Introducción a la Inteligencia Predictiva en R con aplicaciones:

Métodos de regresión.

Diego J. Pedregal

Universidad de Castilla-La Mancha

Diego.Pedregal@uclm.es

Universidad de Sevilla

15-16 de Noviembre de 2018

Regresión lineal

- Una regresión lineal modeliza la relación entre una variable endógena y (un vector que contiene todos los valores de los n elementos de una variable) y un conjunto de exógenas $x_1, x_2, \dots, x_k$.
- Es un modelo muy sencillo, pero muy útil conceptualmente, y sobre él se basan una gran cantidad de técnicas más sofisticadas.
- El modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

- $\beta_i, i = 1,2, \dots, k$ son los parámetros o coeficientes desconocidos del modelo, y ϵ es el término de error.

PRÁCTICA #1.4

Regresión lineal

- Si tuviéramos estimaciones de los coeficientes ($\widehat{\beta_i}, i = 1,2, \dots, k$) podríamos calcular los valores ajustados para la variable endógena

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \cdots + \widehat{\beta_k} x_k$$

- El vector de residuos es $\hat{e} = y - \hat{y}$. Un residuo individual será $\widehat{e_i} = y_i - \widehat{y_i}$
- La suma residual (al cuadrado) es

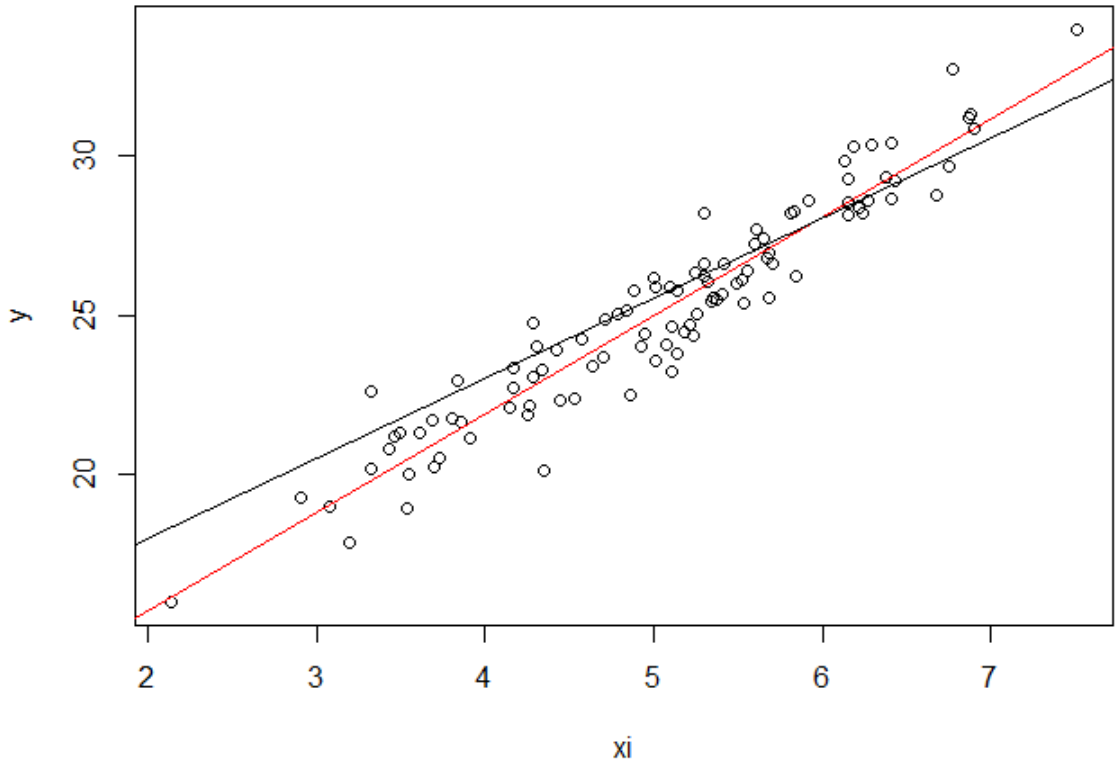
$$SCR = \sum_{i=1}^{n} \widehat{e_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \hat{e}'\hat{e}$$

- La estimación de los parámetros se puede llevar a cabo minimizando esta expresión para una muestra determinada.

