# Métodos de Regresión

Jorge Gallego

Facultad de Economía, Universidad del Rosario

Junio 24 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
- Haremos un rápido repaso de estas técnicas que de seguro ya dominan

# Fundamentos Básicos

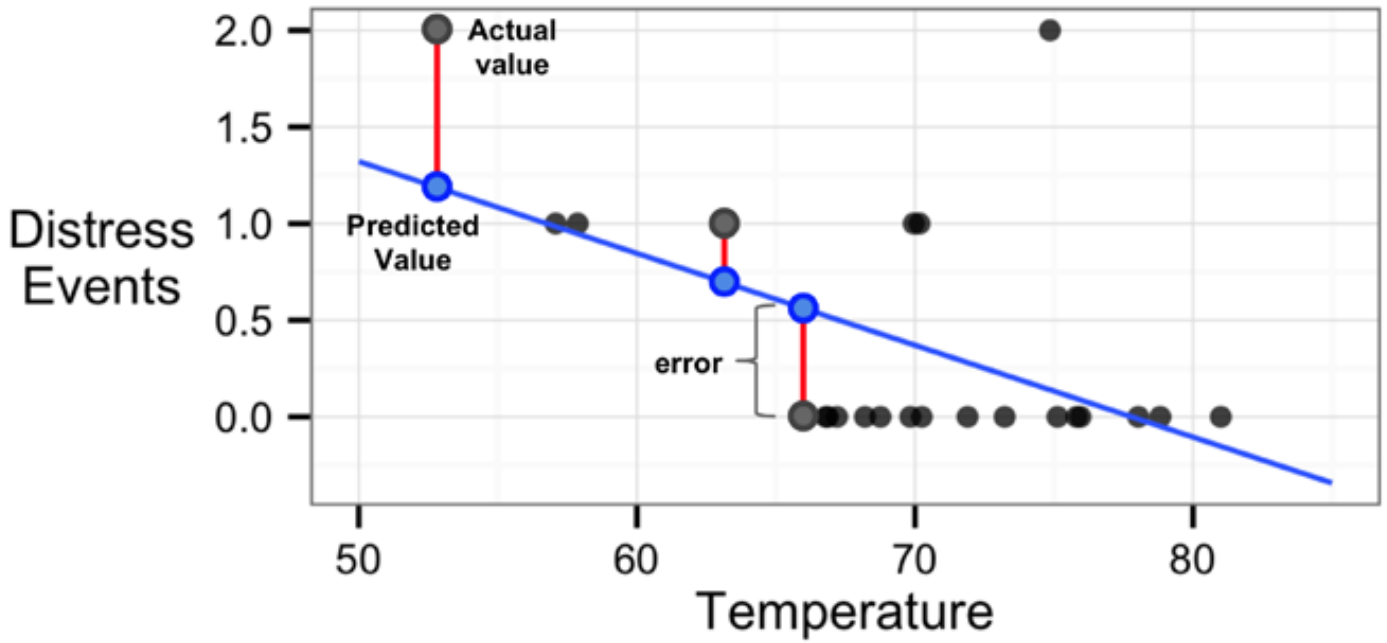
- Por medio de una regresión buscamos especificar la relación entre una **variable dependiente** $y$ y una **independiente** $x$
- Pero el método ha sido usado con enfoques distintos:
  - Examinar cómo poblaciones e individuos *varían* en sus características observables
  - Cuantificar relaciones *causales* 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:

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

- donde $\mathbf{X^T}$ 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

# Variable Dependiente Dicotómica

- El modelo clásico por lo general asume variable dependiente continua
- En predicción categórica, la variable dependiente no es contínua
- Pero, ¿qué pasa si la variable dependiente es dummy?
- Tres modelos: uno simple pero limitado; dos complejos pero más precisos

# Modelo de la probabilidad lineal

Sea $y$ una variable dummy:

$$y = \begin{cases} 1 & \text{si ocurre A} \\ 0 & \text{si no ocurre A} \end{cases}$$

Es posible modelar:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

# Problemas del MPL

Pero el modelo no es perfecto

- El MPL tiene tres grandes problemas:
- 1. Nada garantiza que $0 < P(\mathbf{x}) < 1$.

  ¡Podría haber probabilidades negativas!
- 2. El efecto marginal de $x_j$ sobre $P(\mathbf{x})$ es constante.
- 3. Se viola homocedasticidad

# Modelos No-Lineales

- Tres problemas del MPL: Probabilidades sin sentido, efectos marginales constantes y heterocedasticidad
- El problema 3 no es tan grave: se puede corregir con errores estándar robustos
- Los modelos no-lineales sirven para corregir los problemas 1 y 2
- Los modelos Logit y Probit son los más populares

# Modelos No-Lineales

Nuestra variable de interés:

$$P(y = 1|\mathbf{x})$$

que es la probabilidad de éxito dado $\mathbf{x}$. Consideremos modelos de la forma:

$$P(y = 1|\mathbf{x}) = G(\beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = G(\mathbf{x}\beta)$$

donde $G$ es tal que

$$0 < G(z) < 1$$

para todo $z$.

