Análisis de Regresión Lineal

Universidad Autónoma de la Ciudad de México

Casa Libertad

Carlos E. Martínez Rodríguez

carlos.martinez@uacm.edu.mx
Academia de Matemáticas - Modelación Matemática
Colegio de Ciencia y Tecnología

Semestre 2019-II

Descripción

Nota

- *En muchos problemas hay dos o más variables relacionadas, para medir el grado de relación se utiliza el* ***análisis de regresión****.*
- *Supongamos que se tiene una única variable dependiente, y, y varias variables independientes, $x_1, x_2, \ldots, x_n$.*
- *La variable y es una varaible aleatoria, y las variables independientes pueden ser distribuidas independiente o conjuntamente.*

RLS

- A la relación entre estas variables se le denomina modelo regresión de y en $x_1, x_2, \ldots, x_n$, por ejemplo $y = \phi(x_1, x_2, \ldots, x_n)$, lo que se busca es una función que mejor aproxime a $\phi(\cdot)$.

Supongamos que de momento solamente se tienen una variable independiente x, para la variable de respuesta y. Y supongamos que la relación que hay entre x y y es una línea recta, y que para cada observación de x, y es una variable aleatoria.
El valor esperado de y para cada valor de x es

$$E(y|x) = \beta_0 + \beta_1 x \tag{1}$$

β_0 es la ordenada al orígen y β_1 la pendiente de la recta en cuestión, ambas constantes desconocidas

Mínimos Cuadrados

Supongamos que cada observación y se puede describir por el modelo

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (2)$$

donde ϵ es un error aleatorio con media cero y varianza σ^2. Para cada valor y_i se tiene ϵ_i variables aleatorias no correlacionadas, cuando se incluyen en el modelo 2, este se le llama *modelo de regresión lineal simple*.

Suponga que se tienen n pares de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, estos datos pueden utilizarse para estimar los valores de β_0 y β_1. Esta estimación es por el **métodos de mínimos cuadrados**.

Mínimos Cuadrados

Entonces la ecuación 2 se puede reescribir como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ para } i = 1, 2, \ldots, n. \tag{3}$$

Si consideramos la suma de los cuadrados de los errores aleatorios, es decir, el cuadrado de la diferencia entre las observaciones con la recta de regresión

$$L = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{4}$$

Mínimos Cuadrados

Para obtener los estimadores por mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, es preciso calcular las derivadas parciales con respecto a β_0 y β_1, igualar a cero y resolver el sistema de ecuaciones lineales resultante:

$$\frac{\partial L}{\partial \beta_0} = 0$$
$$\frac{\partial L}{\partial \beta_1} = 0$$

evaluando en $\hat{\beta}_0$ y $\hat{\beta}_1$, se tiene

Mínimos Cuadrados

$$-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right) = 0$$

$$-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)x_i = 0$$

simplificando

$$n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i$$

Mínimos Cuadrados

Las ecuaciones anteriores se les denominan *ecuaciones normales de mínimos cuadrados* con solución

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \tag{6}$$

entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{7}$$

Mínimos Cuadrados

Se intrduce la siguiente notación

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{8}$$

$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{9}$$

y por tanto

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{10}$$

Propiedades de los estimadores

Nota

Propiedades de los estimadores

Nota

- *Las propiedades estadísticas de los estimadores de mínimos cuadrados son útiles para evaluar la suficiencia del modelo.*
- *Dado que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las variables aleatorias y_i, también resultan ser variables aleatorias.*

A saber

$$E\left(\hat{\beta}_1\right) = E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} y_i \left(x_i - \overline{x}\right)\right)$$

Propiedades de los estimadores

$$
\begin{aligned}
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \epsilon_i)(x_i - \overline{x})\right) \\
&= \frac{1}{S_{xx}} \left[\beta_0 E\left(\sum_{k=1}^{n} (x_k - \overline{x})\right) + E\left(\beta_1 \sum_{k=1}^{n} x_k (x_k - \overline{x})\right)\right. \\
&+ \left. E\left(\sum_{k=1}^{n} \epsilon_k (x_k - \overline{x})\right)\right] = \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1
\end{aligned}
$$

por lo tanto

$$E\left(\hat{\beta}_1\right) = \beta_1 \tag{11}$$

Propiedades de los estimadores

Nota

Es decir, $\hat{\beta}_1$ es un estimador insesgado.

