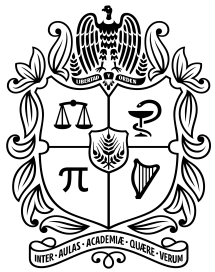
Estadística II - 3006915
Regresión Lineal Simple

Mateo Ochoa Medina

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

Periodo académico 2023-2S



UNIVERSIDAD
NACIONAL
DE COLOMBIA

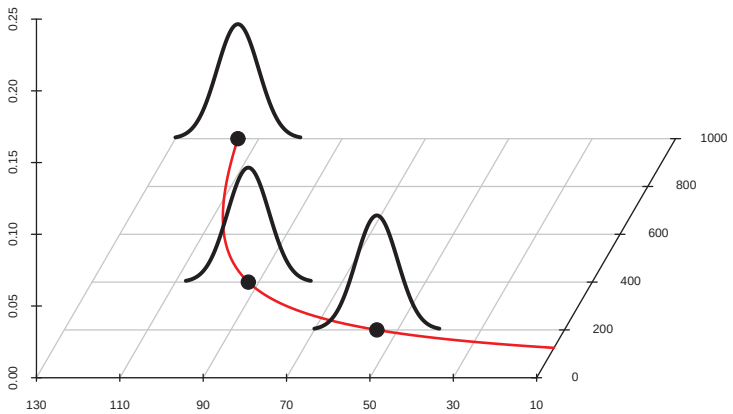
Contenido

Contenido

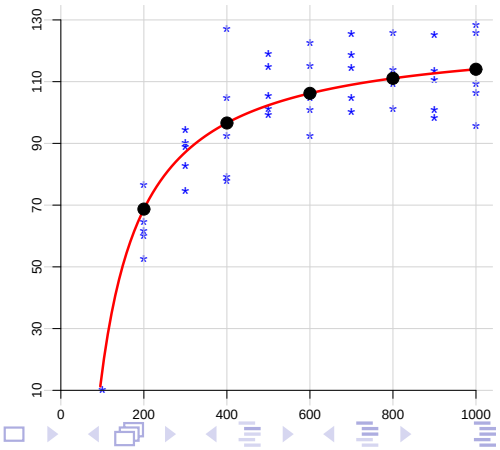
Regresión

La regresión es una técnica estadística para investigar y modelar la relación entre dos o más variables. De esta manera, la regresión tiene dos significados (considerando solo dos variables, X e Y):

- Podemos verla a partir de la distribución conjunta de las variables X e Y, en la cual podemos definir la distribución condicional de $Y \mid X$, $f(Y \mid X)$ y determinar $E(Y \mid X)$. En este caso la regresión pretende ajustar la curva correspondiente a $E(Y \mid X)$.



- Dado un conjunto de pares de datos (x_i, y_i), $i = 1, 2, \ldots, n$, puede asumirse una forma funcional para la curva de regresión y tratar de ajustarla a los datos.



Regresión

Algunos ejemplos donde la regresión puede ser la técnica estadística a aplicar:

1. En un estudio sobre el mejoramiento de vías, se desea analizar la relación que existe entre la temperatura de la superficie de una carretera (X) y la deformación del pavimento (Y).
2. En una investigación de impacto ambiental sobre la salud, se realiza un estudio para investigar la relación entre la exposición al ruido (X) y la hipertensión (Y).
3. En un proceso de fabricación de papel se desea modelar la relación entre la concentración de sulfuro de sodio Na_2S (X) y la producción de papel de una máquina (Y), con el fin de determinar medidas de control y mejoramiento.

Nota:

Debe tenerse presente que los métodos de regresión permiten establecer asociaciones entre variables de interés entre las cuales la relación usual no es necesariamente de causa-efecto.

Si la relación entre X e Y es lineal, se refiere a un análisis de regresión lineal. En este caso, la relación de Y con una sola variable X concierne a una regresión lineal simple.

