

# Regresión múltiple y otras técnicas multivariadas

Javier Santibáñez

IIMAS, UNAM

`jsantibanez@sigma.iimas.unam.mx`

Semestre 2018-2

- ① Introducción
- ② Regresión lineal simple
- ③ Repaso de álgebra lineal y probabilidad
- ④ Regresión lineal múltiple
- ⑤ Diagnósticos del modelo y medidas correctivas
- ⑥ Selección de modelos
- ⑦ Tópicos adicionales

## Actividades a evaluar

- Tareas semanales en equipo (14-15) 30 %
- Exámenes parciales individuales (2) 60 %
- Proyecto final en equipo 10 %

## Fechas importantes

- Primer parcial: 22 de marzo
- Segundo parcial: 31 de mayo
- Proyecto final: 7 de junio

## Página del curso

[http://sigma.iimas.unam.mx/jsantibanez/Cursos/Regresion/2018\\_2/](http://sigma.iimas.unam.mx/jsantibanez/Cursos/Regresion/2018_2/)

# El esquema frecuentista

- Se tiene una población  $\mathcal{U} = \{u_1, u_2, \dots\}$  con un número infinito de elementos e interesa estudiar la distribución de frecuencias de una característica de los elementos de  $\mathcal{U}$ .
- Si  $Y$  representa la cuantificación de la característica de interés, se considera que las mediciones de los elementos son  $Y_1 = Y(u_1), Y_2 = Y(u_2), \dots$
- El interés entonces es dar una descripción de la distribución de frecuencias de las mediciones  $Y_1, Y_2, \dots$ , para ello se utilizan modelos de probabilidad.
- El objetivo bajo este esquema es hacer inferencias sobre el modelo de probabilidad  $F$  que mejor describe la distribución de frecuencias de  $Y_1, Y_2, \dots$ , a partir de observar  $Y$  en una muestra de elementos de la población.

# El esquema frecuentista paramétrico

- Se pueden reducir las opciones para determinar el mejor modelo si se asume que éste tiene una forma conocida salvo por algunas constantes desconocidas, denominadas parámetros.
- Por ejemplo, si se asume que el modelo es normal, basta con hacer inferencias sobre  $\mu$  y  $\sigma$ ; si se asume que el modelo es exponencial, basta con hacer inferencias sobre  $\lambda$ .
- La crítica a este esquema radica en este punto, ya que asumir que el modelo tiene una forma conocida limita el alcance de las conclusiones obtenidas.
- En el esquema frecuentista paramétrico se modela como  $Y \sim F(y|\theta)$ , donde  $\theta$  es el parámetro del modelo (que puede ser un vector) y la búsqueda se restringe a un conjunto de posibles valores  $\Theta \subset \mathbb{R}^p$  conocido como espacio parametral.

# Los modelos de regresión

- En algunos casos interesa dar una descripción de  $Y$  condicional a un conjunto de variables auxiliares  $\mathbf{X} = (X_1, \dots, X_k)$ ,  $k \in \mathbb{N}$ .
- En este caso  $\mathbf{X}_1 = \mathbf{X}(u_1)$ ,  $\mathbf{X}_2 = \mathbf{X}(u_2)$ ,  $\dots$ , son los vectores de mediciones de las variables auxiliares de los elementos de la población.
- En el esquema frecuentista paramétrico se puede representar lo anterior como sigue

$$Y | \mathbf{X} = \mathbf{x} \sim F(y | \mathbf{x}, \theta), \quad \theta \in \Theta \subset \mathbb{R}^p,$$

donde  $\mathbf{X} = (X_1, \dots, X_k)$  es el vector de variables auxiliares y  $\mathbf{x} = (x_1, \dots, x_k)$  es una realización particular de  $\mathbf{X}$ .

- A partir de hacer inferencias en un modelo como el anterior es posible determinar si existe algún tipo de asociación entre  $Y$  y  $\mathbf{X}$ .

# Ejemplo

- En el siglo XIX, Francis Galton estudió la relación que había entre la estatura de padres e hijos adultos.
- Galton notó que los hijos de padres altos tienden a no ser tan altos como sus padres, mientras que los hijos de padres bajos tienen a no ser tan bajos como sus padres y llamo a este fenómeno regresión a la media.
- En el paquete `HistData` de R se encuentra el conjunto de datos llamado **Galton**, que contiene la información de 928 hijos de 205 parejas. La estatura de los padres reportada corresponde al promedio de las estaturas del padre y de la madre.

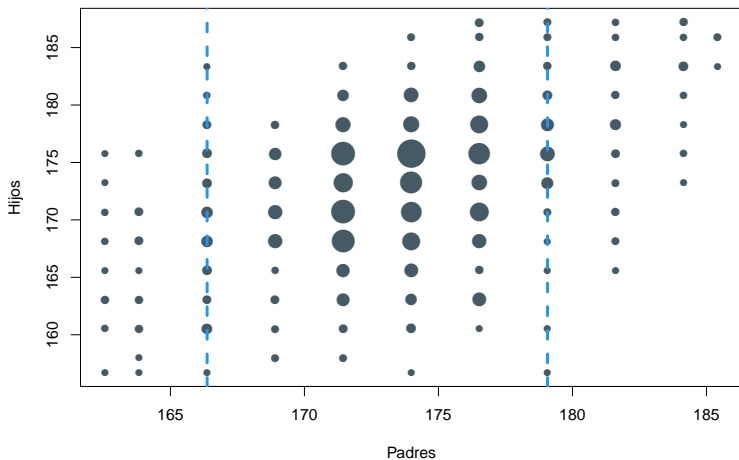


Figura: Gráfico de dispersión de estaturas de padres contra estatura de hijos.

¿Cómo se pueden corroborar las afirmaciones de Galton a partir de los datos?



- El modelo de regresión es un modelo lineal. Se asume que la distribución de  $Y$  está caracterizada por su media y varianza.
- La primera parte del modelo describe la relación que hay entre el valor esperado de  $Y$  y las variables auxiliares.

$$E(Y | \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

- La segunda parte del modelo describe la la varianza de  $Y$ . En caso más sencillo se asume que la forma en que  $Y$  varía alrededor de su media no depende de  $\mathbf{X}$ , esto es

$$V(Y | \mathbf{X} = \mathbf{x}) = \sigma^2.$$

# Regresión lineal simple

- Cuando  $p = 1$  el modelo se nombra *simple*. Cuando  $p \geq 2$  el modelo se llama *múltiple*.
- En el ejemplo de Galton es posible utilizar el modelo de *regresión lineal simple* para explicar las estaturas de los hijos a partir de las estaturas de los padres.

$$E(Y | X = x) = \beta_0 + \beta_1 x$$

donde  $Y$  es la altura del hijos,  $X$  es la altura de los padres,  $\beta_0$  y  $\beta_1$  son los parámetros del modelo.

- Se puede notar que la estatura de los hijos ( $Y$ ) no está completamente determinada por la estatura de los padres ( $X$ ).

# Aplicaciones de los modelos de regresión

- ① Las ventas de un producto pueden ser predichas a partir del gasto en publicidad.
- ② El desempeño de un trabajador en un empleo puede ser predicho a partir de las respuestas de una prueba de aptitudes.
- ③ El tamaño del vocabulario de un niño puede ser predicho a partir de la edad del niño y el grado de escolaridad de sus padres.
- ④ El salario de un trabajador puede ser predicho a partir de su edad, escolaridad y sector de ocupación.
- ⑤ El volumen de madera de un árbol puede ser determinado a partir de variables como el diámetro del tronco y la altura.

*Essentially, all models are wrong, but some are useful.*

George Box

