Fundamentos de regresión lineal

German Rosati
german.rosati@gmail.com

UNTREF - UNSAM - CONICET

27 de marzo de 2019

Hoja de ruta

Temas a tocar

Modelos de regresión lineal y logística

- Forma funcional
- Interpretación de parámetros
- Evaluación
- Problemas...

Repaso...

¿Qué es un modelo?

- Forma de proponer hipótesis sobre la forma en que se combinan variables
- En general, tienen esta forma

$$Y = f(X) + \epsilon \tag{1}$$

- Problema: estimar $f(X)$
  - Suponer que Y es una combinación lineal de las X
  - A su vez Y es una variable cuantitativa
  - Terreno propicio para la **regresión lineal**

- En una regresión lineal, asumimos que la relación entre los predictores X y la respuesta Y toma la siguiente forma

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \quad (2)$$

- En general, buscaremos estimar los parámetros del modelo, por lo cual, dadas las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_j$, definimos una nueva predicción del valor $\hat{Y}$ como

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p + \epsilon \quad (3)$$

- β_0 es el intercepto y β_j es la pendiente para la variable X_j

- Si $\hat{y}_i$ representa la predicción para el i-ésimo caso condicionado a los valores de X_i, $\implies$ $\epsilon_i = (\hat{y}_i - y_i)$ es el *residuo* para ese i-ésimo caso
- Definimos a la suma de los residuos al cuadrado (RSS) como $RSS = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + ... + \epsilon_n^2$. o, de forma equivalente
- Buscamos estimar los valores de $\beta_0, \beta_1, ..., \beta_p$ como el valor que minimiza el RSS

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - ... - \hat{\beta}_p X_p)^2 \end{aligned} \tag{4}$$

- *RSS* es una medida de la variabilidad de la variable dependiente que NO es explicada por nuestro modelo... nuestra "ignorancia"

Regresión Lineal

Ejemplo 1

- Modelo simple: predecir las ventas de una emprsa en función del gasto que realizan en publicidad para TV
- ¿Cuántas variables hay?
- ¿Qué función cumple cada una?

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ sales &= \beta_0 + \beta_1 x TV + \epsilon \end{aligned} \tag{5}$$

- Asumimos que la relación es *lineal*

Regresión Lineal

Ejemplo 1

- Infinitas rectas pasan por esta nube de puntos
- Cada una se caracteriza por un set de parámetros: (β_0, β_1)
- Queremos encontrar la que mejor ajusta: la que minimiza el *RSS*