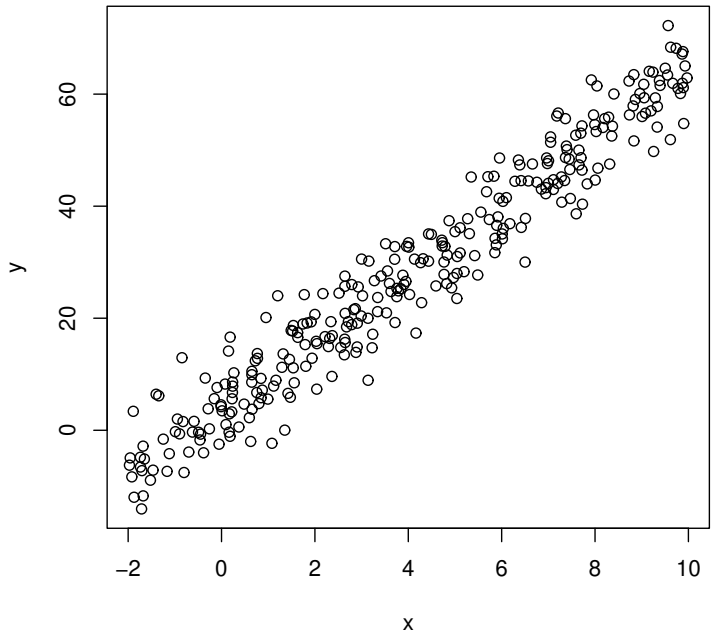


Figura 1: Este gráfico se llama *diagrama de dispersión* y, en este caso (para X e Y) sugiere una tendencia lineal en la nube de puntos. En términos prácticos, estos puntos representan pares de datos.

Contenido

Modelo de regresión lineal simple

El modelo de regresión lineal simple está dado por

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$

donde:

- β_0, β_1: Parámetros de la regresión. β_0 es el intercepto y β_1 la pendiente de la línea recta.
- X: Variable predictora, regresora o independiente.
- Y: Variable respuesta o dependiente.
- ε: Error aleatorio.

Consideraciones del modelo

Notas sobre el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$:

- La variable predictora (X) no es considerada como variable aleatoria, sino como un conjunto de valores fijos que representan los puntos de observación, que se seleccionan con anticipación y se miden sin error o con error despreciable en comparación con los errores aleatorios.
- La variable respuesta (Y) es una variable aleatoria cuyos valores se observan mediante la selección de los valores de la variable predictora en un intervalo de interés.
- El error aleatorio (ε) es una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos. Este error puede estar formado por los efectos de otras variables que no fueron consideradas en el modelo, por errores de medición u otras consideraciones no tenidas en cuenta por el investigador.
- Los datos que se observan constituyen una muestra representativa de un medio acerca del cual se desea generalizar.

Consideraciones del modelo

Notas sobre el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$:

- El modelo de regresión es lineal en los parámetros (β_0 y β_1). Es decir, ningún parámetro de la regresión aparece como el exponente o es dividido o multiplicado por otro parámetro.
- Si la ecuación de regresión ($\beta_0 + \beta_1 X$) es correcta, cualquier variabilidad en la variable respuesta (Y) que no puede ser explicada exactamente por dicha ecuación, es debida al error aleatorio (ε).
- El modelo con n pares de datos está dado por:

$$Y \mid X_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2a}$$

Por simplicidad la ecuación (2a) la escribimos como,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2b}$$

Supuestos del modelo

En el modelo de regresión $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $(i = 1, 2, \ldots, n)$ se supone que los errores aleatorios ε_i:

1. Tienen distribución normal.
2. Tienen media cero, es decir, $E(\varepsilon_i) = 0$.
3. Tienen varianza constante desconocida, es decir, $V(\varepsilon_i) = \sigma^2$.
4. Son estadísticamente independientes.

Notas:

- Los supuestos 1-4 se suelen expresar como:

$$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right), \quad i = 1, 2, \ldots, n. \tag{3}$$

- El supuesto 4 quiere decir que el valor de un error no depende del valor de cualquier otro error. Por otra parte, este supuesto implica que $COV(\varepsilon_i, \varepsilon_j) = 0, \forall\, i \neq j$ (la covarianza entre ε_i y ε_j es cero), por lo que, los errores no están correlacionados.

Características del modelo

Teniendo en cuenta el modelo de regresión $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, con $\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$, $i = 1, 2, \ldots, n$, y asumiendo que X no es variable aleatoria, luego $Y \mid X_i$ o Y_i satisfacen que:

1. Se distribuyen normalmente.
2. Tienen media dada por $E\left(Y \mid X_i\right) = E\left(Y_i\right) = \beta_0 + \beta_1 X_i$.
3. Tienen varianza dada por $V\left(Y \mid X_i\right) = V\left(Y_i\right) = \sigma^2$.
4. Son estadísticamente independientes.

