Tema 5. Remuestreos en Modelos Lineales y Series Temporales

basado en

B. Efron, R. Tibshirani (1993). An Introduction to the bootstrap.

O. Kirchkamp (2019). Resampling methods.

Curso 2022/2023

Introducción a la Regresión Lineal

- En el modelo clásico de regresión lineal se tiene un conjunto de n parejas de observaciones $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ tal que cada $\mathbf{z}_i$ es un par $\mathbf{z}_i = (\mathbf{x}_i, y_i)$.
- Cada $\mathbf{x}_i$ es un vector de dimensión p tal que $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ se suele denominar como vector de *covariables* o *predictores*.
- y_i es un número real denominado *respuesta*.
- Se define la esperanza condicional de la respuesta y_i dado el predictor $\mathbf{x}_i$ como

$$\mu_i = E\left(y_i | \mathbf{x}_i\right)$$

para $i = 1, 2, \ldots, n$.

Introducción a la Regresión Lineal

- La suposición básica de los modelos lineales es que μ_i es una combinación lineal de los componentes del vector $\mathbf{x}_i$

$$\mu_i = \mathbf{x}_i \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij} \beta_j$$

- El vector de parámetros $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ es desconocido de modo que se trata de estimarlo mediante los datos observados $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$.

- El término *lineal* se refiere a la forma lineal de la esperanza, no a que las términos de $\mathbf{x}_i$ puedan estar elevados a un exponente dado.

- La estructura habitual es (para $i = 1, 2, \ldots, n$)

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

Introducción a la Regresión Lineal

- Los términos de error ε_i se asume que proceden de una distribución desconocida F que tiene esperanza igual a 0:

$$F \to (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$$

tal que $E_F(\varepsilon_i) = 0$.

- Esto implica que

$$E\left(y_i|\mathbf{x}_i\right) = E\left(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i|\mathbf{x}_i\right) = E\left(\mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i\right) + E\left(\varepsilon_i|\mathbf{x}_i\right) = \mathbf{x}_i\boldsymbol{\beta}$$

- Ya que al ser ε_i independientes de $\mathbf{x}_i$ entonces

$$E\left(\varepsilon_i|\mathbf{x}_i\right) = E\left(\varepsilon_i\right) = 0$$

Introducción a la Regresión Lineal

- Para estimar los parámetros de la regresión β a partir de los datos originales, se toma un valor inicial, digamos $\mathbf{b}$ de β,

$$\mathrm{ECM}(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i \mathbf{b})^2$$

- De modo que el estimador de mínimos cuadrados de β es el valor $\widehat{\beta}$ de $\mathbf{b}$ que minimiza el error cuadrático medio

$$\boxed{\mathrm{ECM}(\widehat{\beta}) = \min_{\mathbf{b}} \left(\mathrm{ECM}(\mathbf{b})\right).}$$

Introducción a la Regresión Lineal

- Se define la llamada *matriz de diseño* como $\mathbf{X}$, de orden $n \times p$, tal que la fila $i-$ésima es $\mathbf{x}_i$, y se denomina $\mathbf{y}$ al vector $(y_1, y_2, \ldots, y_n)'$
- Entonces el estimador de mínimos cuadrados es la solución de las *ecuaciones normales*

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

- es decir

$$\boxed{\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

Introducción a la Regresión Lineal con R

- En R hay muchos paquetes estadísticos que permiten trabajar con métodos de regresión.
- La orden básica en R es `lm`.
- Ver, por ejemplo, como tutoriales:

  Curso completo sobre métodos de regresión con R:

  http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad3/nreg1.pdf

  Tutorial corto sobre métodos de regresión con R:

  http://www.montefiore.ulg.ac.be/~kvansteen/GBIO0009-1/ac20092010/Class8/Using%20R%20for%20linear%20regression.pdf

Bootstrap en Regresión Lineal

- La aplicación al modelo de regresión lineal simple sirve como base para otros modelos más complejos.
- El modelo de probabilidad $P \to \mathbf{z}$ para la regresión lineal tiene dos componentes: $P = (\beta, F)$ donde β es el vector de parámetros de la regresión y F es la distribución de los errores.
- En principio, se dispone del estimador de $\widehat{\beta}$ de mínimos cuadrados. Pero hace falta estimar F.
- Si β fuese ***conocido*** entonces se podrían calcular los errores como $\varepsilon_i = y_i - \mathbf{x}_i\beta$ para $i = 1, 2, \ldots, n$ y se estimaría F mediante su distribución empírica.

Bootstrap en Regresión Lineal

- Como no se conoce $\boldsymbol{\beta}$ se puede usar $\widehat{\boldsymbol{\beta}}$ para calcular los errores aproximados o *residuos*

$$\widehat{\varepsilon}_i = y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}}$$

  para $i = 1, 2, \ldots, n$

- Se usa la distribución empírica de $\widehat{\varepsilon}_i$

$$\widehat{F} \quad \rightarrow \quad \text{probabilidad igual a } 1/n \quad \text{en } \widehat{\varepsilon}_i$$

  para $i = 1, 2, \ldots, n$, de modo que $\widehat{F}$ tiene esperanza igual a 0.

