



Curso en Análisis de regresión

Unidad 1 • Regresión lineal simple

$$\frac{1}{n}$$

$$t-1$$

$$HV_1 = VAR(S_1) = n-1$$

$$\frac{1}{n} \sum_{t=1}^n X_2^t$$

$$HV_2^? = VAR(S_2) = \frac{1}{n-1}$$

$$(S_1, S_2) = \sqrt{\frac{1}{n-2} \sum_{t=1}^n (x_1^t - \bar{x}_1)(x_2^t - \bar{x}_2)}$$

$$(S_1, S_2) = \sqrt{\frac{1}{n-2} \frac{VAR(S_1)}{HV_2}}$$

$$S_1 = CF$$



LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Autor:
Juan Carlos Rubriche Cárdenas
© 2017. Todos los derechos reservados



Preguntas orientadoras

- ¿Qué herramientas estadísticas hay para investigar si existe alguna relación de dependencia lineal entre dos variables que estemos estudiando en un problema, a partir de una muestra de datos observados?
- ¿Qué es un modelo de regresión lineal simple?
- ¿Qué métodos hay para estimar los parámetros del modelo de regresión lineal simple usando los datos observados?
- ¿Será posible construir intervalos de confianza y hacer pruebas de hipótesis para los parámetros del modelo de regresión lineal simple?
- ¿Cómo podemos medir la bondad de ajuste de un modelo de regresión lineal simple?

Síntesis

En esta unidad se estudia la metodología estadística para investigar la relación de dependencia entre dos variables aleatorias mediante un modelo de regresión lineal simple. Específicamente, se muestra cómo estimar los parámetros del modelo usando el método de mínimos cuadrados ordinarios, con una colección de datos observados de las variables y se muestra cómo construir intervalos de confianza y pruebas de hipótesis para dichos parámetros. Además, se describe el método ANOVA para medir la bondad de ajuste de un modelo.

Palabras clave

- Modelo de regresión lineal simple
- Supuestos del modelo
- Método de mínimos cuadrados
- Inferencia sobre el modelo de regresión
- ANOVA





Tema 1

¿Qué es el análisis de regresión?

El análisis de regresión es un metodología estadística para investigar la relación de dependencia entre variables a partir de un conjunto de datos observados.

En primer lugar vamos a considerar problemas que implican aproximar la relación entre dos variables mediante una línea recta. Estos problemas son comúnmente denominados como regresión lineal simple.

Un primer ejemplo

El gerente de compras de una empresa desea estimar el tiempo promedio que se toman para procesar determinado número de facturas. Con este propósito, recolecta datos sobre el número de facturas procesadas y el tiempo total empleado (en horas) para realizar dicha labor, durante un período de 30 días.

En la tabla se presentan los datos obtenidos:

Día	Facturas	Tiempo	Día	Facturas	Tiempo
1	149	12.1	16	169	2.5
2	60	1.8	17	190	2.9
3	188	2.3	18	233	3.4
4	23	0.8	19	289	4.1
5	201	2.7	20	45	1.2
6	58	1	21	193	2.5
7	77	1.7	22	70	1.8
8	222	3.1	23	241	3.8
9	181	2.8	24	103	1.5
10	30	1	25	163	2.8
11	110	1.5	26	120	2.5
12	83	1.2	27	201	3.3
13	60	0.8	28	135	2
14	25	1	29	80	1.7
15	173	2	30	29	1.5



En este ejemplo tenemos dos variables $X = \text{Facturas}$ y $Y = \text{Tiempo}$, y el objetivo es investigar cómo cambia X cuando Y varía sobre su rango de posibles valores.

La variable X se denomina variable explicativa o predictora y la variable Y se conoce como variable respuesta o dependiente.

Inicialmente podemos explorar si existe una relación de dependencia entre estas dos variables y qué tipo de relación funcional podría tenerse (lineal o curvilínea), graficando un diagrama de dispersión de los datos.

En general, si tenemos n observaciones $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, un diagrama de dispersión es la gráfica obtenida al representar estos pares de mediciones como puntos en el plano cartesiano.