Nota:

Los resultados 1-4 se suelen expresar como:

$$Y \mid X_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 X_i, \sigma^2\right), \quad i = 1, 2, \ldots, n, \tag{4a}$$

o

$$Y_i \overset{ind}{\sim} N\left(\beta_0 + \beta_1 X_i, \sigma^2\right), \quad i = 1, 2, \ldots, n. \tag{4b}$$

Consideraciones de las características del modelo

Notas sobre $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$, para $i = 1, 2, \ldots, n$ (siendo Y_i equivalente a $Y \mid X_i$):

- Al fijar los niveles o valores en que X es observada, los valores observados de Y provienen de una distribución normal. Por tanto, cuando X es X_i, la normal tiene como:
  - Media la altura de la recta de regresión $\beta_0 + \beta_1 X$ evaluada en X_i. Podemos también interpretarla como la respuesta media $E(Y \mid X) = \beta_0 + \beta_1 X$ en X_i (esto quiere decir que la recta de regresión es la función de la media condicional de Y en cualquier valor de X).
  - Varianza la de los errores aleatorios ε_i, es decir, σ^2. Por ende, esta es independiente del punto de observación, en otras palabras, del valor de X.

Consideraciones de las características del modelo

Notas sobre $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$, para $i = 1, 2, \ldots, n$ (siendo Y_i equivalente a $Y \mid X_i$):

- La independencia implica que $COV(Y_i, Y_j) = 0, \forall\, i \neq j$ (la covarianza entre Y_i y Y_j es cero), por lo que, los valores observados de Y (condicionados al valor de X, X_i) no se encuentran estadísticamente correlacionados.

- La interpretación de los coeficientes de regresión: β_1 representa el cambio en la media de Y dado un cambio unitario en X. Si el rango en que se observa X incluye al 0, entonces β_0 corresponde a la media de la distribución de Y cuando $X = 0$. Sin embargo, si $X = 0$ no ha sido observado en los datos, entonces β_0 no tiene interpretación práctica en el modelo de regresión.

Ilustración de $Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$, $i = 1, 2, \ldots, n$ (siendo Y_i equivalente a $Y \mid X_i$)

