Modelo lineal (1^o cuatrimestre 2019)

PRÁCTICA 4, primera parte

1. Vimos en clase que bajo los Supuestos S.1, si consideramos el modelo de proyección lineal

$$y = \mathbf{x}^T\boldsymbol{\beta} + u$$

con $\mathbf{x}, \boldsymbol{\beta} \in \mathbb{R}^p$, y tomamos $\boldsymbol{\beta}$ como el coeficiente óptimo lineal, entonces resulta que $\mathbb{E}(u\mathbf{x}) = 0$. Recíprocamente, supongamos que $y = \mathbf{x}^T\boldsymbol{\beta}+u$ con $\mathbb{E}(u\mathbf{x}) = 0$. Probar que entonces $\boldsymbol{\beta}$ es el coeficiente óptimo lineal

2. Asumamos que la variable aleatoria y y el vector aleatorio $(1, \mathbf{x}^T)^T$ cumplen los Supuestos S.1, donde $\mathbf{x} \in \mathbb{R}^{p-1}$, $\beta_0 \in \mathbb{R}$ es el intercept y $\mathbf{x}$ no contiene una constante. Escribamos el modelo de proyección lineal, separando la constante, del siguiente modo

$$y = \mathbf{x}'\boldsymbol{\beta} + \beta_0 + u, \tag{1}$$

donde $\boldsymbol{\beta} \in \mathbb{R}^{p-1}$ y $\beta_0 \in \mathbb{R}$ es el intercept.

   a) Tomando esperanza en (1), encuentre una relación entre $\mu_y = \mathbb{E}(y)$ y $\boldsymbol{\mu}_\mathbf{x} = \mathbb{E}(\mathbf{x})$. ¿Cuánto vale la $\mathbb{E}(u)$? ¿Y $\mathbb{E}(u\mathbf{x})$? Observe que se puede contestar las preguntas sobre la esperanza de u sin hacer cuentas.

   b) Escribamos la ecuación (1) para las variables o vectores centrados, es decir, la versión que relaciona $y-\mu_y$ con $\mathbf{x}-\mu_\mathbf{x}$, restando la μ_y en ambos miembros de (1). Calcule la $\text{cov}(\mathbf{x}-\mathbb{E}(\mathbf{x}), u)$.

   c) Observe que, como consecuencia del ejercicio 1, la ecuación obtenida en (b) es el modelo de proyección para las variables centradas, y por lo tanto se tiene que

$$\begin{aligned} \boldsymbol{\beta} &= \text{var}(\mathbf{x})^{-1}\,\text{cov}(\mathbf{x}, y) && (2)\\ \beta_0 &= \mu_y - \boldsymbol{\mu}_\mathbf{x}'\boldsymbol{\beta}. && (3) \end{aligned}$$

   En particular, este ítem muestra que hay dos maneras alternativas de calcular a los valores del coeficiente lineal óptimo de los predictores no constantes $\boldsymbol{\beta}$ en el modelo (1).

3. (Sesgo por variables omitidas) Partamos al vector de regresoras de la siguiente forma

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

donde $\mathbf{x}_1 \in \mathbb{R}^{k_1}$ y $\mathbf{x}_2 \in \mathbb{R}^{k_2}$, con $k_1 + k_2 = p$. Consideremos la proyección lineal de y en $\mathbf{x}_1$ solamente. Esto nos da la ecuación

$$\begin{aligned} y &= \mathbf{x}_1'\boldsymbol{\eta}_1 + e \\ \mathbb{E}(\mathbf{x}e) &= 0 \end{aligned} \tag{4}$$

Por otro lado tenemos el modelo de proyección lineal

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + u \\ y &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + u \\ \mathbb{E}(\mathbf{x}u) &= 0 \end{aligned} \tag{5}$$

Hemos notado a los coeficientes del vector $\mathbf{x}_1$ y a los errores en ambos modelos con letras distintas puesto que pueden diferir.

a) Escribir la expresión del coeficiente de proyección óptimo lineal para el modelo (4). Asumir que el modelo (5) es verdadero y expresar a $\boldsymbol{\eta}_{1,OL}$ óptimo lineal en función de $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$.

b) ¿En qué dos casos resulta $\boldsymbol{\eta}_{1,OL} = \boldsymbol{\beta}_1$? ¿Puede darles una interpretación a estos dos casos en términos del modelo? Cuando no son iguales, a la diferencia entre ambos se la denomina el *sesgo por variables omitidas*.

