdata=new,interval="prediction")
lines(x, pred[, 2], col = "blue", lwd = 2)
lines(x, pred[, 3], col = "blue", lwd = 2)
lines(new[,1], pred2[, 2], col = "green", lwd = 2)
lines(new[,1], pred2[, 3], col = "green", lwd = 2)
title("Ejemplo Simulado")

>predict(fit,newdata=data.frame(x=c(5,6)),interval="confidence")
       fit      lwr      upr
1 16.77854 15.59662 17.96046
2 19.89853 18.63624 21.16082

>predict(fit,newdata=data.frame(x=c(5,6)),interval="prediction")
       fit       lwr      upr
1 16.77854  8.338984 25.21810
2 19.89853 11.447340 28.34972
```



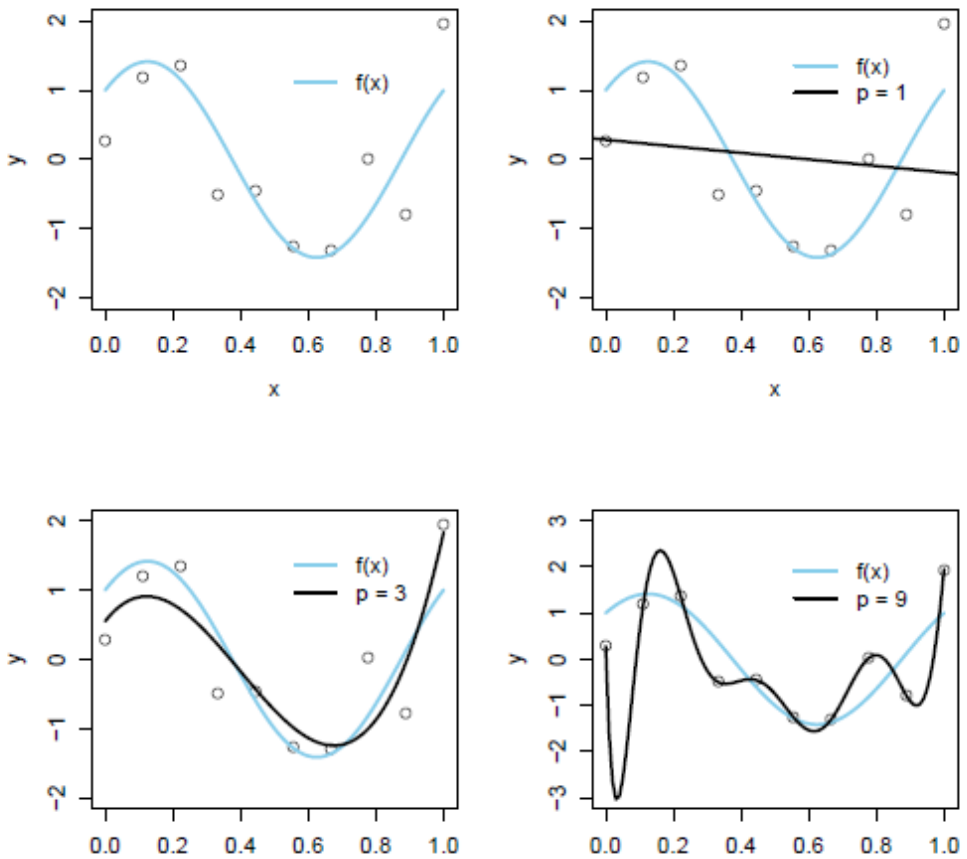
Plan

Introducción. Bishop 2006

Se considera el siguiente modelo de regresión

$$Y = f(X) + \epsilon$$

donde $f(x) = \sin(2\pi x) + \cos(2\pi x)$ y $\epsilon \sim \mathcal{N}(0, 0{,}75^2)$ Queremos ajustar un polinomio del tipo $\widehat{f}(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j$ donde el número de coeficientes p es un parámetro de complejidad del modelo. Se toman $n = 10$ puntos.



El polinomio de grado 1 claramente subajusta los datos: no logra ajustarse bien a la estructura. El polinomio de grado 3 parece ser más adecuado. El de grado 9 sobreajusta los datos ya que los interpola. Esto es un inconveniente a la hora de querer generalizar: el error es nulo sobre la muestra de entrenamiento, pero el error de generalización, es decir sus predicciones sobre nuevas observaciones, será alto.