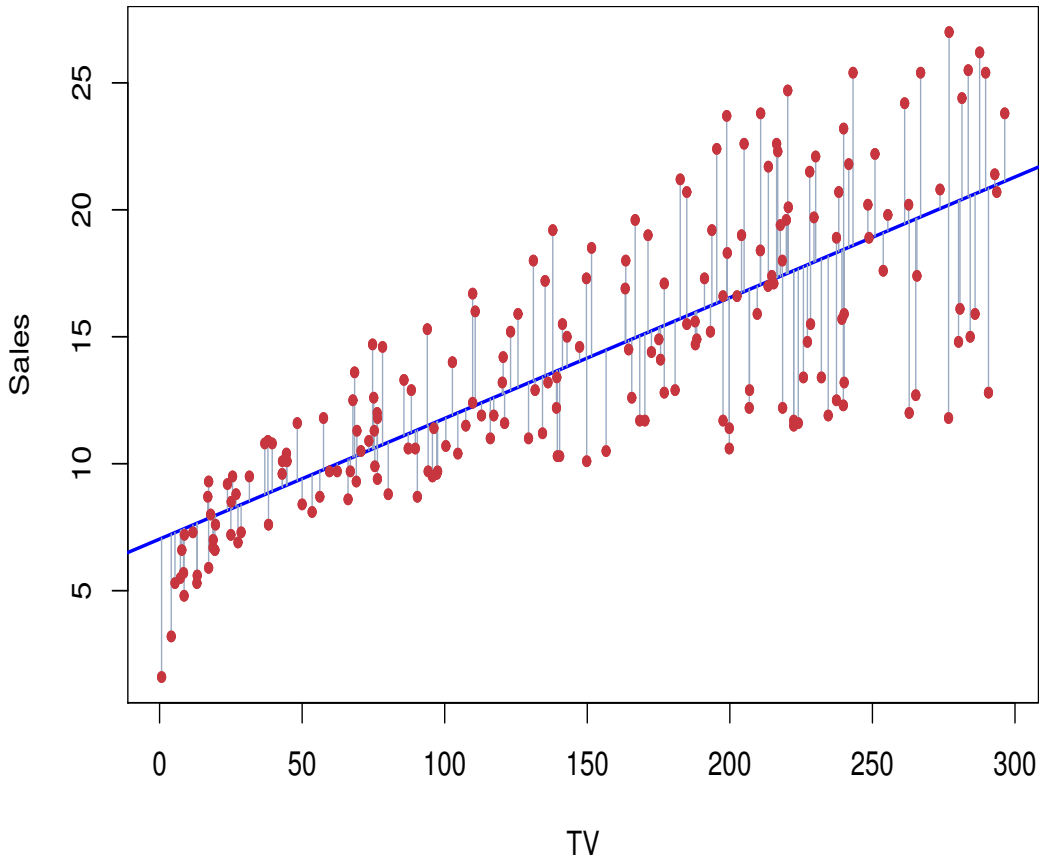

Figura: Scatter plot de gasto en TV y ventas [1]

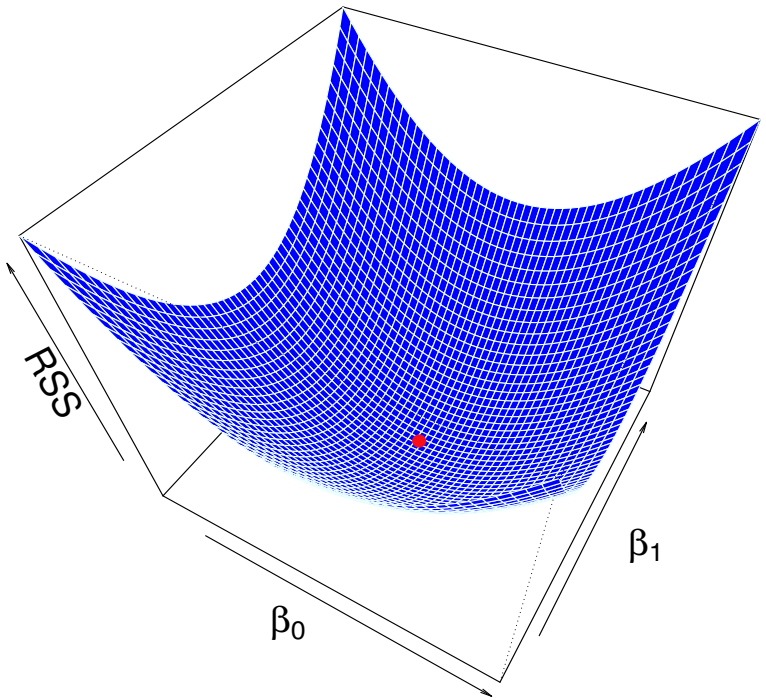

Figura: Esquema de valores de *RSS* en función de β_0 y β_1 [1]

- Para cada combinación (β_0, β_1) se podría calcular su *RSS* correspondiente.
- Intuición: pruebo todas las combinaciones posibles de (β_0, β_1) y elijo la de menor *RSS*

