FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Taller R - 10: Modelo de regresión lineal

Ricardo Aravena C. - Cristian Capetillo C. - Ingrid Guevara R.
Ricardo Olea O. - Bladimir Morales T., Daniel Saavedra M.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2024

Contenidos

Motivación

- **Simplicidad y claridad interpretativa:** Los modelos de regresión lineal tienen una formulación matemática simple ($y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) que permite interpretar fácilmente los coeficientes (β) como el efecto promedio de las variables independientes (x_i) sobre la dependiente (y_i).
- **Base para modelos más complejos:** Es un modelo fundamental en estadística, que sirve como punto de partida para desarrollar otros más avanzados, como la regresión múltiple, la regresión logística o los modelos lineales generalizados (GLMs).
- **Eficiencia computacional:** La implementación computacional de estos modelos es directa y eficiente, lo que los hace adecuados incluso para grandes conjuntos de datos.

- **Aplicaciones prácticas diversas:** Se usa ampliamente en economía, biología, psicología, ingeniería, y otras áreas para estimar relaciones lineales y realizar predicciones, Podemos por ejemplo tener los siguientes temas de interés:
  - Interesa estudiar el efecto del monto utilizado en publicidad sobre el volumen de las ventas.
  - Interesa predecir el desempeño laboral de los ingenieros contratados en base a un test de aptitudes aplicado antes de sucontratación.
  - Interesa estudiar la relación entre el nivel educacional de los padres y el tamaño del vocabulario de un niño, controlando por su edad.
  - Interesa estudiar la demanda de un producto en función del precio.
  - Interesa estudiar la relación entre temperatura y presión en un sistema físico.

Definición

Sea $Y_1, \ldots, Y_n$ variables aleatorias independientes, que está definida en función de X_i, $i = 1, \ldots, n$, así el modelo de regresión lineal simple puede definirse como:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\overset{i.i.d.}{\sim} Normal(0, \sigma^2), \end{aligned} \tag{1}$$

o alternativamente

$$Y_i \sim Normal(\beta_0 + \beta_1 x_i, \sigma^2) \tag{2}$$

La función `lm()` sirve para crear modelos de regresión lineal. Para efectos de este laboratorio se utilizará para estimar el intercepto y la pendiente de una recta que ajuste por mínimos cuadrados los valores de y respecto a x, es decir:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La función se utiliza de la siguiente manera: `lm(y ~ x)`.

Aplicación

La base de datos del archivo `casas_macul_nunoa.xlsx` es el resultado de realizar web scraping el día 3 de mayo de 2020 a una página que contiene avisos de ventas de viviendas de la Región Metropolinada de Chile. Los datos se refieren principalmente al valor de las casas usadas en la Región Metropolitana publicadas en el sitio web `https://chilepropiedades.cl`, correspondiente a *Chile Propiedades*. En este problema se desea estudiar el valor de las casas usadas en Unidades de Fomento de las comunas de Macul y Ñuñoa. Se quiere determinar si la variable de la superficie construida en m^2 es determinante para explicar el valor de la casa.

Se pide realizar el siguiente análisis:

- Realizar un diagrama de dispersión de ambas variables.
- Obtenga el coeficiente de correlación lineal simple y comente sobre su signo y magnitud.
- Ajuste un modelo de regresión a las observaciones y agregue la recta ajustada a la figura.
- Interprete los coeficientes estimados.

Significancia de la regresión

Hipótesis de interés: Para establecer si el valor de $\hat{\beta}_1$ obtenido es evidencia de que $\beta_1 \neq 0$ y, por tanto, existe asociación entre las variables, interesa testear la hipótesis:

$$H_0 : \beta_1 = 0 \qquad vs. \qquad H_1 : \beta_1 \neq 0.$$

Se puede demostrar que

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2},$$

Así la regla de rechazo será:

$$t_0 = \frac{|\hat{\beta}_1|}{\hat{\sigma}/\sqrt{S_{xx}}} \geq t_{1-\alpha/2,n-2},$$

o si el $p-$valor es menor a α, con significancia igual a α.

Supuestos

Si el modelo es cierto, los errores cumplen con

$$\varepsilon_i \overset{i.i.d.}{\sim} Normal(0, \sigma^2).$$

Tomando algunas precauciones, utilizaremos los residuos como representantes de estos errores, que están definidos como

$$e_i = Y_i - \hat{Y}_i.$$

Una propiedad interesante es que la suma de los residuos es igual a cero.

Los residuos estandarizados se definen como:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{3}$$

con $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$, $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.
Los residuos studentizados corresponen a:

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \tag{4}$$

donder $\hat{\sigma}_{(i)}$ es el estimador de σ al ajustar el modelo sin la $i-$ésima observación, así

$$\hat{\sigma}_{(i)} = \frac{\hat{\sigma}}{n-3}(n - 2 - r_i^2)$$

Supuesto1: la media es lineal en el predictor. Si este supuesto es válido, el valor esperado de los residuos debe ser cero, por lo tanto, en un gráfico e_i versus x_i no deben observarse patrones. Los puntos deben ubicarse de manera homogénea en el gráfico. Lo mismo debe ocurrir en un gráfico e_i versus $\hat{y}_i$. También puede revisarse con los residuos estandarizados, r_i.

Supuesto2: todas las observaciones tienen igual varianza (homocedasticidad). Si este supuesto es válido, los residuos estandarizados deben tener igual varianza (e igual a 1). Un gráfico r_i versus x_i debe mostrar variabilidades verticales similares a lo largo del eje de las abscisas. Lo mismo debe ocurrir en un gráfico r_i versus $\hat{y}_i$. Lo mismo aplica para los residuos studentizados, t_i.

Supuesto 3: las observaciones siguen una distribución Normal. Si este supuesto es válido, los residuos estandarizados siguen aproximadamente una distribución Normal estándar. Un qq-plot de los residuos estandarizados, r_i, debe mostrar un buen ajuste.

Supuesto 4: todas las observaciones provienen del mismo modelo (identicamente distribuidos). Si las observaciones provienen del modelo propuesto, sus residuos estandarizados (studentizados) deben ser coherentes con una distribución Normal estándar (t_{n-3}). Valores fuera el área central de estas distribuciones son indicación de observaciones anómalas u outliers.

Supuesto 5: las observaciones son independientes. El test de Durbin-Watson está diseñado para detectar correlación temporal en los errores. Interesa testear

$$\begin{aligned} H_0 &: \rho = 0 \text{ (No hay correlación)} \\ H_1 &: \rho \neq 0 \text{ (Hay correlación)} \end{aligned}$$

En `R` es posible realizar este test a través de la función `dwtest` de la librería `lmtest`.

Predicción

Suponga que interesa el valor de la media de la variable respuesta cuando el predictor toma un valor dado x_0:

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

Un estimador natural para μ_0 está dado por:

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Así el intervalo de confianza para μ_0 es:

$$\left[\hat{\mu}_0 - t_{(1-\alpha/2,n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}; \hat{\mu}_0 + t_{(1-\alpha/2,n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\right]$$

Aplicación

Para la aplicación del valor de las casas de Macul y Ñuñoa, se pide

- Verificar si el coeficiente estimado es significativo para el modelo.
- Verificar los cinco supuestos del modelo.
- Realizar una predicción para un casa que tenga una superficie construida de 200 metros cuadrados

Aplicación

Se pide realizar un modelo de regresión lineal múltiple con dos covariables en el modelo es decir:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

donde: $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, x_{i1} es el coeficiente asociado a la superficie construida en metros cuadrados y x_{i2} a la comuna a la que pertenece. Ajuste el modelo e interprete los resultados verificando su significancia.