Bootstrap en Regresión Lineal

- A partir de $\widehat{P} = (\widehat{\beta}, \widehat{F})$ se calculan los muestras bootstrap $\widehat{P} \to \mathbf{z}^*$

- Para generar $\mathbf{z}^*$ se toma primero una muestra aleatoria de términos de error
$$\widehat{F} \to (\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*) = \varepsilon^*$$

- Cada ε_i^* es igual a cualquiera de los n valores de $\widehat{\varepsilon}_j$ con probabilidad $1/n$

- Así, las respuestas boostrap se generan mediante
$$y_i^* = \mathbf{x}_i \widehat{\beta} + \varepsilon_i^*$$
para $i = 1, 2, \ldots, n$ donde $\widehat{\beta}$ es el mismo para todo i.

Bootstrap en Regresión Lineal

- En conjunto, las muestras bootstrap son $\mathbf{z}_i^* = (\mathbf{x}_i, y_i^*)$
- Se observa que los valores $\mathbf{x}_i$ (vector de covariables) son iguales tanto en los datos originales como en los datos bootstrap. Esto se debe a que $\mathbf{x}_i$ son valores *fijos* y no aleatorios.
- El estimador bootstrap $\widehat{\beta}^*$ es el valor que minimiza el error cuadrático residual

$$\sum_{i=1}^{n} \left(y_i^* - \mathbf{x}_i \widehat{\beta}^* \right)^2 = \min_{\mathbf{b}} \sum_{i=1}^{n} \left(y_i^* - \mathbf{x}_i \mathbf{b} \right)^2$$

- y con las ecuaciones normales aplicadas a los datos bootstrap se obtiene

$$\widehat{\beta}^* = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{y}^*$$

Bootstrap en Regresión Lineal

- El error estándar de los componentes de $\widehat{\boldsymbol{\beta}}^*$ se obtiene de manera directa

$$\begin{aligned}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}^*\right) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathrm{Var}(\mathbf{y}^*)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \widehat{\sigma}_F^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

- ya que $\mathrm{Var}(\mathbf{y}^*) = \widehat{\sigma}_F^2\mathbf{I}$ donde $\mathbf{I}$ es la matriz identidad.

- Así, el estimador bootstrap del error estándar es igual al usual en regresión lineal.

Bootstrap en regresión basado en pares de valores

- Hay otro método alternativo para aplicar el bootstrap en regresión, que es remuestreando las parejas de valores $\mathbf{z}_i = (\mathbf{x}_i, y_i)$

- De este modo, una muestra bootstrap consiste en

$$z^* = \{(\mathbf{x}_{i_1}, y_{i_1}), (\mathbf{x}_{i_2}, y_{i_2}), \ldots, (\mathbf{x}_{i_n}, y_{i_n})\}$$

  para $i_1, i_2, \ldots, i_n$, que es una muestra aleatoria de números enteros entre 1 y n.

- ¿Qué método es **mejor**, el que remuestrea residuos o el que remuestrea parejas?

Bootstrap en regresión basado en pares de valores

- La respuesta es que depende de cómo se considere el modelo de regresión.
- Si en el modelo se asume que el error correspondiente a la diferencia entre y_i y la media $\mu_i = x_i\beta$ no depende de $\mathbf{x}_i$, esto implica que tiene la misma distribución F sin importar cuál sea el valor de $\mathbf{x}_i$.
- El bootstrap con parejas es menos sensible a la suposición anterior y lo único que se requiere es que las parejas originales $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ se remuestreen de manera aleatoria de una distribución F en los vectores $p + 1$ dimensionales $(\mathbf{x}, y)$.

Aplicación del bootstrap a series temporales

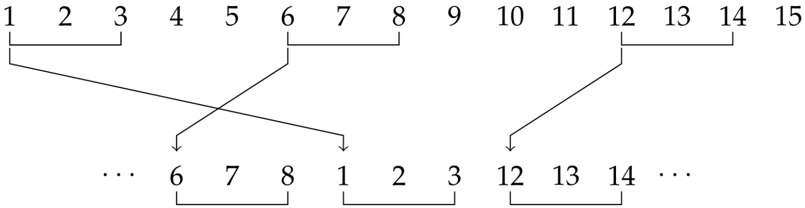
¿Cuáles son los métodos que se podrían aplicar en este caso?

- **Bootstrap de parejas de puntos**: Aquí **NO** se puede hacer porque se rompe la estructura de la serie temporal.
- **Bootstrap de residuos**: se preserva la estructura original de la serie cuando se asume la estrutura de dependencia entre los residuos.
- **Bootstrap de mediante bloques móviles (*moving blocks*)**: se preserva la estructura original de la serie.

Bloques móviles (*moving blocks*)

- En el esquema del bootstrap mediante análisis de residuos se asume que se ***sabe*** cuál es el proceso que genera los datos.
- Pero, en el esquema de bloques móviles se asume solo que un bloque de datos corto tiene un patrón de comportamiento semejante.
- Por ejemplo

```
N = 150
blockLen = 5
```


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
· · · 6 7 8 1 2 3 12 13 14 · · ·