Multiple Linear Regression

Aniel Nieves-González

Considere el ejemplo en cual queremos modelar las ventas en una cadena de tiendas por departamento.

- La v.a. dependiente (Y) es...

Considere el ejemplo en cual queremos modelar las ventas en una cadena de tiendas por departamento.

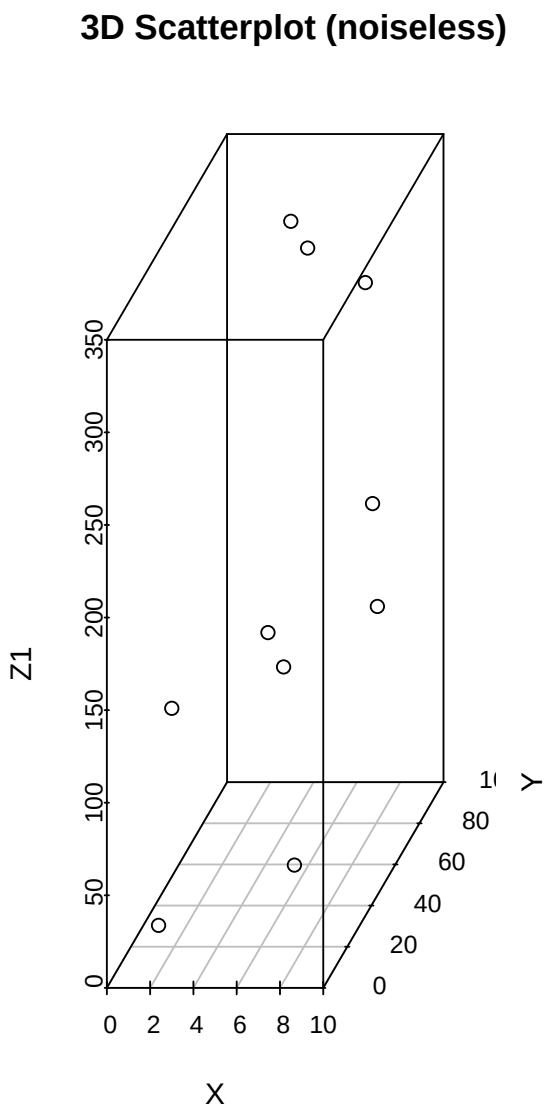
- La v.a. dependiente (Y) es...
- La(s) v.a. independiente es (son)...

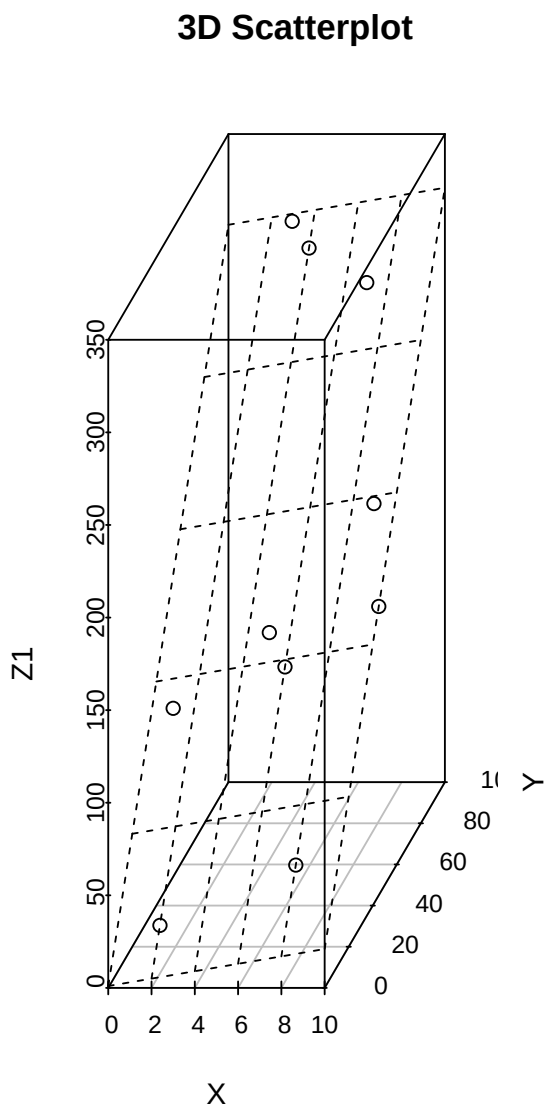
Considere el ejemplo en cual queremos modelar las ventas en una cadena de tiendas por departamento.