Regresión lineal: Estimación

- Estimación: Llamando $X = (1 \quad x_1 \quad x_2 \quad \dots \quad x_k)$, tenemos que la estimación por mínimos cuadrados es

$$\hat{\beta} = (X'X)^{-1}X'y$$



PRÁCTICA #1.5

Regresión lineal: Estimación

- Incertidumbre en estimación de parámetros

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$\sigma^2 = Var(\epsilon)$$

- Intercambiando los valores teóricos por estimados de la varianza del error

$$\hat{\sigma}^2 = \hat{e}'\hat{e}/(n-k-1)$$

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

- Esto nos permite construir intervalos de confianza de los parámetros, puesto que el error estándar de un parámetro individual $\hat{\beta}_i$ será el elemento i-ésimo de la matriz $\hat{\sigma}^2 (X'X)^{-1}$.

Regresión lineal: Inferencia

- El intervalo de confianza al 95% para un parámetro individual será

$$\hat{\beta}_i \pm 2 \times ES(\hat{\beta}_i)$$

- Esto quiere decir que tenemos una probabilidad del 95% de encontrar el verdadero valor del parámetro en dicho intervalo en muestras repetidas.
- En la misma línea se pueden hacer contraste de hipótesis. El más típico es
  - H_0: $\beta_i = 0$. NO hay relación entre la variable exógena i-ésima y la variable endógena.
  - H_A: $\beta_i \neq 0$. SÍ hay relación entre la variable exógena i-ésima y la variable endógena.