En el siguiente cuadro se muestran los distintos coeficientes que se obtienen al ajustar el modelo:

$\hat{\beta}_j$	$p=1$	$p=3$	$p=9$
$\hat{\beta}_0$	0.286	0.548	0.279
$\hat{\beta}_1$	-0.473	6.272	-237.909
$\hat{\beta}_2$	0	-30.338	5486.367
$\hat{\beta}_3$	0	25.346	-46686.042
$\hat{\beta}_4$	0	0	203251.273
$\hat{\beta}_5$	0	0	-509682.308
$\hat{\beta}_6$	0	0	765827.927
$\hat{\beta}_7$	0	0	-680299.555
$\hat{\beta}_8$	0	0	329140.427
$\hat{\beta}_9$	0	0	-66798.508
$\sum_{j=0}^{9} \hat{\beta}_j^2$	0.305	1602.479	1.465×10^{12}

El ejemplo anterior nos muestra que agregar más predictores en el modelo lineal implica a menudo un aumento en el tamaño de los coeficientes. Por lo que sería deseable poder controlar esta complejidad acotando de cierta manera el vector, por ejemplo pidiendo que

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

e integrar esta condición a la minimización de la suma de cuadrados residuales SCR.

Regresión Ridge

El estimador ridge del modelo lineal $Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon$ se obtiene como β^R

$$\widehat{\beta}^R = \underset{\beta}{\text{Argmin}} \left(\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{||Y - X\beta||_2^2} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{||\beta||_2^2} \right)$$

lo cual equivale a

$$\begin{cases} \widehat{\beta}^R = \underset{\beta}{\text{Argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{sujeto a que } \sum_{j=1}^{p} \beta_j^2 \leq s \end{cases}$$

Los parámetros λ o s controlan la complexidad del modelo.
Se prueba que si X es centrada e Y también entonces:

$$\widehat{\beta}_\lambda^R = (X'X + \lambda I_p)^{-1} X'Y$$

- Esta solución existe aún si $X'X$ no es invertible.
- Si $\lambda \to 0$ entonces $\widehat{\beta}^R \to \widehat{\beta}^{MC}$.
- Si $\lambda \to \infty$ entonces $\widehat{\beta}^R \to 0$
- Entre los dos buscamos un compromiso entre ajustar el modelo y contraer los coeficientes.

Regresión Ridge. Desarrollo Teórico

Por convención, suponemos que X es centrada e Y también es un vector centrado. Entonces:

- $\widehat{\beta_0} = \sum_{i=1}^{n} y_i/n = 0$ y entonces

$$\mathbb{E}(Y|X) = \sum_{j=1}^{p} \beta_j X_j$$

.

- X tiene p columnas (en vez de $p+1$) LI, la estimación MC da:

$$\widehat{\beta} = (X'X)^{-1}X'Y, \quad \widehat{Y} = X\widehat{\beta} = X(X'X)^{-1}X'Y$$

Observar que

- $\left(\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{||Y - X\beta||_2^2} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{||\beta||_2^2} \right) = (Y - X\beta)'(Y - X\beta) + \lambda ||\beta||_2^2$ es un problema convexo y tiene una única solución. La misma puede tener un error cuadrático medio menor que $\widehat{\beta}^{MC}$.
- La inclusión de λ en la expresión a minimizar hace que el problema sea no singular mismo si $X'X$ es no invertible.
- Para cada λ se tiene una solución distinta. A λ se le llama parámetro de contracción: controla el tamaño de los coeficientes. Se halla λ utilizando validación cruzada.

Obtención de la forma de $\widehat{\beta}^R$:

Observar que $\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\sum_{j=1}^{p}\beta_j^2=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\sum_{j=1}^{p}\left(\sqrt{\lambda}\beta_j\right)^2$ y

podemos considerar la matriz

$$X_\lambda=\begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ x_{n1} & x_{n2} & x_{n3} & \vdots & x_{np}\\ \sqrt{\lambda} & 0 & 0 & \dots & 0\\ 0 & \sqrt{\lambda} & 0 & \dots & 0\\ 0 & 0 & \sqrt{\lambda} & \ddots & 0\\ 0 & 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \dots & \sqrt{\lambda}\end{pmatrix}=\begin{pmatrix} X\\ \sqrt{\lambda}I_p\end{pmatrix}\qquad Y_\lambda=\begin{pmatrix} y_1\\ y_2\\ \vdots\\ y_n\\ 0\\ \vdots\\ 0\end{pmatrix}=\begin{pmatrix} Y\\ 0\end{pmatrix}$$

Entonces, ya que X_λ tiene rango p si $\lambda\neq 0$, la solución minimos cuadrados $\widehat{\beta}^R=\underset{\beta}{\text{Argmin}}||Y_\lambda-X_\lambda\beta||$ es

$$(X_\lambda' X_\lambda)^{-1}X_\lambda' Y=\left(\begin{pmatrix} X' & \sqrt{\lambda}I_p\end{pmatrix}\begin{pmatrix} X\\ \sqrt{\lambda}I_p\end{pmatrix}\right)^{-1}\begin{pmatrix} X' & \sqrt{\lambda}I_p\end{pmatrix}\begin{pmatrix} Y\\ 0\end{pmatrix}=(X'X+\lambda I_p)^{-1}X'Y$$

Elección de λ - Método de validation cruzada

1. Queremos elegir λ de manera a minimizar el MSE.
2. La idea será de entrenar el modelo $\widehat{f}$ sobre un conjunto de entrenamiento y de testearlo sobre una nueva muestra. Un buen estimador será aquel que tenga un buen desempeño sobre una muestra test.
