Estadística II - 3006915
Regresión Lineal Simple

Mateo Ochoa Medina

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín

Periodo académico 2023-2S

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Contenido

Contenido

Inferencias respecto a la respuesta media $\mu_{Y\,|\,x_0}$ y valores futuros

Debido que los valores ajustados de la variable respuesta son combinaciones lineales de las variables aleatorias $Y_1, Y_2, \ldots, Y_n$, bajo los supuestos de normalidad e independencia entre los errores, podemos afirmar que las variables $\hat{Y}_i$, $i = 1, 2, \ldots, n$, son variables aleatorias normales, con media $\mu_{Y\,|\,x_i} = E\,(Y\,|\,X = x_i) = \beta_0 + \beta_1\,x_i$ y varianza $\hat{\sigma}^2\,\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$, aunque no son independientes. Se sabe además que $\hat{Y}_i$ estima a $\mu_{Y\,|\,x_i} = E\,(Y\,|\,X = x_i)$. Podemos hacer inferencias sobre esta media, así como predecir un valor futuro y_0 de la respuesta en un valor fijo de $X = x_0$.

Inferencias sobre la respuesta media

Bajo los supuestos del modelo se satisface que

$$T = \frac{\hat{Y}_0 - \mu_{Y \mid x_0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}, \tag{1}$$

con $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Por tanto, un intervalo de confianza del $(1 - \alpha)$ % para $\mu_{Y \mid x_0}$ es:

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

Nota:

A partir de la construcción y el análisis del intervalo de confianza definido en (2) se pueden probar hipótesis para la respuesta media de la forma $H_0 : \mu_{Y \mid x_0} = E(Y \mid X = x_0) = c$ vs. $H_1 : \mu_{Y \mid x_0} = E(Y \mid X = x_0) \neq c$, donde $c \in \mathbb{R}$. Por tanto, si el valor c está incluido en el intervalo, entonces no se rechaza H_0, por el contrario, si c no está incluido en el intervalo, entonces se rechaza H_0.

Inferencias sobre la respuesta futura

Bajo los supuestos del modelo se satisface que

$$T = \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}, \tag{3}$$

con $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Por tanto, un intervalo de predicción del $(1 - \alpha)$ % para y_0 está dado por:

$$\hat{y}_0 \pm t_{\alpha/2, n-2} \times \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}. \tag{4}$$

Consideraciones sobre las inferencias para la respuesta media y la respuesta futura

Notas sobre $\mu_{Y\,|\,x_0}$ y/o y_0:

- Mientras que el intervalo de confianza para $\mu_{Y\,|\,x_0}$ proporciona un rango en el cual pudiera estar la media de la respuesta para $X = x_0$, con el nivel de confianza dado, el intervalo de predicción en un valor $X = x_0$, estima, con el nivel de confianza dado, el rango de los posibles valores en el cual podría ser observado el valor de la variable respuesta.
- Asumimos que en el valor particular x_0 obtenemos un valor futuro de la variable aleatoria Y_0, y por tanto, éste no ha sido utilizado en el ajuste del modelo de regresión. La predicción de Y_0 con base en el modelo ajustado con la muestra de pares (x_i, y_i), $i = 1, 2, \ldots, n$, corresponde a $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, y dado que Y_0 no hizo parte de la muestra de ajuste, las variables aleatorias Y_0 y $\hat{Y}_0$ son estadísticamente independientes.

Consideraciones sobre las inferencias para la respuesta media y la respuesta futura

Notas sobre $\mu_{Y \mid x_0}$ y/o y_0:

- La independencia entre Y_0 y $\hat{Y}_0$ desprende que el error del pronóstico: $e_0 = Y_0 - \hat{Y}_0$, es una variable aleatoria normal cuya media es cero y la varianza es igual a

$$V(e_0) = V(Y_0) + V\left(\hat{Y}_0\right) = \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma^2.$$

- En general, no se recomienda realizar extrapolaciones por fuera del rango de variación observado en el conjunto de datos sobre la variable regresora (es decir, por fuera del rango experimental donde el modelo fue ajustado). Por ello es importante que en los datos muestrales se haya cubierto el rango en el que naturalmente pudiera tomar valores la variable explicatoria. De lo contrario, es posible que en un rango fuera del observado, la relación estadística formulada no resulte válida. En conclusión, solo se podrán hacer inferencias sobre la respuesta cuando $X = x_0 \in [X_{\text{mín}}, X_{\text{máx}}]$, donde $X_{\text{mín}}$ y $X_{\text{máx}}$ son los valores mínimo y máximo de la variable regresora, respectivamente, que fueron fijados en la muestra.

Ilustración de extrapolación

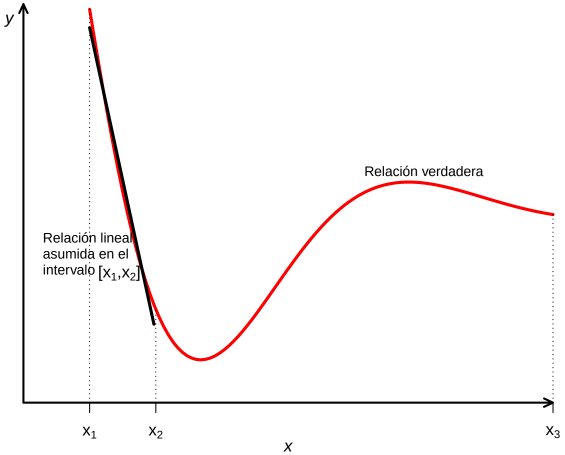

Figura 1: Se consideran las observaciones para el intervalo $[x_1, x_2]$, la recta ajustada en ese tramo causaría un gran error, para todas las extrapolaciones en el intervalo $x \in (x_2, x_3]$.

Contenido

Referencias

Montgomery, D. C., Peck, E. A., y Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley, New Jersey, quinta edición.

Álvarez, N. G. (2022). Notas de Clase Análisis de Regresión - 3006918, Capítulo 2: Regresión Lineal Simple. Notas no publicadas.

Álvarez, N. G. y Gómez, C. M. L. (2018). Notas de Clase - Estadística II (3006918): Análisis de Regresión Lineal e Introducción al Muestreo.