Ahora calculemos la varianza:

$$
\begin{aligned}
V\left(\hat{\beta}_1\right) &= V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left(\sum_{k=1}^{n} y_k\left(x_k - \overline{x}\right)\right) \\
&= \frac{1}{S_{xx}^2} \sum_{k=1}^{n} V\left(y_k\left(x_k - \overline{x}\right)\right) = \frac{1}{S_{xx}^2} \sum_{k=1}^{n} \sigma^2 \left(x_k - \overline{x}\right)^2 \\
&= \frac{\sigma^2}{S_{xx}^2} \sum_{k=1}^{n} \left(x_k - \overline{x}\right)^2 = \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

Propiedades de los estimadores

por lo tanto

$$V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \tag{12}$$

Proposición

$$\begin{aligned} E\left(\hat{\beta}_0\right) &= \beta_0, \\ V\left(\hat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right), \\ Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= -\frac{\sigma^2\overline{x}}{S_{xx}}. \end{aligned}$$

Propiedades de los estimadores

Para estimar σ^2 es preciso definir la diferencia entre la observación y_k, y el valor predecido $\hat{y}_k$, es decir

$$e_k = y_k - \hat{y}_k, \text{ se le denomina } \textbf{residuo}.$$

La suma de los cuadrados de los errores de los reisduos, *suma de cuadrados del error*

$$SC_E = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \tag{13}$$

Propiedades de los estimadores

sustituyendo $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ se obtiene

$$\begin{aligned}
SC_E &= \sum_{k=1}^{n} y_k^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}, \\
E\left(SC_E\right) &= \left(n-2\right)\sigma^2, \text{ por lo tanto} \\
\hat{\sigma}^2 &= \frac{SC_E}{n-2} = MC_E \text{ es un estimador insesgado de } \sigma^2.
\end{aligned}$$

Prueba de Hipótesis

- Para evaluar la suficiencia del modelo de regresión lineal simple, es necesario lleva a cabo una prueba de hipótesis respecto de los parámetros del modelo así como de la construcción de intervalos de confianza.
- Para poder realizar la prueba de hipótesis sobre la pendiente y la ordenada al orígen de la recta de regresión es necesario hacer el supuesto de que el error ϵ_i se distribuye normalmente, es decir $\epsilon_i \sim N\left(0, \sigma^2\right)$.

Prueba de Hipótesis

Suponga que se desea probar la hipótesis de que la pendiente es igual a una constante, $\beta_{0,1}$ las hipótesis Nula y Alternativa son:

H_0: $\beta_1 = \beta_{1,0}$,

H_1: $\beta_1 \neq \beta_{1,0}$.

donde dado que las $\epsilon_i \sim N\left(0, \sigma^2\right)$, se tiene que y_i son variables aleatorias normales $N\left(\beta_0 + \beta_1 x_1, \sigma^2\right)$. De las ecuaciones (5) se desprende que $\hat{\beta}_1$ es combinación lineal de variables aleatorias normales independientes, es decir, $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/S_{xx}\right)$, recordar las ecuaciones (11) y (12).

Prueba de Hipótesis

Entonces se tiene que el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{MC_E/S_{xx}}} \tag{14}$$

que se distribuye t con $n-2$ grados de libertad bajo $H_0 : \beta_1 = \beta_{1,0}$. Se rechaza H_0 si

$$t_0| > t_{\alpha/2,n-2}. \tag{15}$$

Prueba de Hipótesis

Para β_0 se puede proceder de manera análoga para

H_0 : $\beta_0 = \beta_{0,0}$,

H_1 : $\beta_0 \neq \beta_{0,0}$,

con $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$, por lo tanto

$$t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{MC_E\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}, \tag{16}$$

con el que rechazamos la hipótesis nula si

$$t_0| > t_{\alpha/2,n-2}. \tag{17}$$

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno.

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno.