4. Sean $x_i \sim U(0,1)$, con $i \geq 1$ y $y_i = 1 - x_i^2 + \varepsilon_i$, con ε_i variables aleatorias iid con $\mathbb{E}(\varepsilon_1) = 0$ y $\text{var}(\varepsilon_1) = \sigma^2$, independientes de x_i. Supongamos que estimamos el modelo (equivocado) $y_i = \beta_0 + \beta_1 x_i + \eta_i$ por mínimos cuadrados, basándonos en una muestra de tamaño n.

   a) Hallar los valores numéricos de β_0^* y β_1^* a los que converge $\widehat{\beta}_{0,n}$ y $\widehat{\beta}_{1,n}$ en probabilidad cuando n tiende a infinito. Observar que estos valores se pueden calcular sin conocer la distribución de ε.

   b) Verifique los resultados obtenidos en el ítem anterior generando datos en R que sigan el modelo descripto más arriba y calculando el estimador de mínimos cuadrados. Cuando el tamaño de la muestra sea suficientemente grande, el estimador debería estar muy cerca de los resultados analíticos obtenidos en (a). Elija la distribución del error de su preferencia, mientras logre estar bajo los Supuestos S.1 asintóticos. Repita el gráfico visto en clase, que grafica una sucesión de estimadores de β_0 versus n y también una sucesión de β_1 versus n, y les superpone la recta horizontal con el valor verdadero en cada caso. Puede resultarle útil el script utilizado en clase, que está colgado en la página.

5. Escriba (y entregue el script) un programa en R que haga lo siguiente.

   a) Fije la semilla.

      i. Para $n = 10$ genere n datos y_i que sigan el modelo lineal

$$y_i = 4 + 2x_{i1} - 3x_{i2} + 0{,}5x_{i3} + \varepsilon_i, \quad 1 \leq i \leq n$$

      donde

      - $x_{1i} \sim U(-5,5)$, iid
      - $x_{2i} \sim U(-5,5)$, iid
      - $x_{3i} \sim U(-5,5)$, iid
      - $\epsilon_i \sim Exp(\lambda = 1/2) - 2$ independientes, e independientes de los anteriores (¿por qué le restamos 2 a los valores con distribución exponencial con $\lambda = 1/2$?)

- Genere también $x_{i4} \sim U(-5, 5)$, iid

ii. Ajuste el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$$

por mínimos cuadrados, basándose en la muestra de tamaño n.

iii. Guarde los parámetros estimados.

iv. Construya el intervalo de confianza de nivel 0.90 para el parámetro β_1 y para el parámetro β_4 asumiendo normalidad de los errores. ¿Contienen estos intervalos a los verdaderos parámetros para la muestra simulada? Guarde en un nuevo objeto un uno si lo contiene, y un cero sino, para cada uno de los dos intervalos.

v. Construya el intervalo de confianza de nivel asintótico 0.90 para el parámetro β_1 y para el parámetro β_4. ¿Contienen estos intervalos a los verdaderos parámetros para la muestra simulada? Guarde en un nuevo objeto un 1 si lo contiene, y un cero sino, para cada uno de los dos intervalos.

vi. Repita los items a)i) hasta a)v) $B = 1000$ veces, de modo de tener una muestra de tamaño B de los estimadores de cada β_j. ¿Diría que la distribución de los estimadores de β_2 puede aproximarse por la normal? Haga gráficos que le permitan tomar esta decisión. ¿Qué proporción de los B intervalos calculados para β_1 y β_4 basados en una muestra de n observaciones contuvo al verdadero valor del parámetro? Responda para cada tipo de intervalo calculado.

b) Repita (a) para $n = 25$ y $n = 100$.

c) Repita (a) y (b) para el caso de tener errores con distribución $Lognormal(\mu, \sigma^2) - exp(\mu - \sigma^2/2)$, tomando $\mu = 0$ y $\sigma^2 = 1$. Si para alguna de las distribuciones no consigue convencerse de que los $\widehat{\beta}$ tienen distribución que puede ser aproximada por una normal, repita, para errores generados con esa distribución el esquema de simulación anterior, pero con $n = 250, 500, 1000, 1500, 2000, 3000$. Exhiba los resultados en una tabla y comente brevemente sus conclusiones.

d) (para satisfacer la curiosidad, ya que tienen todo programado, sólo se trata de cambiar la distribución de los errores pero NO es para entregar) Repita c) pero ahora con la distribución de errores $U(-3, 3)$, $\chi^2_k - k$ con $k = 3$, y t_k con $k = 3$.