# Modelo Logit

## Definición: Modelo Logit

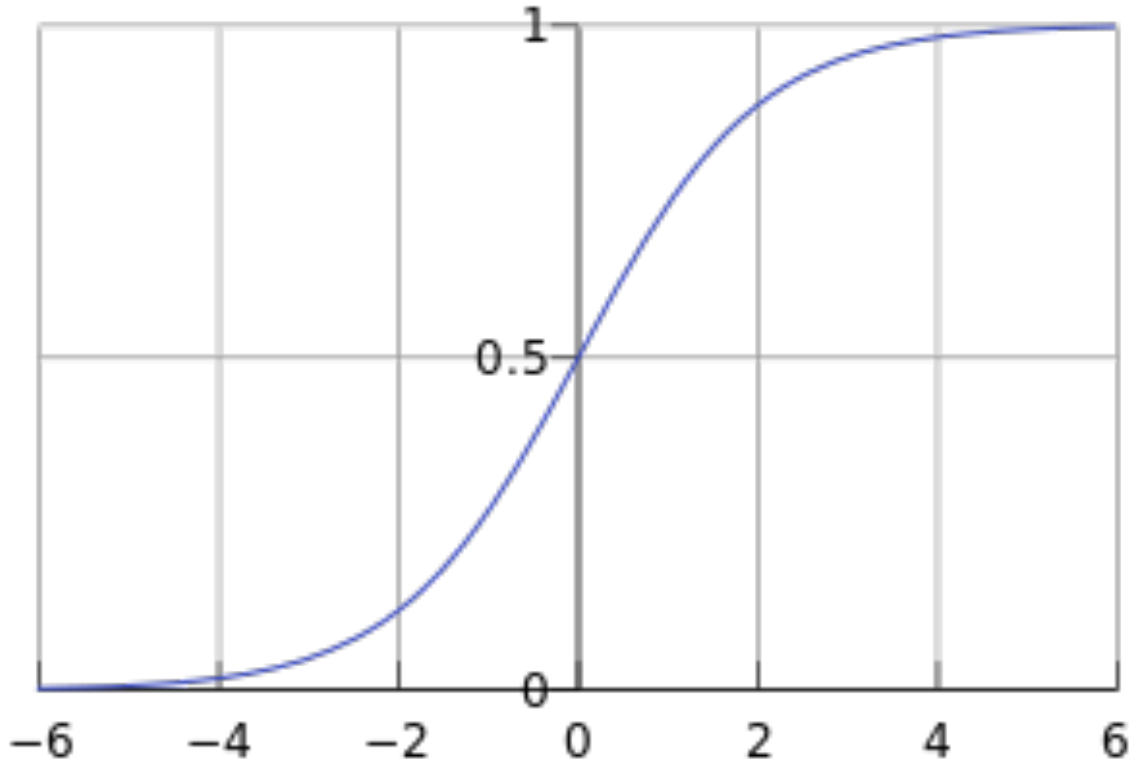
En el modelo logit, la función $G$ es la función logística:

$$G(z) = \Lambda(z) = \frac{exp(z)}{1 + exp(z)}$$

Para esta función, $0 < \Lambda(z) < 1$, para todo $z$.

# Función Logística

Figure: Función Logística



# Modelo Probit

## Definición: Modelo Probit

En el modelo Probit, la función $G$ es la función de distribución acumulada de la normal estándar:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$$