Prueba de Hipótesis

- No rechazar $H_0 : \beta_1 = 0$ es equivalente a decir que no hay relación lineal entre x y y.
- Alternativamente, si $H_0 : \beta_1 = 0$ se rechaza, esto implica que x explica la variabilidad de y, es decir, podrÃa significar que la línea recta esel modelo adecuado.

El procedimiento de prueba para $H_0 : \beta_1 = 0$ puede realizarse de la siguiente manera:

$$S_{yy} = \sum_{k=1}^{n} (y_k - \overline{y})^2 = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

Prueba de Hipótesis

$$\begin{aligned}
S_{yy} &= \sum_{k=1}^{n} (y_k - \overline{y})^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k + \hat{y}_k - \overline{y})^2 \\
&= \sum_{k=1}^{n} [(\hat{y}_k - \overline{y}) + (y_k - \hat{y}_k)]^2 \\
&= \sum_{k=1}^{n} \left[(\hat{y}_k - \overline{y})^2 + 2\,(\hat{y}_k - \overline{y})\,(y_k - \hat{y}_k) + (y_k - \hat{y}_k)^2 \right] \\
&= \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + 2\sum_{k=1}^{n} (\hat{y}_k - \overline{y})\,(y_k - \hat{y}_k) + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2
\end{aligned}$$

Prueba de Hipótesis

$$\sum_{k=1}^{n} (\hat{y}_k - \overline{y}) (y_k - \hat{y}_k) = \sum_{k=1}^{n} \hat{y}_k (y_k - \hat{y}_k) - \sum_{k=1}^{n} \overline{y} (y_k - \hat{y}_k)$$

$$= \sum_{k=1}^{n} \hat{y}_k (y_k - \hat{y}_k) - \overline{y} \sum_{k=1}^{n} (y_k - \hat{y}_k)$$

$$= \sum_{k=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right) \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) - \overline{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right)$$

Prueba de Hipótesis

$$\begin{aligned}
&= \sum_{k=1}^{n} \hat{\beta}_0 \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) + \sum_{k=1}^{n} \hat{\beta}_1 x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&- \bar{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&= \hat{\beta}_0 \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) + \hat{\beta}_1 \sum_{k=1}^{n} x_k \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) \\
&- \bar{y} \sum_{k=1}^{n} \left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k \right) = 0 + 0 + 0 = 0.
\end{aligned}$$

Prueba de Hipótesis

Por lo tanto, efectivamente se tiene

$$S_{yy} = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 + \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 , \tag{18}$$

donde se hacen las definiciones

$$SC_E = \sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2 \cdots \text{Suma de Cuadrados del Error} \tag{19}$$

$$SC_R = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \cdots \text{Suma de Regresión de Cuadrados} \tag{20}$$

Prueba de Hipótesis

Por lo tanto la ecuación (18) se puede reescribir como

$$S_{yy} = SC_R + SC_E \tag{21}$$

recordemos que $SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$

$$\begin{aligned} S_{yy} &= SC_R + \left(S_{yy} - \hat{\beta}_1 S_{xy}\right) \\ S_{xy} &= \frac{1}{\hat{\beta}_1} SC_R \end{aligned}$$

S_{xy} tiene $n-1$ grados de libertad y SC_R y SC_E tienen 1 y $n-2$ grados de libertad respectivamente.

Prueba de Hipótesis

Proposición

$$E\left(SC_R\right) = \sigma^2 + \beta_1 S_{xx} \tag{22}$$

además, SC_E y SC_R son independientes.

Recordemos que $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. Para $H_0 : \beta_1 = 0$ verdadera,

$$F_0 = \frac{SC_R/1}{SC_E/(n-2)} = \frac{MC_R}{MC_E}$$

se distribuye $F_{1,n-2}$, y se rechazaría H_0 si $F_0 > F_{\alpha,1,n-2}$.