Regresión Lineal

Ejemplo 1

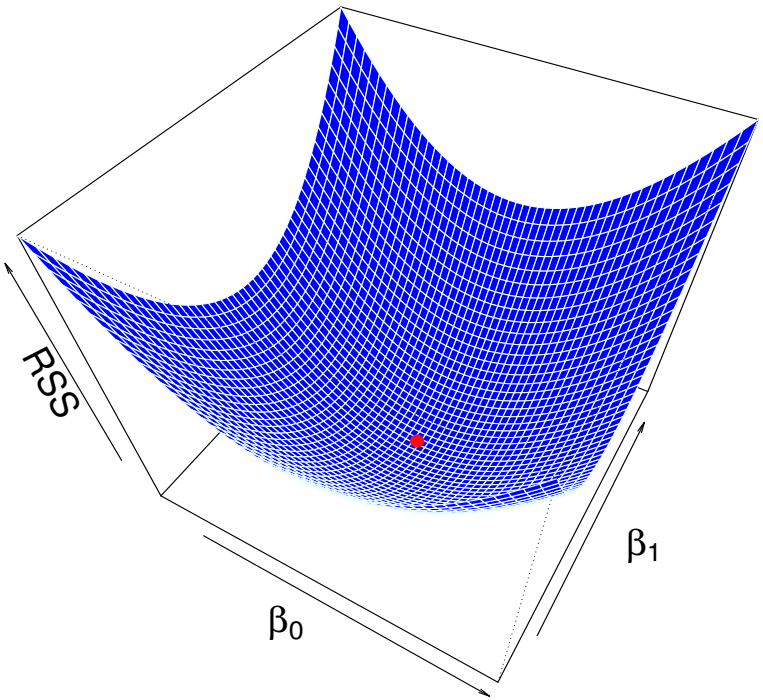

Figura: Esquema de valores de *RSS* en función de β_0 y β_1 [1]

- En regresión lineal forma analítica de resolucion. Derivando sobre la ecuación del *RSS* se obtienen las ecuaciones normales:

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (6)$$
$$\beta_0 = \bar{y} - \beta_1 \bar{X}$$

- Otros métodos... Descenso de gradiente

Regresión Lineal

Ejemplo 1 - Coeficientes

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Figura: Coeficientes β estimados para TV y Sales [1]

- ¿Cómo se interpretan?: β_p Efecto marginal
- ¿Qué significa cada columna de la tabla?

Intervalo de confianza

- Error estándard:

$$SE(\beta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

- Intervalo de confianza:

$$\beta_1 \pm 2 \times SE(\beta_1) \tag{8}$$

- Hay un aproximadamente un 95 % de chances de que el intervalo contenga el valor verdadero del parámetro.
- Para el ejemplo, el intervalo de confianza de 95 % para β_1 es $[0{,}042, 0{,}053]$

Test de hipótesis

- Hipótesis más común:
  - H_O : No hay relación entre X y Y — $\beta_1 = 0$
  - H_A : Hay relación entre X y Y — $\beta_1 \neq 0$
- Estadístico t:

$$t = \frac{\beta_1 - 0}{SE(\beta_1)} \tag{9}$$

Muchas medidas:

- Coeficiente de determinación:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{10}$$

  dónde $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  Es decir, la proporción de la varianza de la variable dependiente que es explicada por el modelo
- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Regresión Lineal Múltiple

Volvamos al comienzo...

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{11}$$

- Interpretamos cada β_j como el efecto promedio sobre Y *mantiendo todos los demás factores -X's- constantes*
- Nuestro modelo se transforma en

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times diarios + \epsilon \tag{12}$$

Regresión Lineal Múltiple

Interpretación

- *Escenario ideal:*
  - Diseño balanceado
  - Cada coeficiente puede ser interpretado y testeado de forma separada
  - Las interpretaciones estilo "efecto de X_j sobre Y manteniendo el resto constante" son viables
- La correlación entre los predictores trae problemas
  - La varianza de los predictores tiende a aumentar
  - Las interpretaciones se vuelven más imprecisas: no funciona el "manteniendo todo lo demás constante" porque cuando X_j cambia, el resto también
- Abstenerse de hacer interpretaciones causales

Regresión Lineal Múltiple

Ejemplo 2

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Figura: Coeficientes β estimados para TV, radio, diarios y Sales [1]

- ¿Es al menos uno de los predictores relevantes para la predicción de y?
  - Estadístico F:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{13}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Figura: Prueba F y R^2 para regresión completa [1]

Regresión Lineal Múltiple

Evaluación: ¿Qué variables son importantes?

- Área más amplia llamada *model selection*
- Muchos enfoques y técnicas
  - **Best Subset Selection:** ajustamos todos los modelos posibles para todos los subsets de variables y elegimos el mejor basado en alguna métrica de error.
    Problema: generalmente no podemos examinar TODOS los modelos posibles. Hay 2^p modelos posibles. Para $p = 40$ hay más de 1.000.000.000 de modelos.
    Necesitamos algún enfoque de selección automático y que achique el espacio de búsqueda

- **Forward Selection:**
  1. Empezamos con modelo nulo: contiene solo β_0
  2. Fiteamos p modelos simples (un predictor) y elegimos el mejor en términos de RSS
  3. Agregamos a ese modelo la variable que mejor funciona en un modelo de dos variables
  4. Seguimos hasta que se cumple algún criterio de corte

- **Backward Selection:**
  1. Empezamos con un modelo con todas las variables
  2. Eliminamos la variable con mayor p-valor en la prueba t
  3. Fiteamos el nuevo modelo con $(p-1)$ variables y volvemos a eliminar la variable con mayor p-valor
  4. Seguimos hasta que se cumple algún criterio de corte

Regresión Lineal Múltiple

Evaluación: ¿Qué variables son importantes?