Este diagrama de dispersión muestra una inclinación de los valores de Y a variar sistemáticamente respecto a los valores de X , ya que cuando aumentan el número de facturas procesadas X , aumenta el tiempo requerido Y para realizar dicha tarea. Además, hay una tendencia de los puntos a dispersarse alrededor de una línea recta denominada línea de regresión.

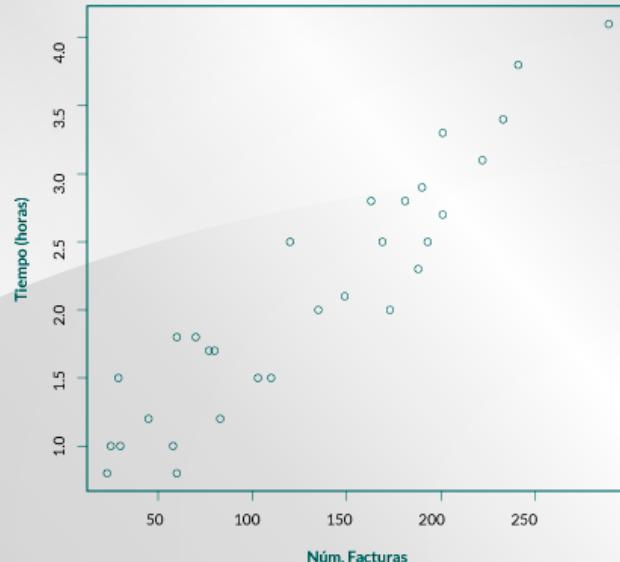


Figura 1. Diagrama de dispersión





Tema 2

Diagrama de dispersión

Modelo de regresión lineal simple

Un modelo estadístico que describe la relación de dependencia entre las variables X y Y es:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Conocido como modelo de regresión lineal simple, donde β_0 y β_1 son los parámetros del modelo y ε es el término del error aleatorio.

Tema 3

Supuestos del modelo de regresión lineal simple

El modelo (1) satisface los siguientes supuestos:

1. $E(Y | X = x) = \beta_0 + \beta_1 x$
2. $Var(Y | X = x) = \sigma^2$

El primer supuesto indica que para cada valor específico X de la variable explicativa X, la variable respuesta Y tiene una distribución de probabilidad con media que depende linealmente de X dada por la expresión: $\beta_0 + \beta_1 x$

Por otra parte, el segundo supuesto nos muestra que la distribución de probabilidad de Y para cualquier valor específico X de X tiene varianza constante σ^2 .





Tema 4

Interpretación de los parámetros del modelo

La pendiente β_1 del modelo de regresión, se interpreta como el cambio en el valor medio de la distribución de probabilidad de Y cuando hay un cambio de una unidad en X, es decir, el cambio en el tiempo medio utilizado para procesar las facturas cuando el número de facturas aumenta en una unidad.

Por otra parte, el intercepto β_0 se interpreta como el valor medio de la distribución de probabilidad de Y cuando $X = 0$. Esta interpretación es válida siempre que 0 esté en el rango de valores de X.

Las desviaciones de los valores observados y_i respecto a la media $E(Y | X = x_i)$ se les conoce como errores estadísticos. Si tenemos un conjunto de n observaciones $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ sus errores correspondientes estarían dados por la expresión:

$$\begin{aligned} e_i &= y_i - E(Y | X = x_i) \\ &= y_i - (\beta_0 + \beta_1 x_i) \end{aligned}$$

para $i = 1, 2, \dots, n$.

Los errores estadísticos e_i son realizaciones de los términos de error aleatorios ε_i en cada x_i .

Tema 5

Método de mínimos cuadrados

Surge el problema de hallar estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros β_0 y β_1 del modelo usando los datos observados $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, de tal manera que la recta que "mejor" se ajuste a los datos observados, sea:





$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

El método de mínimos cuadrados nos permite hallar estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros β_0 y β_1 del modelo de regresión lineal simple, de tal manera que la suma de los cuadrados de los errores e_i para las n observaciones muestrales $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, dados, sea mínima. En otras palabras, se busca determinar los valores de β_0 y β_1 que hagan que la suma:

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] \end{aligned}$$

sea mínima.