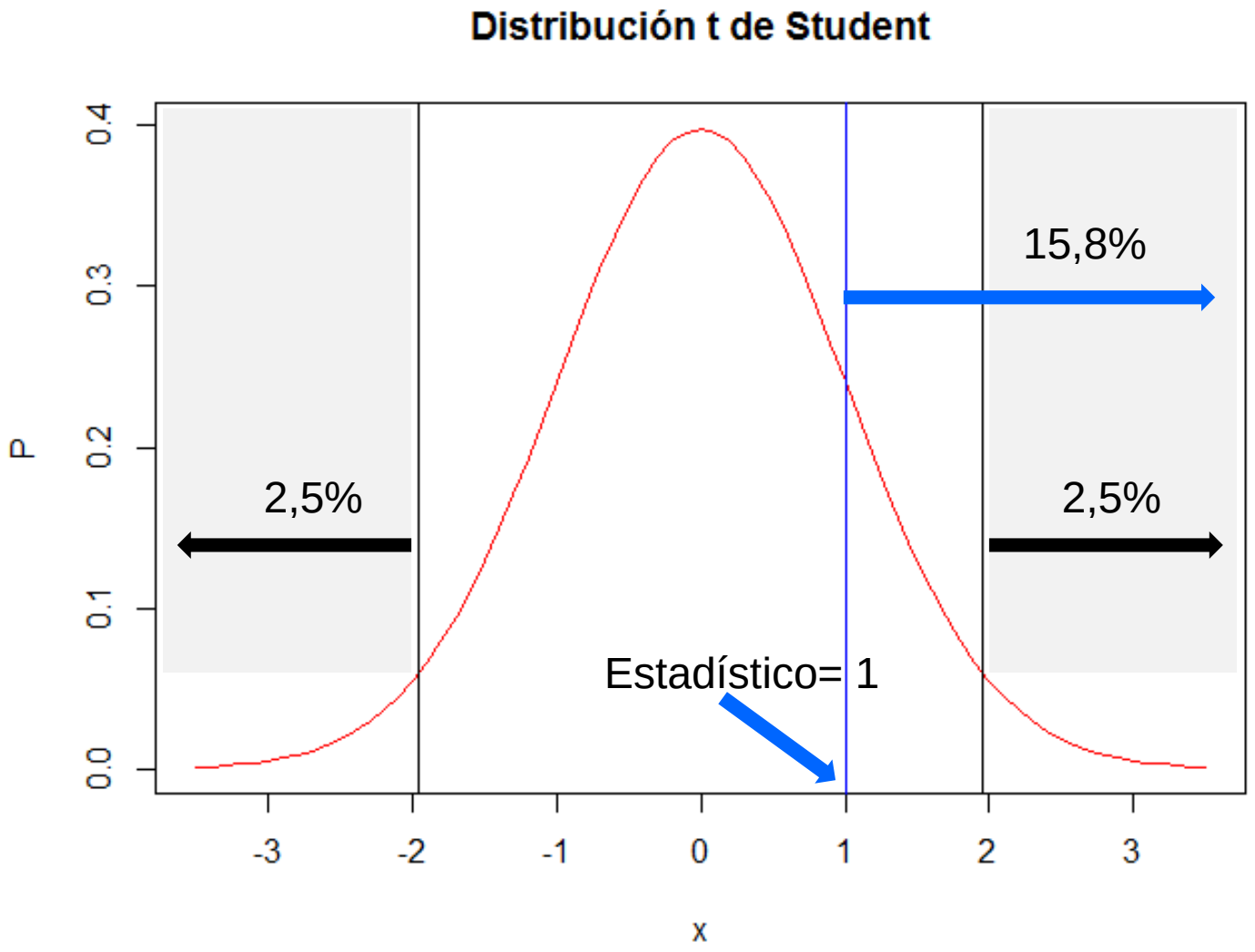
Regresión lineal: Inferencia

- Para contrastar esa hipótesis calculamos el estadístico t

$$t_i = \frac{\hat{\beta}_i - 0}{ES(\hat{\beta}_i)}$$

- Este estadístico se distribuye, bajo la hipótesis nula, como una t de Student con $n - k - 1$ grados de libertad.
- La valor que toma el estadístico se llama también el valor crítico y la probabilidad que deja el estadístico a ambos lados de la distribución correspondiente es el p-valor (p-value en inglés).

Regresión lineal: Inferencia



Regresión lineal: Inferencia

- Considerando $\mathrm{ST} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, podemos definir el coeficiente R^2 como una medida de ajuste comprendida entre 0 (ajuste nulo) y 1 (ajuste perfecto)

$$R^2 = \frac{ST - SCR}{ST} = 1 - SCR/ST$$

- Un último test es el test de significación de todos los parámetros simultáneamente o test de la F.
  - H_0: $\beta = 0$. NO hay relación entre las variables exógenas y la variable endógena.
  - H_A: $\beta \neq 0$. SÍ hay relación entre las variables exógenas y la variable endógena.

PRÁCTICA #1.6

Regresión lineal: Interpretación

- La interpretación de los coeficientes es como derivadas parciales, es decir, la variación media que experimenta cada variable endógena, cuando la exógena varía en una unidad manteniendo todas las demás variables fijas.
- Cuando las variables exógenas son independientes no hay problemas.
- La interpretación con exógenas no independientes es problemática, puesto que cuando una variable cambia, cambian todas las demás.

Regresión lineal: Interpretación

- A menudo se utilizan especificaciones logarítmicas…

$$\log(y) = \beta_0 + \beta_1 \ln(x_1) + \cdots + \beta_k \ln(x_k) + \epsilon$$

- … que implican que el modelo inicial es exponencial

$$y = e^{\beta_0} x_1^{\beta_1} \dots x_k^{\beta_k} e^{\epsilon}$$

- En este modelo los coeficientes se pueden interpretar como elasticidades (variación porcentual que experimenta la variable endógena cuando la exógena varía en 1%)

$$\frac{\partial y}{\partial x_i}\frac{x_i}{y} = \beta_i$$

$$\frac{\partial y}{\partial x_i}\frac{x_i}{y} = \frac{e^{\beta_0} x_1^{\beta_1} \dots \beta_i x_i^{\beta_i - 1} \dots x_k^{\beta_k} e^{\epsilon} x_i}{e^{\beta_0} x_1^{\beta_1} \dots x_k^{\beta_k} e^{\epsilon}} = \beta_i$$

Regresión lineal: Interpretación

- Para la especificación *lineal-log*:

$$y = \beta_0 + \beta_1 \ln(x_1) + \cdots + \beta_k \ln(x_k) + \epsilon$$

$$\Delta y = \beta_i \ln\left(1 + \frac{\%\Delta x_i}{100}\right)$$

- Para la especificación *log-lineal*:

$$\ln(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

$$\%\Delta y = \left(e^{\beta_i \Delta x_i} - 1\right) \times 100$$

Regresión lineal: Exógenas cualitativas

- Las variables cualitativas se definen como variables que toman siempre valores binarios (0/1). La regresión no presenta ningún problema, siempre que este tipo de variables sean explicativas. Por ejemplo, si queremos estimar un modelo que relacione ingresos personales diferenciando entre hombres y mujeres se puede definir la variable

$$x_1 = \begin{cases} 1 & \text{para mujer} \\ 0 & \text{para hombre} \end{cases}$$

- Siendo y el ingreso personal de los individuos de una muestra, ¿cuál es la interpretación del modelo?

$$y = \beta_0 + \beta_1 x_1 + \epsilon = \begin{cases} \beta_0 + \beta_1 & \text{para mujer} \\ \beta_0 & \text{para hombre} \end{cases}$$

Regresión lineal: Exógenas cualitativas

- Cuando se necesitan más de dos categorías se añaden más variables binarias. Siempre una menos que categorías. Por ejemplo, para tres razas (americana, asiática y europea).