- La v.a. dependiente (Y) es...
- La(s) v.a. independiente es (son)...

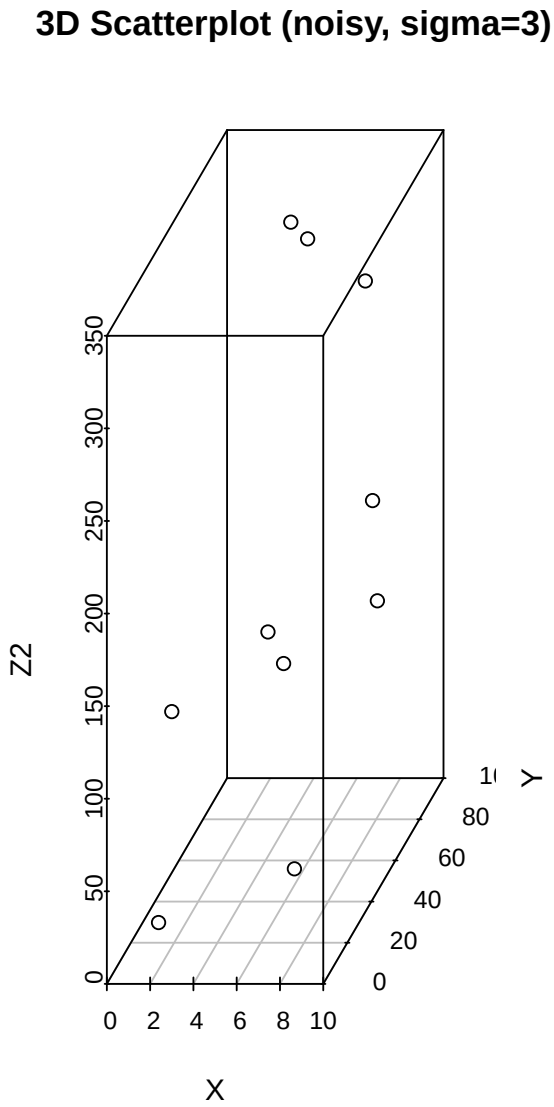
La diferencia entre el simple linear regression y el multiple linear regression es que en el primero se considera una v.a. independiente mientras que en el segundo se considera más de una v.a. independiente.

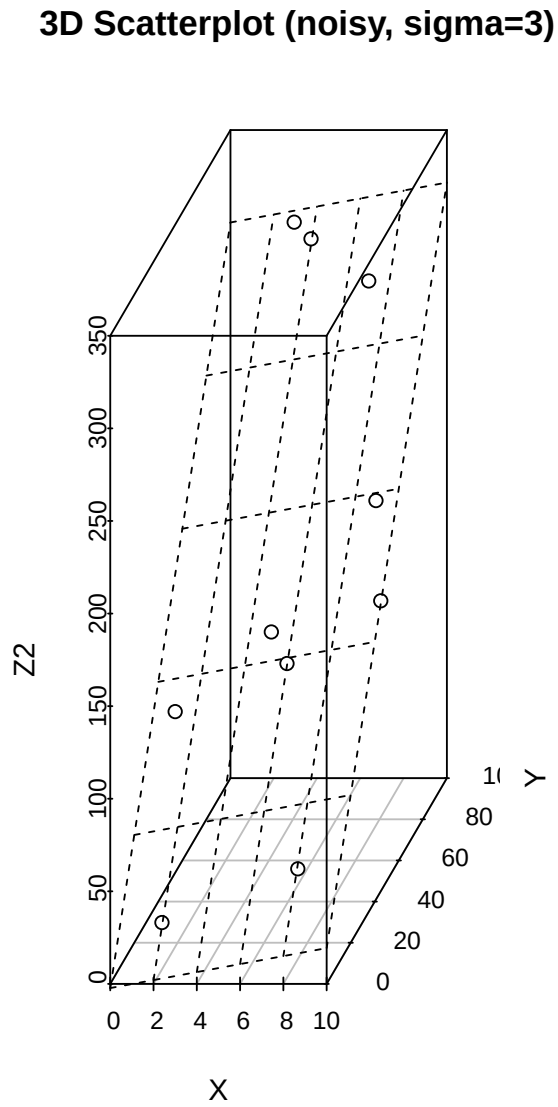
Ejemplo ficticio...



