Estadística (M) - Segundo cuatrimestre 2020

Práctica 7 - Estimadores de mínimos cuadrados y regresión lineal

Se recomienda empezar por los ejercicios marcados sin ⋆.

A) *Propiedades del estimador de mínimos cuadrados en regresión lineal*

Supongamos que tenemos una muestra $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$ (fijos) y que existe un $\boldsymbol{\beta} \in \mathbb{R}^p$ tal que

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1}$$

donde ε_i son errores independientes con media 0 y varianza σ^2. Sea $\mathbf{X} \in \mathbb{R}^{n \times p}$ la matriz con filas $\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}}$ (en ese orden) y sea $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. Notar que si el modelo tiene intercept, $\mathbf{X}$ tendrá una primera columna compuesta por unos. En esta sección asumimos que la matriz de diseño $\mathbf{X}$ tiene rango completo. Llamamos $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ al estimador de mínimos cuadrados y $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = (\widehat{y}_1, \ldots, \widehat{y}_n)^{\mathrm{T}}$ el vector estimado de respuestas.

1. (a) Si el modelo tiene intercept, probar que $\sum_{i=1}^n \widehat{y}_i = \sum_{i=1}^n y_i$.
   (b) Probar que $\sum_{i=1}^n \widehat{y}_i (y_i - \widehat{y}_i) = 0$ (incluso si no hay intercept).

2. (a) En el modelo (1), los errores son homocedásticos, es decir, su varianza es siempre la misma. Estos errores pueden ser estimados por los residuos $r_i = y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. ¿Es cierto que estos residuos también son homocedásticos?
   (b) Si el modelo tiene intercept, calcular la suma $\sum_{i=1}^n r_i$. (Sugerencia: usar el ejercicio 1).

3. Supongamos que con la matriz de diseño $\mathbf{X}$ y vector de respuestas $\mathbf{y}$ obtenemos, por el método de mínimos cuadrados, el vector estimado de respuestas $\widehat{\mathbf{y}}$. Supongamos ahora que, en vez de observar $\mathbf{X}$, hubiéramos observado la matriz de diseño $\mathbf{W}$, donde la columna j de $\mathbf{W}$ es igual a la columna j de la matriz $\mathbf{X}$ multiplicada por una constante $c_j \neq 0$. Con la matriz $\mathbf{W}$ y el vector de respuestas $\mathbf{y}$, se obtiene el vector estimado de respuestas $\widehat{\mathbf{y}}_{\mathbf{W}}$. Probar que $\widehat{\mathbf{y}} = \widehat{\mathbf{y}}_{\mathbf{W}}$. Deducir que el vector estimado de respuestas es invariante respecto a la escala con la que se miden las variables explicativas.

⋆ 4. **(Coeficiente de correlación múltiple)**

   (a) Probar que si en el modelo (1) $\mathbf{X}$ es un vector columna de unos, entonces $\widehat{y}_i = \overline{y}$ (el promedio de las y_i).

   (b) Supongamos que el modelo (1) tiene ordenada al origen. Entonces,

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2.$$

   Esto también se puede escribir como SCT = SCE + SCR donde

   - SCT $= \sum_{i=1}^n (y_i - \overline{y})^2$ es la suma de cuadrados totales, y refleja la variabilidad total de la respuesta $\mathbf{Y}$.
   - SCE $= \sum_{i=1}^n (y_i - \widehat{y}_i)^2$ es la suma de cuadrados de los errores y refleja la variabilidad de $\mathbf{y}$ que no fue explicada por el modelo de regresión.
   - SCR $= \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2$ es la suma de cuadrados de la regresión y refleja la variabilidad de $\mathbf{y}$ explicada por el modelo de regresión (en comparación con lo que obtendríamos si sólamente consideráramos el intercept).

El **coeficiente de correlación múltiple** R se define como la raíz cuadrada de

$$R^2 = \frac{\text{SCR}}{\text{SCT}}.$$

Sugerencia: usar el ejercicio 1.

(c) Probar que si hay ordenada al origen,

i. $0 \leq R^2 \leq 1$.

ii. En que caso $R^2 = 0$?

iii. Cuando $R^2 = 1$?

B) *Inferencia para el vector de regresión e interpretación de salidas del* `R`

1. En el puerto de la Ciudad de Grand Lakes, en Canadá, se quiere ver cómo influye el peso de un cargamento en el tiempo necesario para descargarlo. Para eso, se registra el peso y el tiempo de descarga para 30 cargamentos y se plantea el modelo lineal

$$\text{Tiempo} = \alpha + \text{Peso} * \beta + \varepsilon_i \tag{2}$$

donde se asume que los errores ε_i son independientes y tienen distribución normal con media 0. Los datos se encuentran en el archivo `glakes.csv`. A partir de la salida de la función `lm` de `R`, contestar las siguientes preguntas:

(a) ¿Cuáles son los coeficientes estimados para α y β usando mínimos cuadrados?