Prueba de Hipótesis

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0

Prueba de Hipótesis

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E

Prueba de Hipótesis

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E
Error Residual	SC_E	$n-2$	MC_E	

Prueba de Hipótesis

El procedimiento de prueba de hipótesis puede presentarse como la tabla de análisis de varianza siguiente

Fuente de variación	Suma de Cuadrados	Grados de Libertad	Media Cuadrática	F_0
Regresión	SC_R	1	MC_R	MC_R/MC_E
Error Residual	SC_E	$n-2$	MC_E	
Total	S_{yy}	$n-1$		

Prueba de Hipótesis

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (14), con $\hat{\beta}_{1,0} = 0$, es decir

Prueba de Hipótesis

La prueba para la significación de la regresión puede desarrollarse basándose en la expresión (14), con $\hat{\beta}_{1,0} = 0$, es decir

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MC_E/S_{xx}}} \tag{23}$$

Elevando al cuadrado ambos términos:

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MC_E} = \frac{\hat{\beta}_1 S_{xy}}{MC_E} = \frac{MC_R}{MC_E}$$

Observar que $t_0^2 = F_0$, por tanto la prueba que se utiliza para t_0 es la misma que para F_0.

Intervalos de Confianza

- Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros.

Intervalos de Confianza

- Además de la estimación puntual para los parámetros β_1 y β_0, es posible obtener estimaciones del intervalo de confianza de estos parámetros.
- El ancho de estos intervalos de confianza es una medida de la calidad total de la recta de regresión.

Intervalos de Confianza

Si los ϵ_k se distribuyen normal e independientemente, entonces

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{MC_E}{S_{xx}}}} \quad y \quad \frac{\left(\hat{\beta}_0 - \beta_0\right)}{\sqrt{MC_E\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$$

se distribuyen t con $n-2$ grados de libertad.

Intervalos de Confianza

Si los ϵ_k se distribuyen normal e independientemente, entonces

$$\frac{\left(\hat{\beta}_1 - \beta_1\right)}{\sqrt{\frac{MC_E}{S_{xx}}}} \quad y \quad \frac{\left(\hat{\beta}_0 - \beta_0\right)}{\sqrt{MC_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

se distribuyen t con $n-2$ grados de libertad.Por tanto un intervalo de confianza de $100\,(1-\alpha)\,\%$ para β_1 está dado por

Intervalos de Confianza

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \qquad (24)$$

Intervalos de Confianza

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MC_E}{S_{xx}}}. \qquad (24)$$

De igual manera, para β_0 un intervalo de confianza al $100\,(1-\alpha)\,\%$ es

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MC_E\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)} \quad (25)$$

Predicción

Supongamos que se tiene un valor x_0 de interés, entonces la estimación puntual de este nuevo valor

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{26}$$

Nota

Esta nueva observación es independiente de las utilizadas para obtener el modelo de regresión, por tanto, el intervalo en torno a la recta de regresión es inapropiado, puesto que se basa únicamente en los datos empleados para ajustar el modelo de regresión. El intervalo de confianza en torno a la recta de regresión se refiere a la respuesta media verdadera $x = x_0$, no a observaciones futuras.

Predicción

Sea y_0 la observación futura en $x = x_0$, y sea $\hat{y}_0$ dada en la ecuación anterior, el estimador de y_0. Si se define la variable aleatoria

$$w = y_0 - \hat{y}_0,$$

esta se distribuye normalmente con media cero y varianza

$$V(w) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - x_0)^2}{S_{xx}} \right]$$

dado que y_0 es independiente de $\hat{y}_0$, por lo tanto el intervalo de predicción al nivel α para futuras observaciones x_0 es

Predicción

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MC_E\left[1+\frac{1}{n}+\frac{(x-x_0)^2}{S_{xx}}\right]} \leq y_0$$

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

Coeficiente de Determinación

La cantidad

$$R^2 = \frac{SC_R}{S_{yy}} = 1 - \frac{SC_E}{S_{yy}} \tag{27}$$

se denomina coeficiente de determinación y se utiliza para saber si el modelo de regresión es suficiente o no. Se puede demostrar que $0 \leq R^2 \leq 1$, una manera de interpretar este valor es que si $R^2 = k$, entonces el modelo de regresión explica el $k * 100\%$ de la variabilidad en los datos.

Coeficiente de Determinación

R^2

- No mide la magnitud de la pendiente de la recta de regresión
- Un valor grande de R^2 no implica una pendiente empinada.
- No mide la suficiencia del modelo.
- Valores grandes de R^2 no implican necesariamente que el modelo de regresión proporcionará predicciones precisas para futuras observaciones.