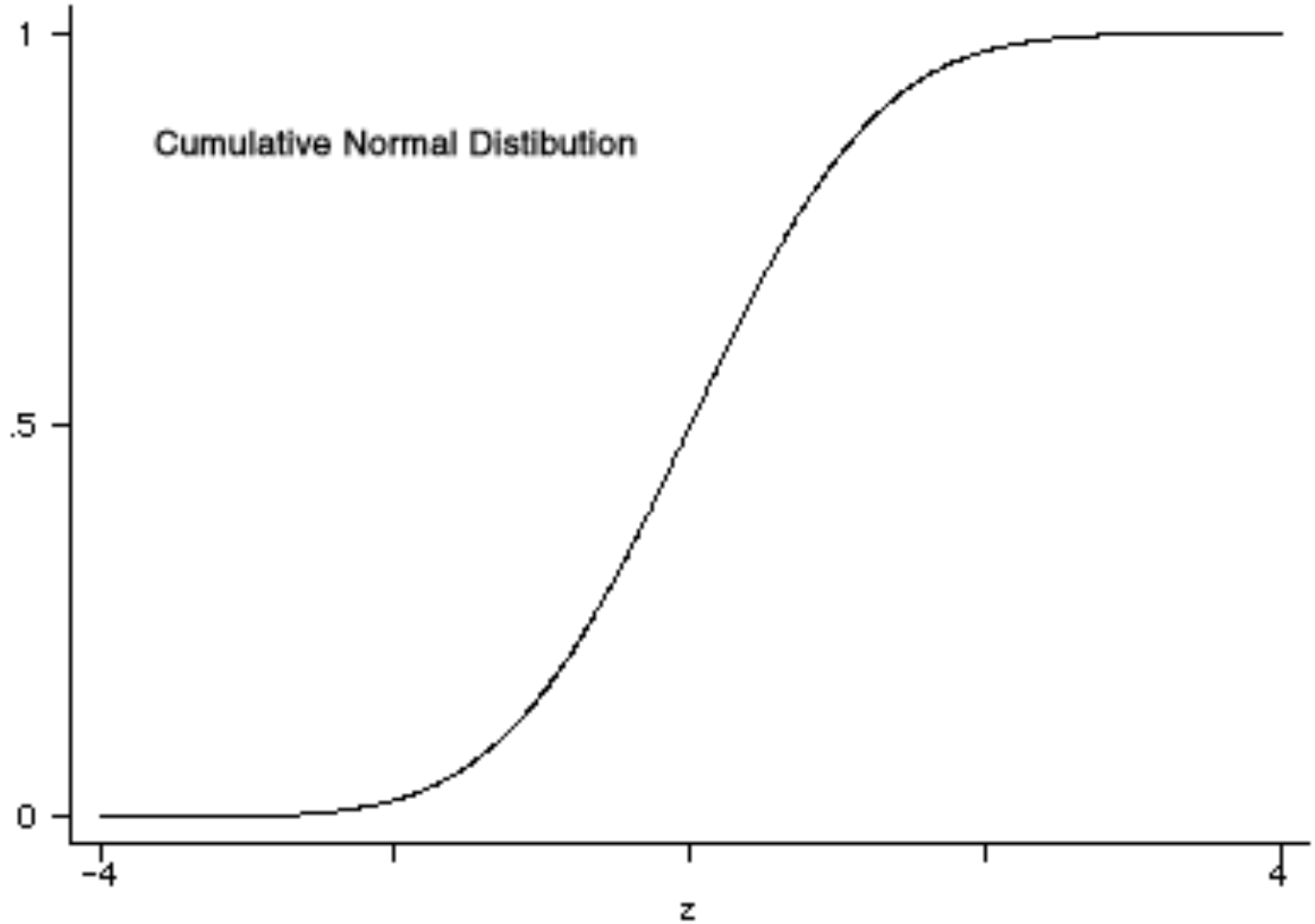
donde $\phi(z)$ es la densidad de la normal estándar:

$$\phi(z) = (2\pi)^{-1/2} exp(-z^2/2)$$

Nuevamente, dada esta función $0 < \Phi(z) < 1$ para todo $z$

# Distribución Acumulada Normal Estándar

Figure: Distribución Acumulada Normal Estándar



# Estimación de los modelos

¿Cómo estimamos los coeficientes $\hat{\beta}$ para estos modelos?

- En el MPL, se usa MCO
- En estos modelos no-lineales, no se puede usar MCO
- Se usa la estimación por máxima verosimilitud (EMV)

## Intuición

Se escogen los parámetros $\hat{\beta}$ para maximizar la probabilidad de pronosticar con el modelo las observaciones que se tienen. Es decir, que si $y_i = 1$ para algún $i$, entonces la idea es que el modelo pronostique una probabilidad cercana a 1 para $i$.

# Bondad de Ajuste

1. Porcentaje de Pronósticos Correctos (PPC)
   - Se estima $P(\mathbf{x}_i)$ para todo $i$.
   - Si $P(\mathbf{x}_i) > .5$, se asume $\hat{y}_i = 1$. En caso contrario $\hat{y}_i = 0$.
   - Se calcula la proporción de aciertos

2. Pseudo $R^2$
   - Se define como $1 - \frac{L_{irr}}{L_0}$
   - Donde $L_{irr}$ es la log-verosimilitud del modelo completo, $L_0$ la del modelo con sólo el intercepto.
   - Intuición: si la verosimilitud (entre 0 y 1) es alta, su log-verosimilitud es baja en valor absoluto. Por eso la fracción está invertida