$$x_1 = \begin{cases} 1 & \text{para americana} \\ 0 & \text{para no americana} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{para asiática} \\ 0 & \text{para no asiática} \end{cases}$$

- El modelo ahora sería

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = \begin{cases} \beta_0 + \beta_1 & \text{americana} \\ \beta_0 + \beta_2 & \text{asiática} \\ \beta_0 & \text{europea} \end{cases}$$

PRÁCTICA #1.7

Regresión lineal: Selección de variables

- Cuando tenemos muchas variables candidatas a ser regresores posibles necesitamos algún procedimiento para decidir cuántas y qué variables son relevantes. Hay varias formas:
  - Selección de un subconjunto óptimo de acuerdo con algún criterio.
  - Shrinkage ("estrujamiento"), por ejemplo, regularización.
  - Reducción de la dimensión mediante proyecciones.
- Estas ideas no son exclusivas de la regresión.

Regresión lineal: Selección de variables

- Criterios de selección que penalizan el uso de muchos parámetros:
  - Cp de Mallow:

$$C_p = \frac{1}{n}(SCR + 2k\hat{\sigma}^2)$$

  - Criterio de información de Akaike(AIC):

$$AIC = \frac{1}{n\hat{\sigma}^2}(SCR + 2k\hat{\sigma}^2)$$

  - Bayesian Information Criterion (BIC):

$$BIC = \frac{1}{n\hat{\sigma}^2}(SCR + \ln(n)k\hat{\sigma}^2)$$

  - R^2 ajustado:

$$R^2\text{ajustado} = 1 - \frac{SCR/(n-k-1)}{ST/(n-1)}$$

Regresión lineal: Selección de variables

- Selección de conjunto de variables óptimo:
  - Bruto: Se estiman todos los modelos posibles con número de regresores creciente y se elige en cada caso el mejor. El mejor modelo será el que dé la mejor función objetivo.
    1. Se estima modelo con constante solo
    2. Para i= 1, 2, …, k
       a. Ajustar todos los modelos posibles con i regresores.
       b. Elegir el mejor de ellos, es decir, el que minimice SCR.
    3. Seleccionar el mejor modelo de todos los anteriores mediante un criterio de selección Cp, AIC, BIC o R^2 ajustado.
  - Forward step wise: El paso 2. anterior se sustituye por estimar solo los modelos con un regresor adicional al modelo seleccionado para i-1.
  - Backward step wise: Igual que el anterior, pero comenzando por el modelo completo y reduciendo variables de una en una.

PRÁCTICA #1.8

Regresión lineal: Selección de variables

- *Shrinkage*: La suma residual (al cuadrado) de una regresión es

$$SCR = \sum_{i=1}^{n} \widehat{e_i}^2 = \sum_{i=1}^{n} \left(y_i - X\widehat{\beta_i}\right)^2 = \hat{e}'\hat{e}$$

  - La función objetivo con *shrinkage* impone una restricción en los valores de los parámetros, de forma que tiendan a cero. Una posibilidad es la regularización:

$$SCR = \sum_{i=1}^{n} \left(y_i - X\widehat{\beta_i}\right)^2 + \lambda \sum_{i=1}^{n} \left(\widehat{\beta_i}\right)^2$$

  - Regularización LASSO:

$$SCR = \sum_{i=1}^{n} \left(y_i - X\widehat{\beta_i}\right)^2 + \lambda \sum_{i=1}^{n} |\widehat{\beta_i}|$$

Regresión lineal: Selección de variables

- Regularización *Elastic-Net*:

$$SCR = \sum_{i=1}^{n} \left(y_i - X\widehat{\beta}_i\right)^2 + \lambda_1 \sum_{i=1}^{n} \left(\widehat{\beta}_i\right)^2 + \lambda_2 \sum_{i=1}^{n} |\widehat{\beta}_i|$$

- Si $\lambda = 0$ se obtiene como caso particular MCO.
- Existe un valor suficientemente grande de λ para el que todos los parámetros son cero.
- Para un $\lambda \neq 0$ implica que algunas variables desaparecen, pero otras están atenuadas respecto a la versión sin restringir.
- λ se puede determinar de muchas formas, siendo una muy habitual la validación cruzada (*cross validation*).

Regresión lineal: Selección de variables

La validación cruzada aproxima el verdadero error de predicción de la siguiente forma:

1. Se divide la muestra en K partes (*folds*).
2. Se ajusta el modelo en 9 de ellas y se predice la que se ha excluido.
3. Se repite el proceso hasta que todas se han predicho.
4. El error de validación cruzada es el que resulta de unir todos los errores de cada una de las K partes.

PRÁCTICA #1.9