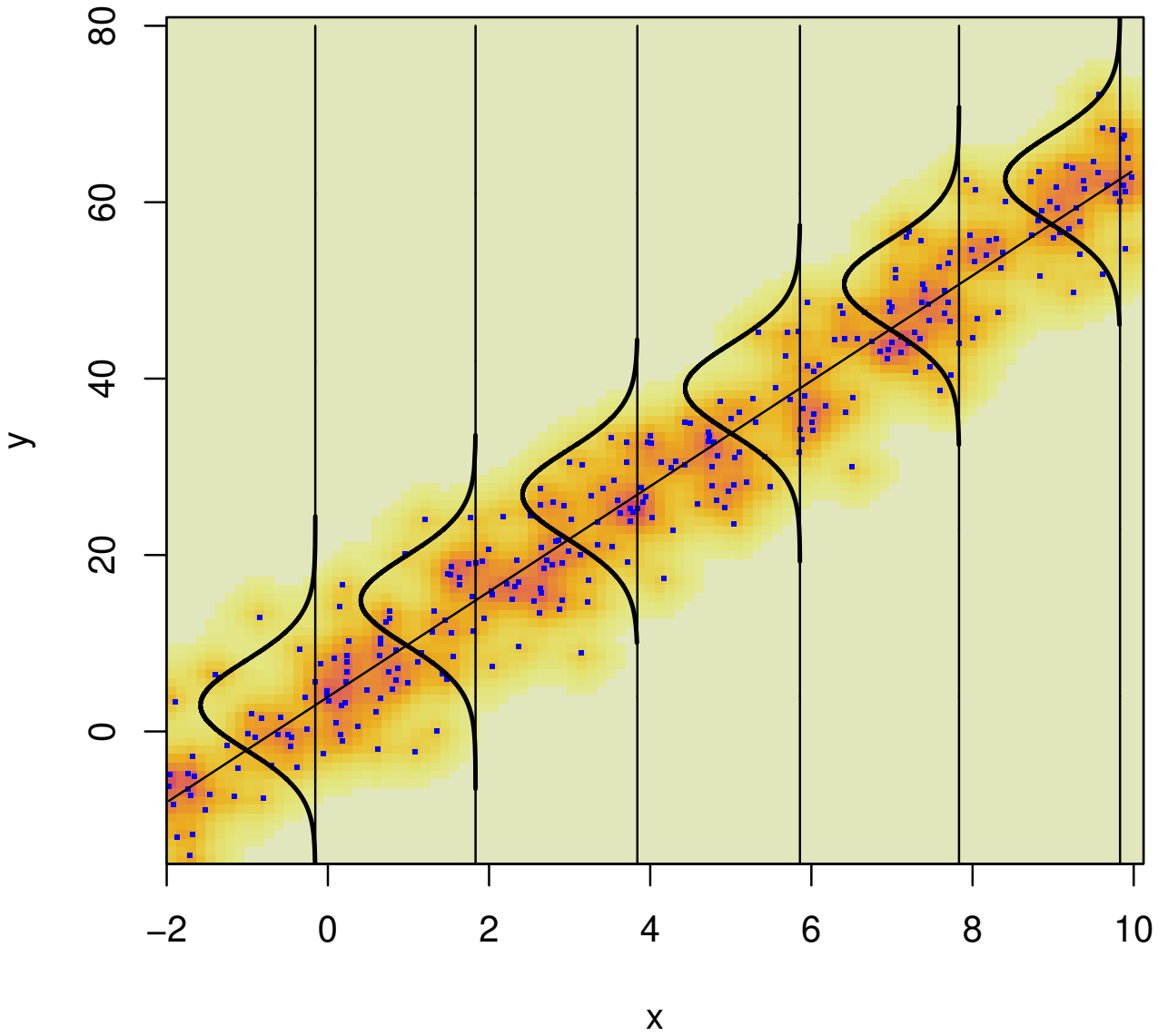


Figura 2: Forma en la que se generan las observaciones de Y en la regresión lineal cuando X es X_i. En este ejemplo estas se originan de una $N\left(4 + 6\,X_i, 25\right)$. Luego, Los valores de Y en cada nivel observado de X tienen la misma medida de dispersión alrededor de la respectiva media condicional, es decir, $E\left(Y \mid X_i\right) = 4 + 6\,X_i$.

Estimación por máxima verosimilitud de los parámetros del modelo

Sea $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ los n pares de datos observados, donde, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, con $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Asumiendo fijos los niveles o valores en que X es observada, $Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$.

Sean $\mathbf{x} = (x_1, \ldots, x_n)$ e $\mathbf{y} = (y_1, \ldots, y_n)$, los valores de X e Y, respectivamente, observados en la muestra de tamaño n. La función de verosimilitud para los parámetros del modelo: $(\beta_0, \beta_1, \sigma^2)$, es denotada por $L(\beta_0, \beta_1, \sigma^2 \,|\, \mathbf{x}, \mathbf{y})$, y es hallada a partir de la distribución conjunta de las variables Y_i, evaluada en las observaciones $y_1, y_2, y_3, \ldots, y_n$: $f(y_1, \ldots, y_n \,|\, \beta_0, \beta_1, \sigma^2)$.

El objetivo es hallar el valor de los parámetros desconocidos $\beta_0, \beta_1, \sigma^2$, que maximicen L, o equivalentemente, que maximicen el logaritmo natural de la verosimilitud, que denotaremos por $\log L$.

Estimación por máxima verosimilitud de los parámetros del modelo

De esta manera,

$$
\begin{aligned}
L\left(\beta_0,\, \beta_1,\, \sigma^2 \,|\, \mathbf{x},\, \mathbf{y}\right) &= f\left(y_1, \ldots,\, y_n \,|\, \beta_0,\, \beta_1,\, \sigma^2\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right] \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2\right].
\end{aligned}
$$

Luego,

$$
\log L = -\frac{n}{2} \log\left(2\pi\right) - \frac{n}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2.
$$

Sean $\tilde{\beta}_0$, $\tilde{\beta}_1$ y $\tilde{\sigma}^2$ los valores de los parámetros que maximizan a $\log L$ y por tanto a L.

Estimación por máxima verosimilitud de los parámetros del modelo

Así, $\tilde{\beta}_0$, $\tilde{\beta}_1$ y $\tilde{\sigma}^2$ deben satisfacer que

$$\left.\frac{\partial \log L}{\partial \beta_0}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial \log L}{\partial \beta_1}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right) x_i = 0$$

y

$$\left.\frac{\partial \log L}{\partial \sigma^2}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2 = 0.$$

Estimación por máxima verosimilitud de los parámetros del modelo

Al solucionar $\left.\frac{\partial \log L}{\partial \beta_0}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = 0$, $\left.\frac{\partial \log L}{\partial \beta_1}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = 0$ y $\left.\frac{\partial \log L}{\partial \sigma^2}\right|_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2} = 0$, se obtienen las estimaciones de máxima verosimilitud de β_0, β_1 y σ^2, respectivamente:

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} \tag{5}$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i \, (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{6}$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2}{n}, \tag{7}$$

donde $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ y $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Nota:

Las estimaciones (5), (6) y (7), son los valores de β_0, β_1 y σ^2, respectivamente, para los cuales la muestra observada es más probable.

Contenido

Referencias

Kutner, M. H., Nachtsheim, C. J., Neter, J., y Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin, New York, quinta edición.

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

Posit team (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., y De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. CRC Press, Boca Raton.

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