# Métodos de Regresión

Jorge Gallego

Facultad de Economía, Universidad del Rosario

marzo 28 de 2017

# Introducción

- Hasta el momento hemos hecho solo predicciones categóricas
- Pronosticar la clase a la que pertenece un ejemplo
- Pero en muchos casos es fundamental hacer una predicción numérica
- Los métodos de regresión son de lejos el método más utilizado para este propósito
- Veremos un rápido repaso de estas técnicas que de seguro ya dominan

# Fundamentos Básicos

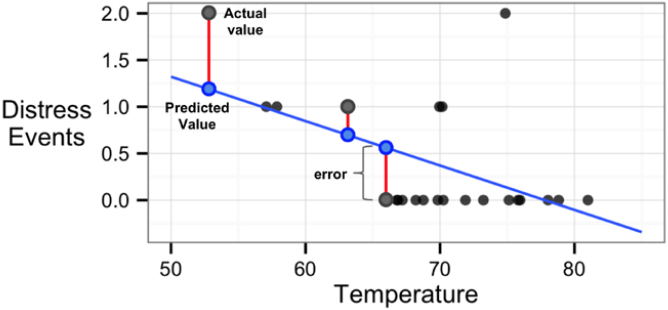
- Por medio de una regresión buscamos especificar la relación entre una **variable dependiente** $y$ y una **independiente** $x$
- Pero el método ha sido usado con enfoques distintos:
  - Examinar cómo poblaciones e individuos *varían* en sus características observables
  - Cuantificar relaciones *causales* entre un evento y la respuesta
  - Identificar patrones para *predecir* el comportamiento futuro dados unos criterios conocidos
- Enfatizaremos en el enfoque predictivo de los métodos de regresión

# Regresión Lineal Simple

- Bajo regresión lineal simple $y$ depende de un único predictor $x$, de forma lineal: $y = \alpha + \beta x + \varepsilon$
- ¿Cómo estimamos $\alpha$ y $\beta$?
- Mínimos Cuadrados Ordinarios (OLS) es el método más usado
- Se busca minimizar la suma de los residuos al cuadrado:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

# Regresión Lineal Simple



# Regresión Lineal Simple

- Puede demostrarse que bajo OLS en regresión simple:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

- Y para el intercepto:

$$\bar{y} = a + b\bar{x}$$

# Correlaciones

- La correlación indica qué tanta asociación (lineal) existe entre dos variables
- El indicador más usado es el coeficiente de correlación de Pearson:

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- El coeficiente está entre -1 y 1. Negativo para correlación negativa, y vice versa para positivo
- Cuánto más se aleje de 0, mayor correlación
- Débil entre 0.1 y 0.3; moderada entre 0.3 y 0.5. Fuerte arriba de 0.5. Similar para negativos

# Regresión Lineal Múltiple

- Es natural extender el enfoque a múltiple predictores:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

- Los coeficientes se estiman con la misma lógica OLS: minimizar la suma de residuos al cuadrado

- Si $\mathbf{X}$ es la matriz de predictores, $\mathbf{y}$ el vector de observaciones de la var. dependiente, $\beta$ el de coeficientes y $\varepsilon$ el de errores:

$$\mathbf{y} = \beta\mathbf{x} + \varepsilon$$

- Este es el modelo en forma matricial

# Regresión Lineal Múltiple

- Con un poco de álgebra lineal puede demostrarse que:

$$\hat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

- donde $\mathbf{X}^\mathsf{T}$ es la transpuesta de $\mathbf{X}$
- De esta forma, podemos estimar a partir de los datos los coeficientes del modelo
- Y con él, podemos hacer las predicciones de interés: $\hat{\mathbf{y}} = \hat{\beta}\mathbf{x}$
- Así, el modelo de regresión múltiple es un algoritmo más de *machine learning*

# Regresión Lineal Múltiple

Las principales ventajas del modelo son:

1. El método más popular para modelar datos numéricos
2. Se puede adaptar prácticamente para cualquier tarea
3. Genera estimación tanto de la fortaleza como del tamaño de la relación entre predictores y *outcome*

# Regresión Lineal Múltiple

Las principales desventajas son:

1. Supuestos fuertes sobre los datos
2. La especificación del modelo debe ser hecha ex ante
3. No tiene en cuenta los datos ausentes