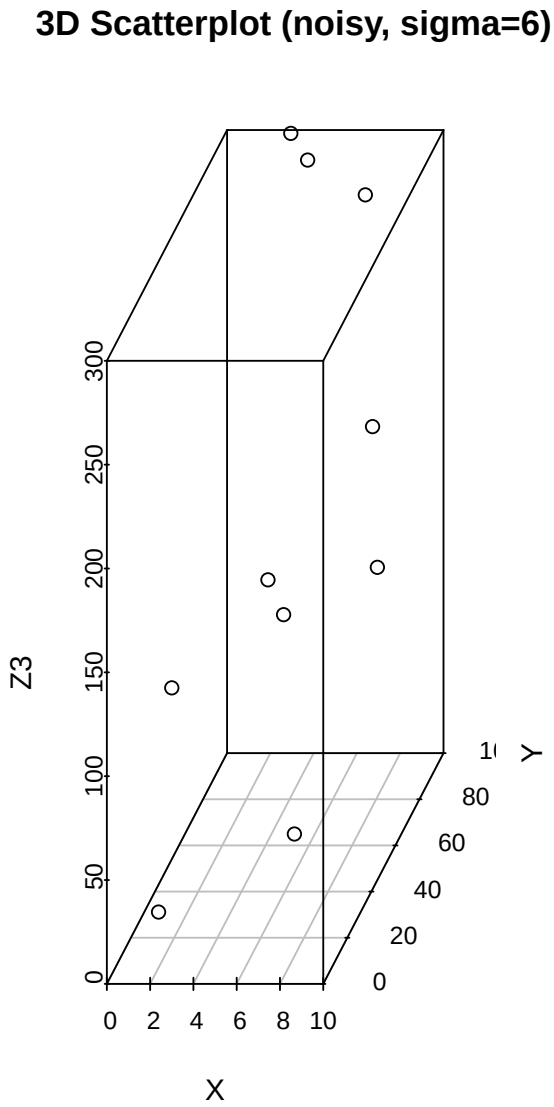


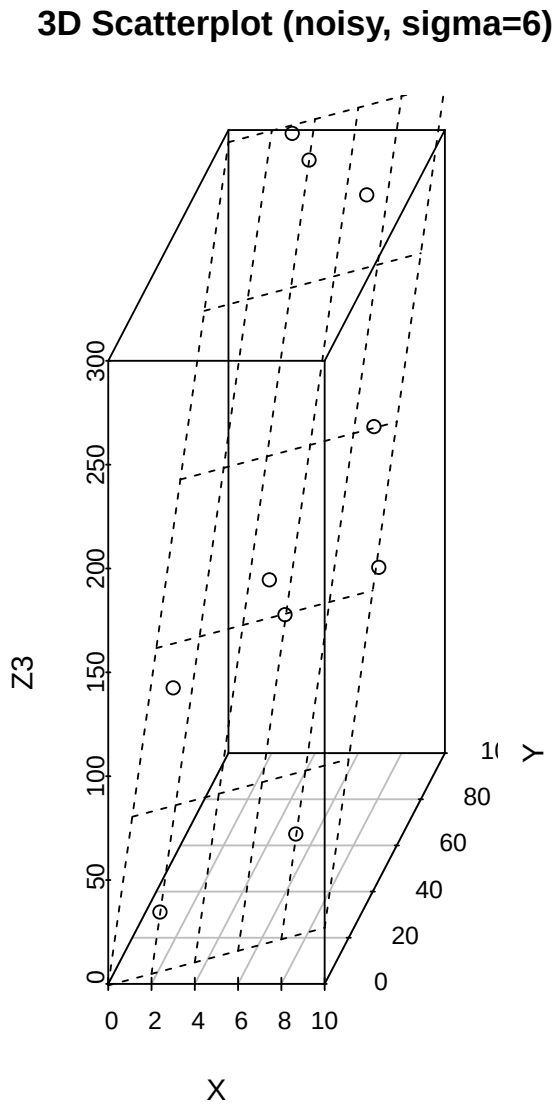
Ejemplo ficticio...





