Análisis de Regresión Lineal Simple

ANÁLISIS DE REGRESIÓN

Métodos Estadísticos

MET. Alejandra Cerda

Regresión Lineal Simple

Un modelo de regresión lineal simple es un modelo de regresión donde interviene una variable regresora X, que tiene una relación con una respuesta Y, donde la relación es una línea recta. Este modelo es:

$$y = \beta_0 + \beta_1 x + e$$

Donde la ordenada al origen β_0 y la pendiente β_1 son constantes desconocidas, y e es un componente aleatorio de error. A los parámetros β_0 y β_1 se les suele llamar coeficientes de regresión.

La diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ se llama error o residual y se denota por e_i . Se supone que los errores tienen promedio cero y varianza σ^2 desconocida. Además, se suele suponer que los errores no están correlacionados.

Datos para regresión lineal simple

Supongamos que se dispone de n observaciones, con y_i la i-ésima respuesta observada, y x_i la i-ésima observación.

OBSERVACIÓN	RESPUESTA	REGRESOR
i	y	x
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
.	.	.
.	.	.
.	.	.
n	y_n	x_n

Estimación del modelo

MÉTODO DE MÍNIMOS CUADRADOS

Métodos Estadísticos

MET. Alejandra Cerda

Estimación por mínimos cuadrados

Los parámetros β_0 y β_1 se deben estimar con los datos de la muestra, buscando que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima. Se puede escribir en la siguiente forma el modelo de regresión:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

De tal manera que la función de mínimos cuadrados es

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Se debe minimizar la función S respecto a β_0 y β_1. Los estimadores de mínimos cuadrados deben satisfacer las ecuaciones:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

Ecuaciones normales

Al simplificar las ecuaciones anteriores se obtienen las ecuaciones normales de mínimos cuadrados

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

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

y su solución es :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

Prueba de Hipótesis

REGRESIÓN LINEAL SIMPLE

Métodos Estadísticos

MET. Alejandra Cerda

Prueba de significancia

La prueba de significancia de la regresión es para determinar si hay una relación lineal entre la respuesta y y la variable regresora. Las hipótesis correspondientes son:

$H_0: \beta_1 = 0$ (la regresión no es significativa, las variables involucradas no muestran relación)

$H_a: \beta_1 \neq 0$ (la regresión es significativa, las variables involucradas muestran relación)

Rechazando H_0 si $EP: F_0 > F_{\alpha,1,n-2}$ o bajo el criterio del p valor rechazo H_0 si $p\,valor < \alpha$ donde α es el nivel de significancia de la prueba.

TABLA ANOVA

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrado medio	F0
Regresión	1	$SSR=\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR=SSR/1	MSR/MSE
Residuales	n-2	$SSE=\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE=SSE/(n-2)	
Total	n-1	$SST=\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Intervalos de Confianza

Para construir intervalos de confianza de los coeficientes de regresión β_0 y β_1 , se considera que los errores están distribuidos normal e independientemente, entonces las distribuciones de muestreo tanto de

$$\frac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1-\beta_1}{\sqrt{MSE/\sum(x_i-\bar{x})^2}} \qquad \text{y} \qquad \frac{\hat{\beta}_0-\beta_0}{se(\hat{\beta}_0)} = \frac{\hat{\beta}_0-\beta_0}{\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)}}$$

Donde el denominador es el error estándar, es t con n-2 grados de libertad. Así el intervalo de confianza de 100(1- α)% para la pendiente β_1 queda definido como:

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{MSE/\sum(x_i-\bar{x})^2} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{MSE/\sum(x_i-\bar{x})^2}$$

Y un intervalo de confianza del $100(1-\alpha)\%$ para la ordenada al origen β_0 queda definido como:

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)}$$

Coeficiente de determinación

El coeficiente de determinación definido como la cantidad

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

se llama con frecuencia la proporción de variación explicada por el regresor x, dado que SST es una medida de variabilidad de y sin considerar el efecto de la variable regresora x y SSE es una medida de la variabilidad de y que queda después de haber tenido en consideración a x.

Ya que $0 \leq SSE \leq SST$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.

El cuadrado del coeficiente de correlación muestral da el valor del coeficiente de determinación que resultaría de ajustar el modelo de regresión lineal simple.

Nivel de desempeño, según el valor de R^2

Nivel de desempeño	Coeficiente de determinación
MUY BUENO	$0.9 < R^2 < 1$
BUENO	$0.7 < R^2 < 0.9$
MODERADO	$0.4 < R^2 < 0.7$
BAJO	$0.2 < R^2 < 0.4$
NULO	$0 \leq R^2 < 0.2$