# Regresión Lineal Simple: Inferencia Estadística

## Clase 16

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

**Probabilidad y Estadística II**
**Departamento de Ingeniería Industrial**
**Universidad de Los Andes, Bogotá, Colombia**

2020-20

# Outline

# Motivación

En un modelo de regresión lineal, se asume que la función de regresión es lineal en las $X$'s:

$$E(Y|X = x) = f(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Y además por conveniencia pondremos el supuesto de que $Y$ tiene distribución normal:

$$Y \sim Normal\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \ , \ \sigma^2\right)$$

donde la varianza no depende de $X$. Esto se puede reescribir como:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$
$$\epsilon \sim Normal\left(0 \ , \ \sigma^2\right)$$

# Motivación

## Que se logra con un modelo de regresión?

1. Explicar: Se pueden responder preguntas científicas (parecidas al diseño de experimentos) sobre la influencia de las variables independientes sobre la respuesta:
   - Aumenta el salario de un egresado obtener un mejor promedio académico?
   - Vale la pena estudiar más para el parcial?
2. Predecir: Se puede predecir el comportamiento de la variable de respuesta si se fijan los niveles de $X$:
   - Si mi promedio es 3.99, qué puedo esperar de mi salario?
   - Qué pasará si estudio sólamente 4 horas para el próximo parcial?
3. Precisión en Inferencia: Parecido al diseño de experimentos, al tener en cuenta el efecto de las $X$'s, se reduce la varianza, con lo cual la inferencia acerca de $Y$ es más precisa (con más potencia).
   - Al hacer un prueba para saber si mi salario será mayor a \$2 millones, esta será mucho más exacta a respuesta si conozco mi promedio.

# Outline

# Planteamiento

Vamos a concentrarnos hoy en el caso en que solo se tiene una variable explicativa. Este caso se denomina el modelo de regresión lineal simple:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Donde $\epsilon \sim_{iid} N(0, \sigma^2)$. Bajo estos supuestos se considera que:

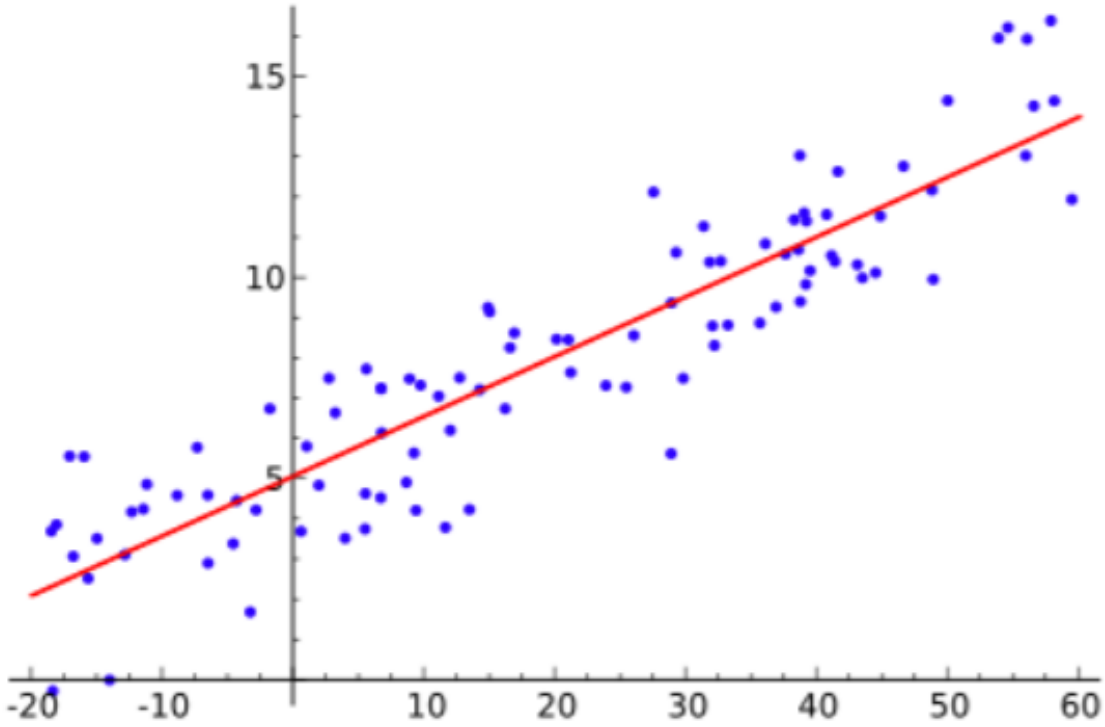
$$Y|X_1 \sim Normal\left(\beta_0 + \beta_1 X_1\ ,\ \sigma^2\right)$$

