

Estadística noparamétrica. Trabajo 1.

Máster en Técnicas Estadísticas. Curso 2009-2010

Regresión lineal simple

Supongamos que tenemos el modelo de regresión lineal simple (con diseño fijo)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

donde ε_i son variables aleatorias independientes con media cero y varianza σ^2 . Dado un conjunto de observaciones $(x_1, y_1), \dots, (x_n, y_n)$ los estimadores de mínimos cuadrados de β_0 y β_1 vienen dados por las fórmulas

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{SS_x}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

donde $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$. Recordemos que σ^2 se estima insesgadamente mediante

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2, \quad i = 1, \dots, n,$$

donde

$$e_i = y_i - \hat{\mu}_i, \quad i = 1, \dots, n$$

son los residuos y

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n,$$

es la estimación de $\mu_i = \beta_0 + \beta_1 x_i$. Es conocido que

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_x}.$$

Si los errores siguen una distribución normal entonces

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)},$$

sigue una distribución t de Student con $n - 2$ grados de libertad, siendo

$$\hat{\sigma}^2(\hat{\beta}_1) = \frac{s^2}{SS_x}.$$

Así el intervalo de confianza para β_1 de nivel $(1 - \alpha)$ vendría dado por

$$[\hat{\beta}_1 - c_u \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 - c_l \hat{\sigma}_{\hat{\beta}_1}],$$

donde

$$P(c_l = -t_{n-2, \alpha/2} \leq T_{n-2} \leq c_u = t_{n-2, \alpha/2}) = 1 - \alpha.$$

Si la hipótesis de normalidad no es cierta los valores de c_l y c_u pueden ser diferentes de los valores críticos de una t de Student. Por supuesto, por el teorema central del límite, esto no ocurrirá si la muestra es grande. Sin embargo, para muestras pequeñas y con errores claramente no normales el intervalo anterior puede no ser adecuado. En esta situación puede ser útil utilizar el bootstrap para aproximar la distribución de T . ¿Cómo podemos generar la muestra bootstrap? Una primera idea, si los errores ε_i siguen una distribución normal, sería seleccionar los errores bootstrap $\varepsilon_1^*, \dots, \varepsilon_n^*$ aleatoriamente según una $N(0, s^2)$ y generar y_i^* mediante la fórmula

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^*, \quad i = 1, \dots, n.$$

Esta idea se puede seguir usando en un contexto no paramétrico. Para ello necesitamos tener una buena aproximación de la distribución de los errores. Los valores de los residuos $\{e_1, \dots, e_n\}$ nos dan una idea de esa distribución. Sin embargo, su distribución no es del todo fiel a la de los errores originales ya que, por ejemplo, su varianza no es constante. Se tiene que

$$\text{Var}(e_i) = \sigma^2(1 - h_i),$$

donde,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x}, \quad i = 1, \dots, n.$$

Para corregir este problema se construyen los residuos modificados

$$r_i = \frac{e_i}{(1 - h_i)^{\frac{1}{2}}}, \quad i = 1, \dots, n.$$

Estos residuos ya tienen varianza constante σ^2 , como los errores ε_i . Sin embargo no tienen media cero. Por ello los errores bootstrap se escogen al azar del conjunto $\{r_1 - \bar{r}, \dots, r_n - \bar{r}\}$. El plan de remuestreo bootstrap para construir un intervalo de confianza para β_1 sería el siguiente:

1. Para $i = 1, \dots, n$

- a) Poner $x_i^* = x_i$.
- b) Seleccionar al azar ε_i^* del conjunto $\{r_1 - \bar{r}, \dots, r_n - \bar{r}\}$.
- c) Hacer $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i^*$

2. Estimar $\hat{\beta}_0^*, \hat{\beta}_1^*$ y los residuos e_1^*, \dots, e_n^* a partir de $(x_1^*, y_1^*), \dots, (x_n^*, y_n^*)$
3. Evaluar T^* en la muestra bootstrap. Para cada muestra bootstrap se obtiene

$$t^* = \frac{\hat{\beta}_1^* - \hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1^*)},$$

donde

$$\hat{\sigma}^2(\hat{\beta}_1^*) = \frac{s^{*2}}{SS_x}, \quad s^{*2} = \frac{1}{n-2} \sum_{i=1}^n e_i^{*2},$$

4. Repetir los pasos anteriores B veces obteniendo t_1^*, \dots, t_B^*
5. Ordenar de menor a mayor los valores calculados de T^* y tomar el valor que ocupa la posición $\alpha/2 * B$, c_l^* , y el que ocupa la posición $(1 - \alpha/2) * B$, c_u^* . El intervalo bootstrap para β_1 de nivel $(1 - \alpha)$ es

$$\left[\hat{\beta}_1 - c_u^* \hat{\sigma}(\hat{\beta}_1), \hat{\beta}_1 - c_l^* \hat{\sigma}(\hat{\beta}_1) \right]$$

Ejercicio: Comprueba el funcionamiento del método anterior. Para ello toma como valores de x 15 puntos equiespaciados en el intervalo $[0, 1]$, $x = 0, 1/15, 2/15, \dots$. Supongamos además que los errores de medida, ε_i , siguen una distribución t de Student con 3 grados de libertad.

1. Genera el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, 15,$$

con $\beta_0 = \beta_1 = 1$.

2. Calcula el intervalo bootstrap para β_1 . Toma $B = 500$ y $\alpha = 0,05$
3. Comprueba si β_1 está en el intervalo construido.
4. Repite los pasos anteriores $M = 500$ veces. Calcula el porcentaje de veces en que β_1 está en el intervalo bootstrap. Este porcentaje debería estar próximo al 95 %
5. Con las mismas M muestras generadas anteriormente, calcula el porcentaje de veces que β_1 está contenido en el intervalo construido suponiendo que los errores son normales. ¿Qué método da mejores aproximaciones del nivel de confianza?
6. Realiza la misma comparación cuando los errores siguen una distribución normal de media cero y $\sigma^2 = 3$.