3. Se usa el procedimiento de validación cruzada:
   1. Particionamos el conjunto de muestra T en $T_1, \ldots, T_K$ partes de igual tamaño (en general $K = 10$).
   2. Para cada $k = 1, \ldots, K$:
      - Se entrena el modelo $\widehat{f}^{(\lambda)}_{(-k)}(x)$ sobre $T \setminus T_k$
      - Se predice a partir de $\widehat{f}^{(\lambda)}_{(-k)}$ los valores de las observaciones de T_k.
      - Se calcula el error de validación cruzada: $\mathrm{CV}_k^{(\lambda)} = \frac{1}{|T_k|} \sum_{(x,y)\in T_k} (y - \widehat{f}^{(\lambda)}_{(-k)}(x))^2$
   3. El error del modelo es
   $$\mathrm{CV}^{(\lambda)} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CV}_k^{(\lambda)}$$
   4. Se elige λ^* el valor que minimiza $\mathrm{CV}^{(\lambda)}$
   5. Se reestima el modelo $\widehat{f}^{(\lambda^*)}(x)$ usando toda la muestra T
   6. Se testea $\widehat{f}(x)^{(\lambda^*)}$ sobre un conjunto test para evaluar el error de predicción.

Si $K = n$, el proceso se llama *leave one out cross validation* o *Jacknife*.

Regresión Lasso

La regresión Lasso (*Least Absolute Shrinkage and Selection Operator*) combina la regressión Ridge con la selección de variables. En efecto la regresión Ridge incluye en el modelo final todos los predictores, contrayendo varios de ellos a 0, pero no todos serán iguales a 0. Esto no es tanto un problema en cuanto a la predicción, si no más bien en cuanto a la interpretación del modelo.
Regresión lasso:

$$\widehat{\beta}^L = \underset{\beta}{\text{Argmin}} \left(\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{||Y - X\beta||_2^2} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{||\beta||_1} \right)$$

lo cual equivale a

