Regresión múltiple y otras técnicas multivariadas

Tarea 07

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Barrios Georgina

Abril 03, 2019

Ejercicio 1

En el análisis del modelo de RLM, calcular ECM($\hat{\sigma}^2_{MCO}$) y ECM($\hat{\sigma}^2_{MV}$). A partir de estos resultados decidir qué estimador es mejor.

Demostración. Recordemos que si una variable aleatoria se distribuye ji-cuadrada con n grados de libertad $X \sim \chi^2_{(n)}$, entonces cumple que $\mathrm{E}(X) = n$ y $\mathrm{V}(X) = 2n$.

Sabemos que $\dfrac{(n-p-1)\hat{\sigma}^2_{MCO}}{\sigma^2} \sim \chi^2_{(n-p-1)}$, por lo tanto

$$(n-p-1) = \mathrm{E}\left[\frac{(n-p-1)\hat{\sigma}^2_{MCO}}{\sigma^2}\right] = \frac{(n-p-1)}{\sigma^2}\mathrm{E}\left(\hat{\sigma}^2_{MCO}\right) \quad \Rightarrow \quad \mathrm{E}\left(\hat{\sigma}^2_{MCO}\right) = \sigma^2$$

$$2(n-p-1) = \mathrm{V}\left[\frac{(n-p-1)\hat{\sigma}^2_{MCO}}{\sigma^2}\right] = \frac{(n-p-1)^2}{\sigma^4}\mathrm{V}\left(\hat{\sigma}^2_{MCO}\right) \quad \Rightarrow \quad \mathrm{V}\left(\hat{\sigma}^2_{MCO}\right) = \frac{2\sigma^4}{(n-p-1)}$$

Con lo anterior, se procede a calcular el error cuadrático medio (ECM) para $\hat{\sigma}^2_{MCO}$.

$$\mathrm{ECM}\left(\hat{\sigma}^2_{MCO}\right) = \mathrm{B}^2\left(\hat{\sigma}^2_{MCO}\right) + \mathrm{V}\left(\hat{\sigma}^2_{MCO}\right) = 0 + \frac{2\sigma^4}{(n-p-1)} = \frac{2}{(n-p-1)}\sigma^4$$

De forma análoga se procede con $\hat{\sigma}^2$. Sabemos que $\dfrac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$, por lo tanto

$$(n-p-1) = \mathrm{E}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = \frac{n}{\sigma^2}\mathrm{E}\left(\hat{\sigma}^2\right) \quad \Rightarrow \quad \mathrm{E}\left(\hat{\sigma}^2\right) = \frac{(n-p-1)}{n}\sigma^2$$

$$2(n-p-1) = \mathrm{V}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = \frac{n^2}{\sigma^4}\mathrm{V}\left(\hat{\sigma}^2\right) \quad \Rightarrow \quad \mathrm{V}\left(\hat{\sigma}^2\right) = \frac{2(n-p-1)}{n^2}\sigma^4$$

Con lo anterior, se procede a calcular el error cuadrático medio (ECM) para $\hat{\sigma}^2$.

$$\begin{aligned}\text{ECM}\left(\hat{\sigma}^2\right) &= \text{B}^2\left(\hat{\sigma}^2\right) + \text{V}\left(\hat{\sigma}^2\right) = \left(\frac{(n-p-1)}{n}\sigma^2 - \sigma^2\right)^2 + \frac{2(n-p-1)}{n^2}\sigma^4 \\ &= \frac{(p+1)^2}{n^2}\sigma^4 + \frac{2(n-p-1)}{n^2}\sigma^4 \\ &= \frac{p^2+2n-1}{n^2}\sigma^4\end{aligned}$$

□

Tenemos así que $\text{ECM}\left(\hat{\sigma}^2_{MCO}\right) = \dfrac{2}{(n-p-1)}\sigma^4$ y $\text{ECM}\left(\hat{\sigma}^2\right) = \dfrac{p^2+2n-1}{n^2}\sigma^4$.

Notemos que $n^2 > (n-p-1)$ implica que $\dfrac{p^2+2n-1}{n^2} < \dfrac{p^2+2n-1}{n-p-1}$.

Ejercicio 2

En el análisis del modelo RLM, la suma de cuadrados de regresión se define como

$$SCR = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$$

donde $\hat{y}_i$ es la i-ésima componente del vector $\hat{\boldsymbol{y}}$ y $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$. Mostrar que

$$SCR = \boldsymbol{y}^T\left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n\right)\boldsymbol{y}$$

Ejercicio 3

En el análisis del modelo RLM, mostrar que $\left(\boldsymbol{I}_n - \boldsymbol{H}\right)\left(\boldsymbol{H} - \dfrac{1}{n}\boldsymbol{J}_n\right) = \boldsymbol{0}_{n\times n}$

Ejercicio 4

El conjunto de datos `house_selling_prices_OR.csv` contiene información sobre el precio de venta y características de una muestra de 200 observaciones. El objetivo es ajustar un modelo RLM paraexplicar la distribución del precio de venta en miles de dólares (`house_price`) como función del tamaño de la vivienda en metros cuadrados (`house_size`) y del número de habitaciones (`bedrooms`).

Con los resultados obtenidos, responder lo siguiente.

Inciso 4.a)

Reportar e interpretar las estimaciones de los coeficientes del modelo.

Inciso 4.b)

Calcular intervalos de confianza simultáneos 95% para los coeficientes del modelo e interpretar los resultados.

Inciso 4.c)

¿Tiene algún efecto el tamaño de la vivienda en el precio de venta?

Inciso 4.d)

¿Tiene algún efecto el número de habitaciones en el precio de venta de la vivienda?

Inciso 4.e)

Reportar la estimación de σ^2 y calcular un intervalo de confianza 95%.

Inciso 4.f)

Estimar puntualmente y por intervalo la media del precio de venta de las viviendas de 250 metros cuadrados y tres habitaciones.

Ejercicio 5

El conjunto de datos `fl_crime.csv` contiene información sobre los 67 del estado de Florida, EUA. Para este ejercicio se debe ajustar un modelo RLM para explicar la distribución de la tasa de crímenes por cada 1000 habitantes (`crime_rate`) como función del porcentaje de adultos con educación superior (`edu`) y del grado de urbanización (`urban`). Con los resultados obtenidos, responder lo siguiente.

Inciso 5.a)

Reportar e interpretar las estimaciones de los coeficientes del modelo

Inciso 5.b)

Calcular intervalos de confianza simultáneos 95% para los coeficientes del modelo e interpretar los resultados

Inciso 5.c)

¿Tiene algún efecto la educación en la tasa de crímenes de los condados de Florida?

Inciso 5.d)

¿Tiene algún efecto la urbanización en la tasa de crímenes de los condados de Florida?

Inciso 5.e)

Calcular la matriz de correlaciones de las tres variables involucradas en el modelo y reportar los resultados. Tratar de explicar los resultados de los incisos b) y c) a partir de estas correlaciones.

Inciso 5.f)

Reportar la estimación de σ^2 y calcular un intervalo de confianza 95%.

Inciso 5.g)

Estimar puntualmente y por intervalo la media de la tasa de crímenes para un 65% de adultos con educación superior y un grado de urbanización de 50%.