- Hay muchos criterios para elegir un modelo "óptimo" en el camino de modelos construidos por la selección Forward o Backward
  - *Cp de Mallow*
  - *Akaike Information Criteria* (AIC)
  - *Bayesian Information Criteria* (BIC)
  - R^2 *ajustado*
  - *Cross-Validation*

- Pese a su simplicidad, es posible agregar complejidad a un modelo lineal
- Podemos modelar la interacción entre predictores a través del producto $X_1 X_2$
- El modelo adquiere la siguiente forma

$$\begin{aligned} y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + +\beta_3 X_1 X_2 + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{14}$$

- *Principio jerárquico:* si se incluye una interacción en el modelo, debe incluirse también los efectos de orden menor (aún cuando los p-valores no sean significativos)

Extensiones del Regresión Lineal

Interacciones entre predictores - Ejemplo 3

- Supongamos que en el modelo anterior que el impacto de un incremento del gasto en TV afecta de alguna forma la efectividad de gasto en radio
- En esta situación, dado un presupuesto fijo quizás sea más efectivo alocarlo en ambas variables
- Efecto interacción

- Nuestro modelo:

$$\begin{aligned} sales &= \beta_0 + \beta_1 \times TV + \beta_2 \times radio + +\beta_3 \times radio \times TV + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon \end{aligned} \tag{15}$$

- Resultados

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Figura: Resultados de regresión con términos de interacción[1]

- ¿Es relevante la interacción?
- $R^2 = 0{,}98$ ¿Qué pueden decir del ajuste respecto al modelo anterior?
- Los coeficientes sugieren que un incremento de la publicidad en TV de \$1.000 se asocia con un incremento en las ventas de $(\beta_1 + \beta_3 \times \textit{radio}) \times 1{,}000 = 19 + 1{,}1 \times \textit{radio}$

- Diferentes formas funcionales:
  - Log-Log:

$$ln(y_i) = \beta_0 + \beta_1 ln(X_i) + \epsilon \tag{16}$$

  - Inversas:

$$y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \epsilon \tag{17}$$

- Regresiones polinómicas

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + ... + \beta_p X_i^p + \epsilon \tag{18}$$

Extensiones del Modelo Lineal

Predictores cualitativos

- Algunos predictores no son cuantitativos sino categóricos
- Ejemplo: pensemos en modelar el monto de deuda de una persona en función de la condición de estudiante y del ingreso
  - Creamos una nueva variable

$$genero_i = \begin{cases} 1 & \text{si la i-ésima persona es mujer} \\ 0 & \text{si la i-ésima persona es estudiante} \end{cases}$$

  - El modelo resulta en

$$deuda_i \approx \beta_0 + \beta_1 \times ingreso + \beta_2 \times estud$$

- Entonces,

$$\begin{aligned} deuda \approx & \beta_0 + \beta_1 \times ingreso + \begin{cases} \beta_2 & \text{if } estud = 1 \\ 0 & \text{if } estud = 0 \end{cases} \\ \approx & \beta_1 \times ingreso + \begin{cases} \beta_0 + \beta_2 & \text{if } estud = 1 \\ \beta_2 & \text{if } estud = 0 \end{cases} \end{aligned} \tag{19}$$

- Si hubiera interacción...,

$$
\begin{aligned}
deuda \approx & \beta_0 + \beta_1 \times ingreso + \begin{cases} \beta_2 + \beta_3 \times income & \text{if } estud = 1 \\ 0 & \text{if } estud = 0 \end{cases} \\
\approx & \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times income & \text{if} estud = 1 \\ \beta_0 + \beta_1 \times income & \text{if } estud = 0 \end{cases}
\end{aligned} \tag{20}
$$