Ejemplo ficticio...





Definition (Multiple Linear Regression model)

Considere un SRS de n observaciones independientes de la variables independientes $X_1, \ldots, X_p$ y de la variable dependiente Y, donde $Y \sim N(\mu_y, \sigma)$:

$$\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{i1}, \ldots, x_{ip}, y_i), \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$$

El **modelo estadistico de regresión lineal** es

$$y_i(x_{i1}, \ldots, x_{ip}) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ip} + \epsilon_i$$

donde $i = 1, 2, \ldots, n$, $\epsilon_i \sim N(0, \sigma)$ y son independientes. Los parámetros del modelo son $\beta_0, \beta_1, \ldots, \beta_p, \sigma$. La respuesta media (mean response) es

$$\mu_y = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ip}.$$

- La meta es obtener de la muestra estimados para β_k y σ^2. A dichos estimados le llamaremos b_k y s^2_{reg} respectivamente.

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

- $\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$ son los datos (note las $p+1$ variables $X_1 \ldots, X_p$ y Y)

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

- $\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$ son los datos (note las $p+1$ variables $X_1 \ldots, X_p$ y Y)
- $\mathbf{b} = (b_0, \ldots, b_k)$ son los parámetros (constantes del modelo matemático).

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

- $\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$ son los datos (note las $p+1$ variables $X_1 \ldots, X_p$ y Y)
- $\mathbf{b} = (b_0, \ldots, b_k)$ son los parámetros (constantes del modelo matemático).
- $F(x_{i1}, \ldots, x_{ip}, \mathbf{b})$ es el modelo matemático. Note que F depende de los datos y de los parámetros.

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

- $\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$ son los datos (note las $p+1$ variables $X_1 \ldots, X_p$ y Y)
- $\mathbf{b} = (b_0, \ldots, b_k)$ son los parámetros (constantes del modelo matemático).
- $F(x_{i1}, \ldots, x_{ip}, \mathbf{b})$ es el modelo matemático. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las variables $X_1 \ldots, X_p$ y Y.

Least-squares regression (regresión por cuadrados mínimos)

Como vimos antes, En esencia el problema de cuadrados mínimos (o least squares (LS) regression) es un problema de minimización, que se resuleve para encontrar los párametros del modelo:

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (F(x_{i1}, \ldots, x_{ip}, \mathbf{b}) - y_i)^2$$

donde:

- $\{(x_{11}, \ldots, x_{1p}, y_1), \ldots, \ldots, (x_{n1}, \ldots, x_{np}, y_n)\}$ son los datos (note las $p+1$ variables $X_1 \ldots, X_p$ y Y)
- $\mathbf{b} = (b_0, \ldots, b_k)$ son los parámetros (constantes del modelo matemático).
- $F(x_{i1}, \ldots, x_{ip}, \mathbf{b})$ es el modelo matemático. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las variables $X_1 \ldots, X_p$ y Y.
- En LSP se buscan los parámetros del modelo que **minimizan la distancia (euclideana)** entre modelo y datos.

Regression standard error

La desviación estandar del modelo, esto es σ, se estima con el "standard error" de la regresión (s_{reg}):

$$s_{\text{reg}} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i(\mathbf{x}_i)\right)^2}{n - p - 1}}$$

donde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ y

$$\hat{y}_i(\mathbf{x}_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

es el modelo matemático evaluado en $\mathbf{x}_i$ (modelo lineal en este caso). Más aun note que $y_i - \hat{y}_i(\mathbf{x}_i)$ es el i-ésimo residual y $\sum_{i=1}^{n} \left(y_i - \hat{y}_i(\mathbf{x}_i)\right)^2$ es el "sum of squares residuals or error" (SSE).