Es decir, la media de Y dado X es $E(Y|X) = \beta_0 = \beta_1 X_1$, luego:

- $\beta_0$ es el valor esperado de Y dado un valor nulo de la variable X
- $\beta_1$ es el aumento en Y,dado un aumento de 1 en X

# Planteamiento

Gráficamente se tendría la siguiente situación:




El error aleatorio simplemente describe las desviaciones sobre la recta que no se pueden explicar por medio de X.

# Outline

# Estimación

En la práctica, sucede que el modelo es desconocido, esto es, no tenemos conocimiento de los coeficientes $\beta_0$ y $\beta_1$, por tanto debemos estimarlos.

Dada una muestra de $n$ datos, que en este caso corresponden a pares de observaciones de la variable de interés $Y$, y la variable explicativa $X_1$:

| i | $Y$ | $X$ |
|---|---|---|
| 1 | $Y_1$ | $X_1$ |
| 2 | $Y_2$ | $X_2$ |
| ⋮ | ⋮ | ⋮ |
| i | $Y_i$ | $X_i$ |
| ⋮ | ⋮ | ⋮ |
| n | $Y_n$ | $X_n$ |

Nuestro objetivo es intentar recrear la relación entre ellas

# Estimación

## Como podemos hacer el ajuste?

La recta de regresión ajustada se puede denotar como $\hat{Y} = \hat{\beta}_0 = \hat{\beta}_1 X_1$, luego podemos definir el error de estimación como $e_i = Y_i - \hat{Y}_i$. La idea es que este error sea pequeño, por tanto se puede minimizar:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2$$

La suma simplemente agrega el error de todas las observaciones, y el cuadrado hace desaparecer el signo para así minimimizar la magnitud del error. De ahí viene el nombre de estimación por mínimos cuadrados ordinarios.

# Estimación

Como hallamos el mínimo? Derivar, igualar a 0 y despejar.

## Las ecuaciones normales

Derivando con respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$ e igualando a 0, se tienen las llamadas ecuaciones normales:

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1) = 0$$

# Estimación

Resolviendo el sistema de ecuaciones normales, se obtienen las siguientes expresiones para los estimadores:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

# Estimación

Haciendo álgebra es fácil mostrar las siguientes igualdades:

$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = \sum_{i=1}^{n}(X_i - \bar{X})Y_i = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})X_i = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

- La primera es para facilidad de cálculos.
- La segunda servirá para mostrar propiedades del estimador más fácilmente.
- La tercera es para efectos de interpretación.

# Outline

# Propiedades de los estimadores

Para hallar las propiedades de los estimadores, escribimos a $\hat{\beta}_1$ como:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

Si definimos $w_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ podemos reescribir el estimador como $\hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i$, con lo cual resulta fácil hallar el valor esperado y la varianza del mismo:

$$E\left(\hat{\beta}_1\right) = \sum_{i=1}^{n} w_i E(Y_i) \qquad Var\left(\hat{\beta}_1\right) = \sum_{i=1}^{n} w_i^2 Var(Y_i)$$

# Propiedades de los Estimadores

Realizando los procedimientos, se tiene que $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen una distribución Normal con párametros:

$$E\left(\hat{\beta}_1\right) = \beta_1 \qquad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}}$$

y

$$E\left(\hat{\beta}_0\right) = \beta_0 \qquad Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)$$

Además,

$$Cov(\hat{\beta}_1, \hat{\beta}_0) = \frac{-\sigma^2\bar{X}}{S_{xx}}$$

Se puede mostrar que los estimadores son insesgados y consistentes.