Tema 6

Cálculo de estimadores de mínimos cuadrados

Usando un poco de cálculo, podemos determinar que los valores de β_0 y β_1 que minimizan la suma $S(\beta_0, \beta_1)$ son:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \end{aligned}$$

donde:





$$\bar{X} = \frac{1}{n} \sum X_i, \quad \bar{Y} = \frac{1}{n} \sum Y_i \\ S_{XX} = \sum (X_i - \bar{X})^2, \quad S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

La recta dada $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, se denominada ecuación de estimación, ya que \hat{Y} estima puntualmente a $E(Y | X = x)$. Además, se dice que:

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

es el valor estimado del valor observado y_i .

Observando el diagrama de dispersión de los datos en la figura 1 (tema 1), la relación de dependencia existente entre el número de facturas procesadas X y el tiempo requerido para ejecutar esta labor Y , se puede describir mediante un modelo de regresión lineal simple dado por:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Ahora, unas estimaciones puntuales de los parámetros de este modelo β_0 y β_1 usando el método de los mínimos cuadrados, son:

$$\hat{\beta}_0 = 0,6417 \\ \hat{\beta}_1 = 0,0113$$

Por lo tanto, la ecuación de estimación es:

$$\hat{y} = 0,6417 + 0,0113X$$

Aquí, la pendiente 0,0113 significa que un aumento de una factura para procesar produce un cambio en el tiempo medio de 0,0013 horas aproximadamente. Por otra parte, el intercepto 0,6417 horas se interpreta como el tiempo de alistamiento medio necesario para procesar cualquier número de facturas.





Tema 7 Estimación de la varianza

Considere el modelo de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_I + \varepsilon_i$$

cuyos términos de error aleatorio satisfacen que $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$.

Queremos estimar la varianza constante σ^2 de los términos de error aleatorios ε_i . Con este fin vamos a calcular los residuales \hat{e}_i asociados a los n datos observados, los cuales se definen como:

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_I)$$

Los residuales asociados en el ejemplo que estamos considerando de facturas y tiempo, se muestran en la tabla. Entonces un estimador puntual $\hat{\sigma}^2$ de la varianza del modelo σ^2 se calcula promediando los cuadrados de los residuales, es decir:

$$\hat{\sigma}^2 = \frac{I}{n - 2} \sum \hat{e}_i^2$$

Dividimos entre $n-2$ porque este es el número de observaciones independientes o grados de libertad en esta estimación, ya que en el cálculo de residuales se estiman los parámetros β_0 y β_1 a partir de los datos y como consecuencia se pierden dos grados de libertad.





Obs.	y_i	\hat{y}_i	\hat{e}_i
1	2.1	2.3242	-0.2242
2	1.8	1.3192	0.4808
3	2.3	2.7645	-0.4645
4	0.8	0.9014	-0.1014
5	2.7	2.9113	-0.2113
6	1	1.2966	-0.2966
7	1.7	1.5112	0.1888
8	3.1	3.1485	-0.0485
9	2.8	2.6855	0.1145
10	1	0.9805	0.0195
11	1.5	1.8838	-0.3838
12	1.2	1.5789	-0.3789
13	0.8	1.3192	-0.5192
14	1	0.9240	0.0760
15	2	2.5952	-0.5952

Tema 8

Inferencia acerca de los parámetros β_0 y β_1

Vamos a construir intervalos de confianza y realizar pruebas de hipótesis para los parámetros del modelo de regresión lineal simple β_0 y β_1 . Para ello necesitamos considerar cuidadosamente los supuestos sobre los términos de error del modelo, a saber:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

- La variable respuesta Y y la variable predictora X están relacionadas por el modelo
- Los términos de error ε_i tienen media 0 y varianza común constante σ^2
- Los términos de error ε_i están distribuidos normalmente
- Los términos de error ε_i no están correlacionados

Estos dos últimos supuestos implican que los errores son independientes.