Intervalo de confianza para los β_i's

Definition

El IC de confianza de nivel C para β_k es

$$b_k \pm t^* \text{SE}_{b_k}$$

donde b_k es el valor que se encuentra resolviendo el LSP, SE_{b_k} es el standard error de b_k, y t^* es el valor para la curva de densidad de la distribución $t(n-p-1)$ con area bajo la curva y sobre el intervalo $[-t^*, t^*]$ igual a C.

Test de significancia para los β_i's

Definition

Considere $H_0 : \beta_k = 0$. Calcule la estadística t:

$$t = \frac{b_k}{\text{SE}_{b_k}}.$$

Suponga que la v.a. $T \sim t(n-p-1)$, luego para:

$$\begin{aligned} H_a : \beta_k > 0, &\quad P_{\text{value}} = Pr(T \geq t) \\ H_a : \beta_k < 0, &\quad P_{\text{value}} = Pr(T \leq t) \\ H_a : \beta_k \neq 0, &\quad P_{\text{value}} = 2Pr(T \geq |t|) \end{aligned}$$

Observe que no rechazar $H_0 : \beta_k = 0$ implica que X_k no tiene ningún importancia para predecir Y.

CI & PI

- La inferencia para la predicción (confidence interval y prediction interval) se construye y se interpreta de forma similar que para el caso de regresión lineal simple.

ANOVA y regresión lineal multiple

- Análogo al caso de regresión lineal simple y ANOVA, la idea es que la variación total se rompe en dos componentes: el atribuido a la regresión (SSR) y el que se atribuye a los residuales.
- La tabla ANOVA en este caso es:

Source	DF	SS	MS
Regression	$\text{DFR} = p$	$\text{SSR} = \sum(\hat{y}_i - \bar{y})^2$	$\text{MSR} = \frac{\text{SSR}}{\text{DFR}}$
Residual (Error)	$\text{DFE} = n - p - 1$	$\text{SSE} = \sum(y_i - \hat{y}_i)^2$	$\text{MSE} = \frac{\text{SSE}}{\text{DFE}}$
Total	$\text{DFT} = n - 1$	$\text{SST} = \sum(y_i - \bar{y})^2$	

$$F = \frac{\text{MSR}}{\text{MSE}}$$

- Donde DF es dregrees of freedom, SS es sum of squares, y MS es mean sum of squares.
- Observe que

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\text{DFT} = \text{DFR} + \text{DFE}$$

ANOVA y regresión lineal multiple

Definition (ANOVA F test)

Plantée el test de hipótesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \beta_i \neq 0 \quad \text{para algún} \quad i$$

La estadística F para ANOVA es:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

El P-value es la probabilidad de que una v.a. con distribución $F(p, n-p-1)$ sea mayor o igual que la estadística F.

F test para una colección de coeficientes de regresión

En un modelo de regresión lineal multiple con p variables independientes podemos plantear el test:

$$H_0 : q \quad \text{variables específicas tienen coeficiente cero}$$
$$H_a : \text{al menos uno de esos coeficientes no es cero}$$

y usar una forma de la estadística F. El P-value es la probabilidad de que una v.a. con distribución $F(q, n-p-1)$ sea mayor o igual que la estadística F.

F test para una colección de coeficientes de regresión

El R^2 en el caso de regresión lineal multiple se puede calcular

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Esta es la proporción de variación en la variable dependiente que se explica por las variables independientes en el modelo de regresión lineal multiple.

F test para una colección de coeficientes de regresión

Si el software no provee la funcionalidad para este test se puede hacer lo siguiente:

1. Haga la regresión para las p variables originales, y llámele al R^2 obtenido R_1^2
2. Haga la regresión para las $p-q$ variables que quedan luego de remover las q variables, y llámele al R^2 resultante R_2^2.
3. La estadística F es

$$F = \left(\frac{n-p-1}{q}\right)\left(\frac{R_1^2 - R_2^2}{1-R_1^2}\right)$$

Ahora, retornemos al ejemplo inicial...