# Estimación de la Varianza

Las expresiones anteriores involucran a $\sigma^2$, luego es necesario estimarlo. Para este punto ya sabemos como:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{gl_E}$$

Y a vimos que el SSE se define como:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1)^2$$

Donde los graos de libertad del error, son $gl_E = n - 2$ .

# Pruebas de Hipótesis

Una de las preguntas más importantes que queremos contestar, es si la variable X es o no relevante para explicar el comportamiento de Y. En nuestra notación, esto es equivalente a preguntar si:

$$H_0: \quad \text{La Variable NO es significativa} \Leftrightarrow \beta_1 = 0$$
$$H_1: \quad \text{La Variable SI es significativa} \Leftrightarrow \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim t_{glE}$$

# Outline

# Pruebas de Hipótesis

Otra pregunta de interés, está asociado a si se desea conocer si el valor medio de Y toma un valor particular, dado cierto valor de X. El mejor estimador de $E(Y|X)$ está dado por la recta de regresión:

$$\hat{E}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$Var(\hat{E}(Y|X)) = Var(\hat{\beta}_0) + X^2 Var(\hat{\beta}_1) + 2XCov(\hat{\beta}_0, \hat{\beta}_1)$$

luego

$$H_0 : E(Y|X) = \theta_0$$

$$H_1 : E(Y|X) \neq \theta_0$$

$$t = \frac{\hat{E}(Y|X) - \theta_0}{\sqrt{\hat{Var}(\hat{E}(Y|X)}} \sim t_{glE}$$

# Outline

# Intervalos de Confianza

## Intervalo de confianza para la media de $Y$ dado X

$$IC_{1-\alpha}\left(E(Y|X)\right) = \hat{E}(Y|X) \pm t_{gl_E;1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{E}(Y|X))}$$

Este intervalo nos da un margen de error para nuestra estimación de la media de $Y$ dado $X$.

## Intervalos de predicción para $Y$

$$IP_{1-\alpha}\left(Y|X\right) = \hat{E}(Y|X) \pm t_{glE;1-\frac{\alpha}{2}} \sqrt{MSE + \hat{Var}(\hat{E}(Y|X))}$$

Este intervalo nos da unos límites en donde se encontrará el valor que efectivamente observemos de Y dado X.

# Intervalos de Confianza

## Intervalo de confianza para la media de $Y$ dado $X$

$$IC_{1-\alpha}(\mu_p) = \hat{\mu}_p \pm t_{(n-2);1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x_p - \bar{X})^2}{S_{xx}} \right)}$$

## Intervalos de predicción para $Y$ dado $X$

$$IP_{1-\alpha}(Y_p) = \hat{\mu}_p \pm t_{(n-2);1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{X})^2}{S_{xx}} \right)}$$

El significado de este intervalo es completamente diferente, dado que no se infiere sobre el valor de un parámetro (constante).

# Ejemplo

En el contexto de valoración de un activo financiero es de vital importancia determinar la tasa de descuento apropiada para encontrar el valor presente de los flujos futuros de efectivo que genera dicho activo. Una de las metodologías más populares para cumplir con este objetivo es el Capital Asset Pricing Model (CAPM) que calcula la tasa de descuento a partir del "beta del activo" $\beta_A$ utilizando la siguiente expresión:

$$E[r_A] = r_f + \beta_A E[RPM]$$

Donde $r_A$ es el retorno del activo, $r_f$ es la tasa libre de riesgo y $RPM$ es el *risk premium* del mercado. En este modelo la tasa de descuento buscada corresponde al valor medio del retorno del activo $E[r_A]$. En este orden de ideas, para estimar el parámetro $\beta_A$ se plante el siguiente modelo de regresión lineal:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Usted tiene información histórica del año bursátil de 2015 acerca del retorno diario de la empresa Apple ($y$) y el retorno diario correspondiente al *risk premium* ($x$).