

# Multiple Linear Regression

Aniel Nieves-González

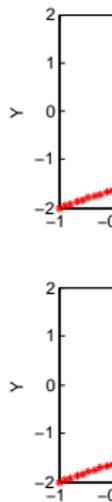
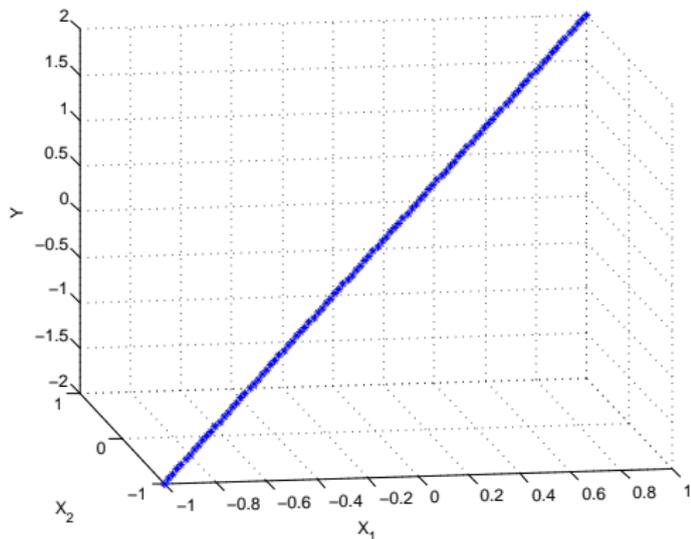
Abril 2013

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. Ver ejemplo 11.5.

# Ejemplo ficticio...



## Definition (Multiple Linear Regression model)

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

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

El **simple linear regression statistical model** es

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ip} + \epsilon_i$$

donde  $i = 1, 2, \dots, n$ ,  $\epsilon_i \sim N(0, \sigma)$  y son independientes. Los parámetros del modelo son  $\beta_0, \beta_1, \dots, \beta_p, \sigma$ . 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 de  $\beta_i$   $\sigma$ . A dichos estimados le llamaremos  $b_i$  y  $s_{\text{reg}}$  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 resuelve para encontrar los parámetros del modelo:

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

donde:

- $\{(x_{11}, \dots, x_{1p}, y_1), \dots, \dots, (x_{n1}, \dots, x_{np}, y_n)\}$  son los datos (note las  $p + 1$  variables  $X_1 \dots, X_p$  y  $Y$ )
- $\mathbf{p}$  son los parámetros (constantes del modelo matemático).
- $F(x_{i1}, \dots, x_{ip}, \mathbf{p})$  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  $X_1 \dots, X_p$  y  $Y$ .
- En LSP se buscan los parámetros del modelo que minimizan la distancia (euclídeana) entre modelo y datos.

## Regression standard error

La desviación estandar del modelo, esto es  $\sigma$ , se estima con el standard error de la regresión:

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

# Intervalo de confianza para los $\beta_i$ 's

## Definition

El IC de confianza de nivel  $C$  para  $\beta_i$  es

$$b_i \pm t^* SE_{b_i}$$

donde  $b_i$  es el valor que se encuentra resolviendo el LSP,  $SE_{b_i}$  es el standard error de  $b_i$ , 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 $\beta_i$ 's

### Definition

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

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

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

$$H_a : \beta_i > 0, \quad P_{\text{value}} = \text{Pr}(T \geq t)$$

$$H_a : \beta_i < 0, \quad P_{\text{value}} = \text{Pr}(T \leq t)$$

$$H_a : \beta_i \neq 0, \quad P_{\text{value}} = 2\text{Pr}(T \geq |t|)$$

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

- 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 múltiple

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

---

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

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

$$SST = SSR + SSE$$

$$DFT = DFR + DFE$$

# ANOVA y regresión lineal múltiple

## Definition (ANOVA F test)

Plantée el test de hipótesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a : \beta_i \neq 0 \quad \text{para algún } i$$

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 múltiple con  $p$  variables independientes podemos plantear el test:

$H_0$  :  $q$  variables específicas tienen coeficiente cero

$H_a$  : 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 múltiple se puede calcular

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

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

## 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

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