Se sabe que el estadístico muestral $\hat{\beta}_1$ dado tiene una distribución de probabilidad normal con media β_1 y varianza σ^2/S_{xx} . Así, estandarizando $\hat{\beta}_1$ tenemos:

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$$

Si se conociera σ^2 , usando el estadístico, podríamos hacer inferencia sobre el parámetro β_1 . Pero como usualmente este no es el caso, debemos estimar σ^2 mediante $\hat{\sigma}^2$ y se tiene que el estadístico:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}$$

Usando el estadístico (7) tenemos que:

$$\hat{\sigma}^2 = \frac{I}{n-2} \sum \hat{e}_i^2$$

Un intervalo de confianza de $(1 - \alpha) 100\%$ para β_1 es:

$$\hat{\beta}_1 - t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n-2$ grados de libertad que deja un área de $\alpha/2$ a su derecha.





En relación con las hipótesis, Nos interesa probar la hipótesis $H_0: \beta_1 = 0$ contra la hipótesis $H_1: \beta_1 \neq 0$ a un nivel de significancia α dado. Para esta prueba se usa el estadístico de prueba:

$$T = \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{XX}}}$$

el cual tiene una distribución t con $n - 2$ grados de libertad cuando suponemos que la hipótesis nula es verdadera. Si no se rechaza la hipótesis nula $H_0: \beta_1 = 0$ en contra de $H_1: \beta_1 \neq 0$, esto nos indicaría que no hay una relación de tipo lineal entre las variables X e Y.

Por otra parte, se sabe que el estimador $\hat{\beta}_0$ dado $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ tiene una distribución normal con media β_0 y varianza:

$$\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

Así, estandarizando a β_0 obtenemos:

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim N(0, 1)$$

Este estadístico estandarizado es útil para hacer inferencia sobre el intercepto β_0 cuando se conoce la varianza poblacional σ^2 . No obstante, como usualmente se desconoce la varianza σ^2 debemos estimarla usando $\hat{\sigma}^2$ y al sustituirla en este estadístico, obtenemos el estadístico muestral:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim t_{(n-2)}$$

Usando el estadístico anterior, podemos construir intervalos de confianza y hacer pruebas de hipótesis para β_0 . Así,





1. Un intervalo de confianza de $(1 - \alpha)100\%$ para β_0 es:

$$\hat{\beta}_0 - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}} < \beta_0 < \hat{\beta}_0 + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n-2$ grados de libertad que deja un área de $\alpha/2$ a su derecha.

2. Una prueba de hipótesis de interés sobre el intercepto es $H_0: \beta_0 = 0$ contra $H_1: \beta_0 \neq 0$. Esta prueba se puede realizar usando el estadístico de prueba cuando se supone que H_0 es verdadera:

$$T = \frac{\hat{\beta}_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}}$$

el cual tiene una distribución t con $n-2$ grados de libertad. Si no se rechaza H_0 podemos concluir con cierto nivel de significancia que la recta de regresión pasa a través del origen.

Tema 9

Predicción

Los modelos de regresión pueden ser usados para predecir el valor real de Y en un valor específico x^* de X o para predecir el valor medio de Y en un valor particular x^* de X ; bajo el supuesto de que el modelo de regresión lineal simple es válido para hacer estas predicciones usando los datos hasta ahora observados.

Denotaremos como Y^* al valor real de la variable Y que deseamos predecir en el valor x^*





de X . Usamos la recta de regresión estimada, a saber $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ para predecir el valor y^* .

El error que se comete al hacer esta predicción $Y^* - \hat{y}^*$ es una variable aleatoria que tiene una distribución de probabilidad normal con media 0 y varianza dada por:

$$Var(Y^* - \hat{y}^*) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right)$$

Como se observa en la expresión anterior hay dos fuentes de variación en la predicción del valor Y^* , a saber, la variación debida al término de error ε^* y la variación debida al error de estimación de los parámetros del modelo.

Estandarizando la variable $Y^* - \hat{y}^*$ y sustituyendo $\hat{\sigma}$ por σ , obtenemos:

$$T = \frac{Y^* - \hat{y}^*}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

Usando el estadístico anterior concluimos que un intervalo de predicción de $(1 - \alpha)100\%$ para Y^* es:

$$\hat{y}^* - \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}} < Y^* < \hat{y}^* + \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n-2$ grados de libertad que deja un área de $\alpha/2$ a su derecha.

