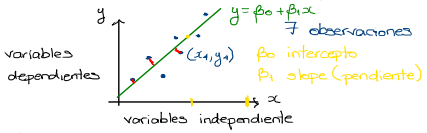
Regresión lineal múltiple

Nota. Los modelos de regresión funcionan tanto para datos numéricos como para datos aleatorios cuantitativos

El caso simple:

1.- ¿Qué tan fuerte es la relación entre dos variables?

2.- Conocer el valor de la variable dependiente de un cierto valor de la variable independiente.



Modelo de regresión simple

$y = \beta_0 + \beta_1 X$ (intercepto, coef. de regresión, slope (pendiente))

variable dependiente — variable independiente

Siempre es posible realizar una regresión lineal simple

Error total del modelo

$e = y - (\beta_0 + \beta_1 X)$

Hipótesis sobre el error: $\mathbb{E}[e] = 0$, $Var[e] = 1$

Meta: A partir de un conjunto dado de datos $\{(x_i, y_i)\}_{i=1}^n$ encontrar

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores para β_0 y β_1 respectivamente

- $\hat{\beta}_0$ y $\hat{\beta}_1$ dependen de $\{(x_i, y_i)\}$
- "Aproximan" a β_0 y a β_1

Hipótesis:

(i) Homogeneidad de la varianza: el tamaño del error en la predicción no cambia (significativamente) conforme los valores de la variable indep cambia

(ii) Independencia entre las observaciones

(iii) Normalidad. Los datos siguen una distribución normal

(X, Y) vector gaussiano:

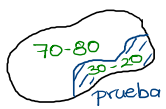
$aX + bY \sim N(\mu, \sigma^2)$ para cualquier a, b que no sean 0 simultáneamente. Por un resultado en prob, existe una v.a. que se llama esperanza condicional

$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$ $\quad \beta_0 = \mathbb{E}[Y - \beta_1 X]$

Muestral $\quad \beta_1 = \frac{Cov(X,Y)}{Var(X)}$

$\beta_0 = \bar{Y} - \beta_1 \bar{X}$

$\beta_1 = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)}$



split
- train
- prueba

Conjunto de datos al cual le estamos ajustando un modelo } Evaluaciones se hacen sobre el conjunto prueba

---

Múltiple: $k < n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad i = 1, \dots, n \quad (1)$$

El sistema de ecs lineales (1), lo podemos reescribir como

$$y = X\beta + e$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad n \times 1$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \quad n \times (k+1)$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad (k+1) \times 1$$

Este vector es el que necesitamos

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Obs. Los coeficientes de la matriz X pueden ser variables aleatorias pero si son variables numéricas el vector e de errores se aproxima (en promedio) a 0

Problema de optimización:

$(x_1, \dots, x_n) = x$
$(y_1, \dots, y_n) = y$

$x^T y$ producto punto

$(A+B)^T = A^T + B^T$
$(AB)^T = B^T A^T$

$$S(\beta) = \sum_{i=1}^{n} e_i^2 = e^T e = (y - X\beta)^T (y - X\beta)$$

$$S(\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta$$

$$S(\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^T y + 2X^T X\beta = 0 \qquad \frac{d\,\beta^T\beta}{d\beta} = 2\beta$$

$$\Rightarrow \boxed{X^T X\beta = X^T y}$$ Ecuaciones normales

Si $X^T X$ es invertible $\Rightarrow \boxed{\hat{\beta} = (X^T X)^{-1} X^T y}$

Coeficiente de determinación

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Medida de qué tan bien se ajusta el modelo regresión

$R^2 \sim 1$ mejor ajuste

$SS_{res} = \sum e_i^2$

$SS_{tot} = \sum (y_i - \bar{y})^2$ numerador que aparece en la varianza