Extensiones del Modelo Lineal

Predictores cualitativos

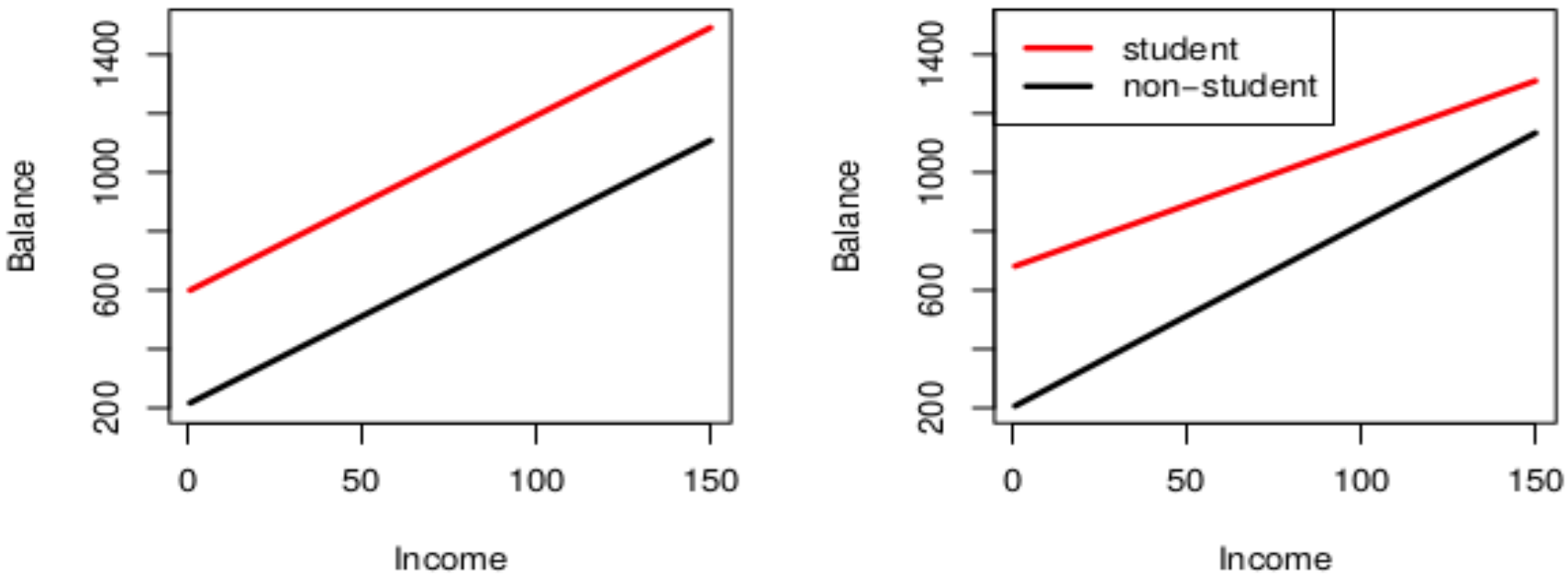

Figura: IZQ: sin interacción ; DER: con interacción [1]

- ¿Qué hacemos con el overfitting?
- Una forma es hacer *model selection*...
- Otra es utilizar técnicas de regularización.
- El objetivo es introducir una restricción en la función de costo (*RSS* - la función que se minimiza) y con eso forzar a los β a reducir su valor
- **Ridge:**

$$CF = \sum_{i=1}^{N}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \beta_j^2 = RSS + \lambda \beta_j^2 \qquad (21)$$

- **LASSO:**

$$CF = \sum_{i=1}^{N}(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda |\beta_j| = RSS + \lambda |\beta_j| \qquad (22)$$

Modelo Lineal y Overfitting

Regularización

- Se parte de la minimización de *RSS* habitual + una restricción $\lambda|\sum_{j=1}^{p}\beta_j|$ que tiene el efecto de reducir los coeficientes β_j estimados
- λ es un hiperparámetro del modelo que controla el impacto de la penalización y se estima mediante *cross validation*
- Ridge se inventó originalmente para lidiar con el problema de la multicolinealidad. Sesga los coeficientes para reducir la varianza
- Ambos "encogen" los coeficientes hacia cero. LASSO, además, hace que algunos sean iguales a cero
- LASSO, entonces, realiza *model selection*
- Es una formalización de un proceso que muchas veces se hace artesanalmente

Referencias bibliográficas I

James, G., Witten, D., Hastie, T., and Tibshirani, R.
An Introduction to Statistical Learning – with Applications in R, vol. 103 of *Springer Texts in Statistics*.
Springer, New York, 2013.