$$\begin{cases} \widehat{\beta}^L = \underset{\beta}{\text{Argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{sujeto a que } \sum_{j=1}^{p} |\beta_j| \leq s \end{cases}$$

Como en la regresión Ridge, la regresión Lasso contrae alguno de los coeficientes a ser exactamente 0 para valores de λ suficientemente grande y por lo tanto puede ser considerado como un método de selección de variable y hace que el modelo sea fácil de interpretar. Se dice que los modelos lasso producen *modelos esparses* (*sparse modelos*), que involucran únicamente un subconjunto de variables.

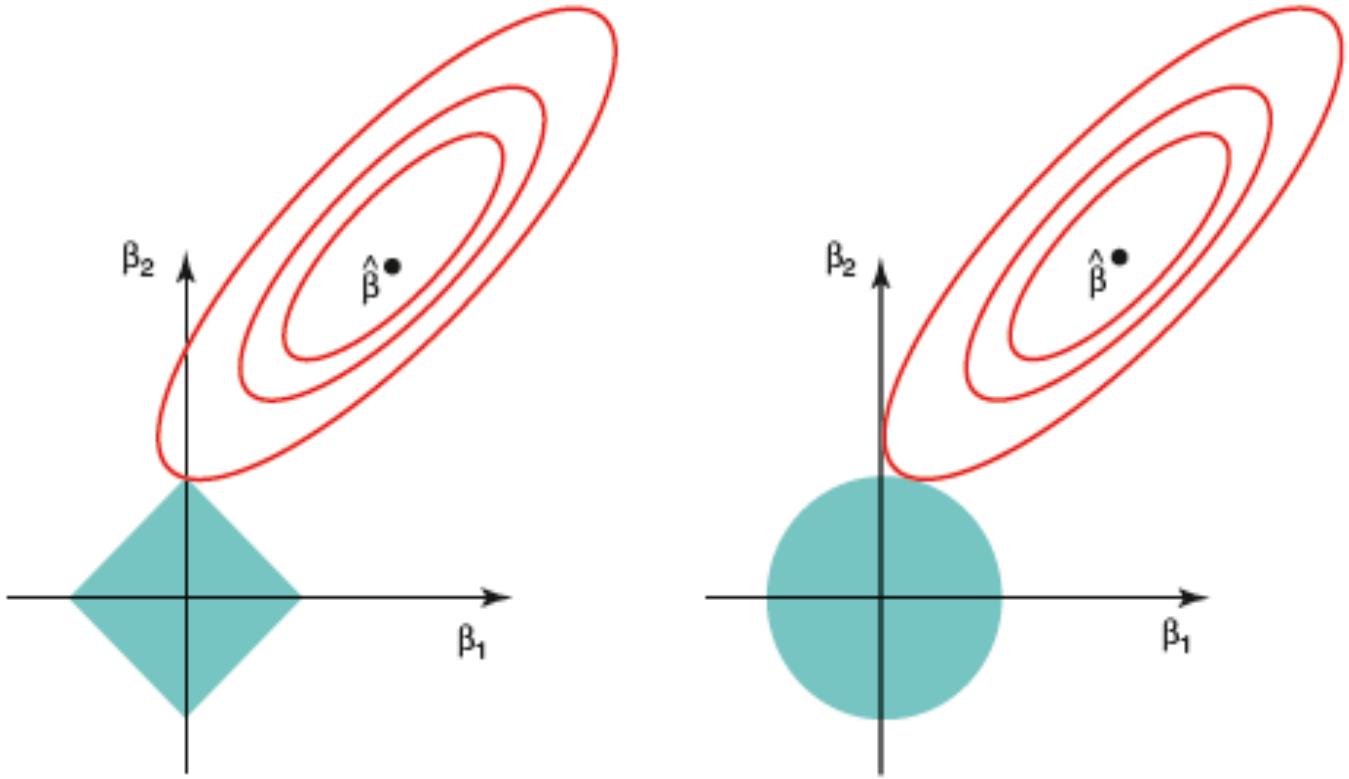


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

ISLR, pág 222.

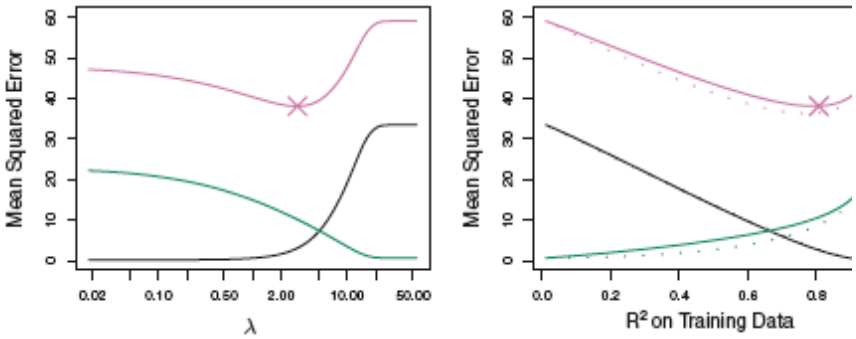


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test (purple) for the lasso on a simulated data set.* Right: *Comparison of squared variance and test MSE between lasso (solid) and ridge (dashed). Both are pl against their R^2 on the training data, as a common form of indexing. The cr in both plots indicate the lasso model for which the MSE is smallest.*

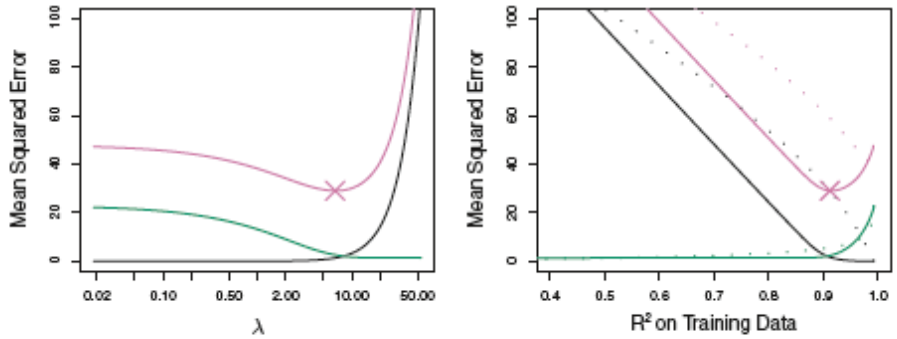


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

ISLR, pág 222-223.

Referencias

- A. I. Izenman, *Modern Multivariate Statistical Techniques*, Springer, 2008.
- F. Carmona, *Modelos Lineales*, notas de curso, Universitat de Barcelona, 2003.
- C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
- S.J. Sheater, *A Modern Approach to Regression with R*, Springer, 2009.
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.
- M. Bourel. Apuntes curso Estadística Multivariada Computacional 2018, 2019. Facultad de Ingeniería.