(b) ¿Hay evidencia suficiente a nivel 0.01 para decir que $\beta \neq 0$? ¿Cuál es el estadístico del test correspondiente y cuál es su distribución bajo H_0? ¿Cuánto vale este estadístico para estos datos? Hallar el $p-$valor. Interpretar la conclusión de este test.

(c) Considerar las hipótesis $H_0 : \alpha = 10$ vs. $H_1 : \alpha > 10$. ¿Hay evidencia suficiente para rechazar H_0 a nivel 0.05? Hallar el $p-$valor.

(d) Hallar un intervalo de confianza para α de nivel 0.95.

* (e) ¿Cuánto vale el R^2 en este caso y cómo interpretaría esto?

(f) Llega un nuevo cargamento con peso 1000. ¿Cómo estimaría el tiempo que se tarda en descargarlo?

2. En este conjunto de datos, queremos explicar la deuda (variable `Balance`) en tarjeta de crédito de 400 clientes en función de varias características de cada uno. Para cada cliente, tenemos las variables

- `Income` (sueldo anual),
- `Rating` (rating crediticio),
- `Limit` (lmite de crédito),
- `Cards` (número de tarjetas),
- `Age` (Edad),
- `Education` (número de años de educación),
- `GenderFemale` (vale 1 si el cliente es mujer, 0 si no),
- `StudentYes` (vale 1 si el cliente es estudiante, 0 si no),

- `MarriedYes` (vale 1 si el cliente está casado, 0 si no).

Se asume válido un modelo lineal con ordenada al origen, variable respuesta `Balance` y las variables explicativas ya descriptas. Los datos se encuentran en el archivo `credit.txt`. Supongamos que los errores ε_i del modelo tienen distribución $N(0, \sigma^2)$.

(a) Utilizando la función `ggpairs` indique qué variables les parece que pueden ser las más informativas en este modelo.

(b) Haga una estimación de los parámetros utilizando cuadrados mínimos. Indique cuál es el estimador de σ^2.

(c) Sea β_{Age} el coeficiente correspondiente a la variable `Age`. Hallar el $p-$valor del test para $H_0 : \beta_{Age} = 0$ vs. $H_1 : \beta_{Age} \neq 0$. ¿Cuál sería el $p - valor$ del test si nos interesara estudiar $H_0 : \beta_{Age} = 0$ vs. $H_1 : \beta_{Age} < 0$.

(d) Supongamos que dos clientes Juan y Pedro comparten las mismas características (en términos de las variables explicativas consideradas), salvo que Juan tiene tres años más que el otro. Sea B_J el valor esperado de la deuda de Juan y B_P el valor esperado de la deuda de Pedro. Hallar $B_P - B_J$ y estimarlo en este caso.

(e) Hallar un intervalo de confianza de nivel 0.9 para $\beta_{Education}$ (el coeficiente correspondiente a la variable `Education`).

3. (Para hacer con el R) Para el conjunto de datos `aircraft` del paquete `robustbase` en R, se tienen las siguientes variables:

- Y: Costo (en unidades de \$100,000)
- X_1: Relación de aspecto
- X_2: Relación elevación-arrastre
- X_3: Peso del avión (en pounds)
- X_4: Impulso máximo

(a) Realizar un boxplot de las variables X_2 e Y. Qué observa?

(b) Realizar un diagrama de dispersión de Y vs X_2

(c) Se propone el siguiente modelo:

$$Y = \beta_0 + \beta_1 \, X_2 + \varepsilon_i \qquad \mathbb{E}(\varepsilon_i) = 0 \qquad \text{Var}(\varepsilon_i) = \sigma^2$$

Realizar un ajuste por cuadrados mínimos y graficar la recta de regresión en el mismo gráfico que el ítem (b). Hallar el $p-$valor del test para $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

(d) Eliminar la observación número 22, realizar nuevamente el ajuste por cuadrados mínimos y graficar la recta predicha sobre el mismo gráfico, utilizando otro color.

(e) Explorar la función `lmrob` del paquete `robustbase` en R y utilizarla para realizar un ajuste robusto para este mismo modelo. Graficar la recta ajustada, utilizando un tercer color. Qué observa? Qué puntos identifica como outliers el método robusto?

(f) Realizar un último ajuste de cuadrados mínimos con la submuestra obtenida quitando los puntos identificados como outliers en el ítem anterior y graficarla junto con los ajustes anteriores. Hallar el $p-$valor del test para $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Concluir acerca de los resultados observados en el punto (c).