Por otra parte, podemos construir un intervalo de confianza para $E(Y | X = x^*) = \beta_0 + \beta_1 x^*$ la respuesta media en el valor x^* de X . Para ello, considere el estimador $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ que se distribuye normalmente con media $\beta_0 + \beta_1 x^*$ y varianza dada por:





$$Var(\hat{y}^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right)$$

Estandarizando el estadístico muestral \hat{y}^* y reemplazando σ por $\hat{\sigma}$ obtenemos que:

$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

Por lo tanto, un intervalo de confianza de $(1-\alpha)100\%$ para $E(Y|X=x^*) = \beta_0 + \beta_1 x^*$ es:

$$\hat{y}^* - t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}} < E(Y|X=x^*) < \hat{y}^* + t_{\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n-2$ grados de libertad que deja un área de $\alpha/2$ a su derecha.

Tema 10

Análisis de varianza

En un modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$





la variación total originada por los valores Y_i en una muestra $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, se puede descomponer en la suma de dos componentes, la variación explicada por el modelo y la variación inexplicada por el modelo.

$$\text{Variación total de los } y = \text{Variación explicada por el modelo} + \text{Variación inexplicada por el modelo}$$

La variación total originada por los y observados en una muestra se denominada suma total de cuadrados (corregida) (STC), y se cuantifica así:

$$SCR = \sum(y_i - \hat{y}_i)^2$$

$$SCE = \sum(\hat{y}_i - \bar{y})^2$$

Por otra parte, la variabilidad explicada por el modelo y la variabilidad no explicada por el mismo se denominan suma de cuadrados de la regresión (SCR) y suma de cuadrados del error (SCE), y se miden respectivamente así:

$$STC = \sum(y_i - \bar{y})^2$$

Por lo tanto, las medidas de variabilidad STC , SCR y SCE están relacionadas mediante la siguiente ecuación fundamental:

$$STC = SCR + SCE$$

Una medida descriptiva de la fuerza de la relación lineal entre X e Y es el coeficiente de determinación R^2 , el cual se define como:





$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC}$$

R^2 se interpreta como la proporción de la variabilidad total en los y observados que es explicada por el modelo de regresión. Esta medida no depende de la escala de medición de los datos y satisface que $0 \leq R^2 \leq 1$. Por lo tanto, un valor de R^2 cercano a 1 muestra que la relación lineal entre las variables es fuerte mientras que un valor de R^2 cercano a 0 indica que la relación lineal entre las variables es débil.

Una prueba muy usada para probar la hipótesis nula $H_0: \beta_1 = 0$ contra la hipótesis alternativa $H_1: \beta_1 \neq 0$ usando el estadístico de prueba:

$$F = \frac{SCR / 1}{SCE / (n - 2)}$$

se denomina la prueba F, porque bajo los supuestos del modelo y la premisa de que H_0 es verdadera SCR/σ^2 y SCE/σ^2 son estadísticos independientes que tienen distribuciones de probabilidad χ^2 con 1 y $n-2$ grados de libertad respectivamente. Por lo tanto, concluimos que:

$$F = \frac{SCR / 1}{SCE / (n - 2)} \sim F_{(1, n-2)}$$

cuando H_0 se supone verdadera. Así, para un nivel de significancia dado α , se rechaza H_0 si el valor calculado de este estadístico de prueba F^* cumple alguno de los siguientes criterios:

- $F^* > F_\alpha$, donde F_α es el valor de F con 1 y $n-2$ grados de libertad que deja un área de α a su derecha
- $p\text{-valor}(F^*) < \alpha$



Bibliografía

- Montgomery D., Peck E. y Vining G. (2006). Introducción al análisis de regresión lineal. Tercera edición. México: CECSA
- Kutner M., Nachtsheim C., Neter J. y Li W. (2005). Applied Linear Statistical Models. Quinta edición. New York: McGraw-Hill.
- Sheather S. (2009). A Modern Approach to Regression with R. New York: Springer.
- Weisberg S. (2005). Applied Linear Regression. Tercera edición. New Jersey: Wiley.

