Regresión Lineal Múltiple

Regresión lineal múltiple

Es una generalización del modelo de regresión lineal simple, para el caso de varias variables explicativas:

$$y = \beta X + e = y_i = \beta_1 + \beta_2 \times x_{2i} + \beta_3 \times x_{3i} + ... + \beta_k \times x_{ki} + e_i,\ i=1,2,\ldots,n$$

Por lo que vamos a tener una expresión matricial:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{21} & X_{31} & \ldots & X_{k1} \\ 1 & X_{22} & X_{32} & \ldots & X_{k2} \\ . & . & . & \ldots & . \\ 1 & X_{2n} & X_{3n} & \ldots & X_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ e_n \end{bmatrix}$$

Dónde:

y es el vector de las observaciones de la variable endógena,

X es la matriz de observaciones de las variables exógenas,

β es vector de coeficientes; y

e es el vector de los errores.

Estimación con R

Para encontrar modelos de regresión lineal, tanto simple como múltiple, vamos a utilizar la función de la instalación base de R: *lm()* para ajustar modelos lineales.

Los parámetros que vamos a utilizar son:

formula: vamos a indicar que variable va a ser la dependiente y cuáles las explicativas, con el operador **~**.

data: el data.frame del que se especifican las variables (opcional).

Para utilizar el modelo con datos nuevos y predecir el valor de la variable de respuesta, vamos a utilizar la función *predict()*.

Los parámetros que vamos a utilizar son:

object: objeto tipo modelo “lm”.

newdata: el data.frame con los datos nuevos (opcional).

Ejercicios

Ejercicio 1 – Cargue el archivo “supermercados.csv” utilizado en la guía anterior, que contiene información de ventas y los gastos de publicidad que se realizaron un producto.

a. Analice la correlación entre las variables exógenas y la variable objetivo. Haga lo mismo con las correlaciones parciales. ¿A que conclusión llega?

b. Encontrar un modelo de regresión lineal, pero en este caso utilizando todas las variables explicativas, para predecir las ventas según el tipo de publicidad utilizada. ¿Qué variables explicativas son las significativas para el modelo?

c. Compare el modelo obtenido con los tres modelos que obtuvo en la guía anterior. ¿Es más adecuado para predecir las ventas?

d. Prediga las ventas para los mismos valores de publicidad de la guía anterior.

Ejercicio – 2. En un estudio ambiental sobre la diversidad de especies de tortuga en las islas Galápagos, se recogió información sobre el número de especies endémicas (Endemics), así como el área de la isla (área), la altura del pico más alto de la isla (Elevation), la distancia a la isla más cercana (Nearest), la distancia a la isla de Santa Cruz (Scruz), y el área de la isla más próxima (Adjacent). Estos datos se encuentran en el archivo “tortugas.csv”.

a. Explore los datos.

b. Encuentre la matriz de correlación entre las variables y verifique el resultado utilizando la correlación parcial entre las variables. ¿Qué relación encuentra entre las variables? Grafique.

c. Encuentre un modelo de regresión para predecir el número de especies endémicas de tortuga.

d. ¿Cuáles son las variables más significativas que explican el modelo? Compare el resultado con los datos de la matriz de correlación obtenida en el punto b.

Ejercicio – 3. La producción de cereales está determinada, principalmente por las condiciones climáticas previas a la cosecha.

En el archivo “produccion.csv” se han registrado las siguientes variables para distintos años:

- año: año de la medición.
- premay: precipitación de mayo.
- tempmay: temperatura de mayo.
- prejun: precipitación de junio.

- tempjun: temperatura de junio.
- prejul: precipitación de julio.
- tempjul: temperatura de julio.
- preago: precipitación de agosto.
- tempago: temperatura de agosto.
- produccion: producción de cereales.

a. Explore los datos y las relaciones entre las variables (matriz de correlación y correlación parcial).

b. Encuentre un modelo de regresión utilizando los datos de cada mes (temperatura y precipitaciones) para predecir la producción. ¿Con los datos de qué mes se genera el modelo más adecuado?

c. Encuentre un modelo de regresión lineal que explique la producción de cereales utilizando los datos de todos los meses registrados. ¿El coeficiente de determinación del modelo con todos los datos mejora respecto a los obtenidos con los meses individuales?

d. ¿Cuáles son las variables más significativas en la producción de cereales?

Ejercicio – 4. Cargue el conjunto de datos “state” de la instalación base de R y utilice state.x77.

a) Explore los datos. ¿Qué relación encuentra entre las variables?

b) Encuentre un modelo para predecir la expectativa de vida.

c) ¿Es significativo el modelo? ¿Cuáles son las variables más significativas que lo explican?

d) Encuentre un nuevo modelo con las variables más significativas encontradas en el punto anterior. ¿El nuevo modelo explica mejor la variable objetivo?

Referencias

1. “Estadística y Machine Learning con R”. Disponible en: https://bookdown.org/content/2274/portada.html

2. “An Introduction to corrplot Package”. Disponible en: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

3. “ppcor: Partial and Semi-Partial (Part) Correlation”. Disponible en: https://cran.r-project.org/web/packages/ppcor/ppcor.pdf