Regresión Lineal Simple: Introducción y Estimación

Clase 15

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

1. Introducción
2. Regresión Lineal Simple
   - Planteamiento
   - Estimacion
3. Ejemplo

Recordemos

En nuestra clase 1 dijimos que ... En este curso nos enfocaremos en desarrollar métodos para entender que fenómenos externos influyen en el valor que toma la media μ, así como cuantificar su efecto.

La idea de Proba. II

Determinar como la media de nuestra variable de interés Y cambia ante factores externos $X_1, ..., X_k$ según una relación de la forma:

$$\mu = f(X_1, ..., X_k)$$

Si los factores X_j son categóricos, estamos en el contexto del diseño de experimentos. Si son variables continuas, estamos en el contexto de regresión.

Motivación

Hasta este punto que hemos hecho?

Ya sabemos como μ se altera cuando tenemos factores categóricos (i.e estructura de grupos), por medio del diseño de experimentos, pero no todos los factores vienen en categorías:

1. La venta mensual se sombrillas y/o paraguas depende del nivel de pluviosidad. Al mirar los datos históricos, se puede ver que hay dependencia directa.
2. La nota que un estudiante obtiene en el parcial con relación a las horas de estudio dedicadas.
3. El salario de un estudiante recién graduado depende del promedio acumulado (GPA) obtenido durante la carrera (X_1) y del nivel de ingreso familiar (X_2).

Ahora queremos explicar nuestra variable de interés Y por medio de variables continuas, es decir variables cuya naturaleza ya no son valores puntuales, sino que pueden tomar todo un infinito de valores.

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

1 Introducción

2 Regresión Lineal Simple
- Planteamiento
- Estimacion

3 Ejemplo

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:

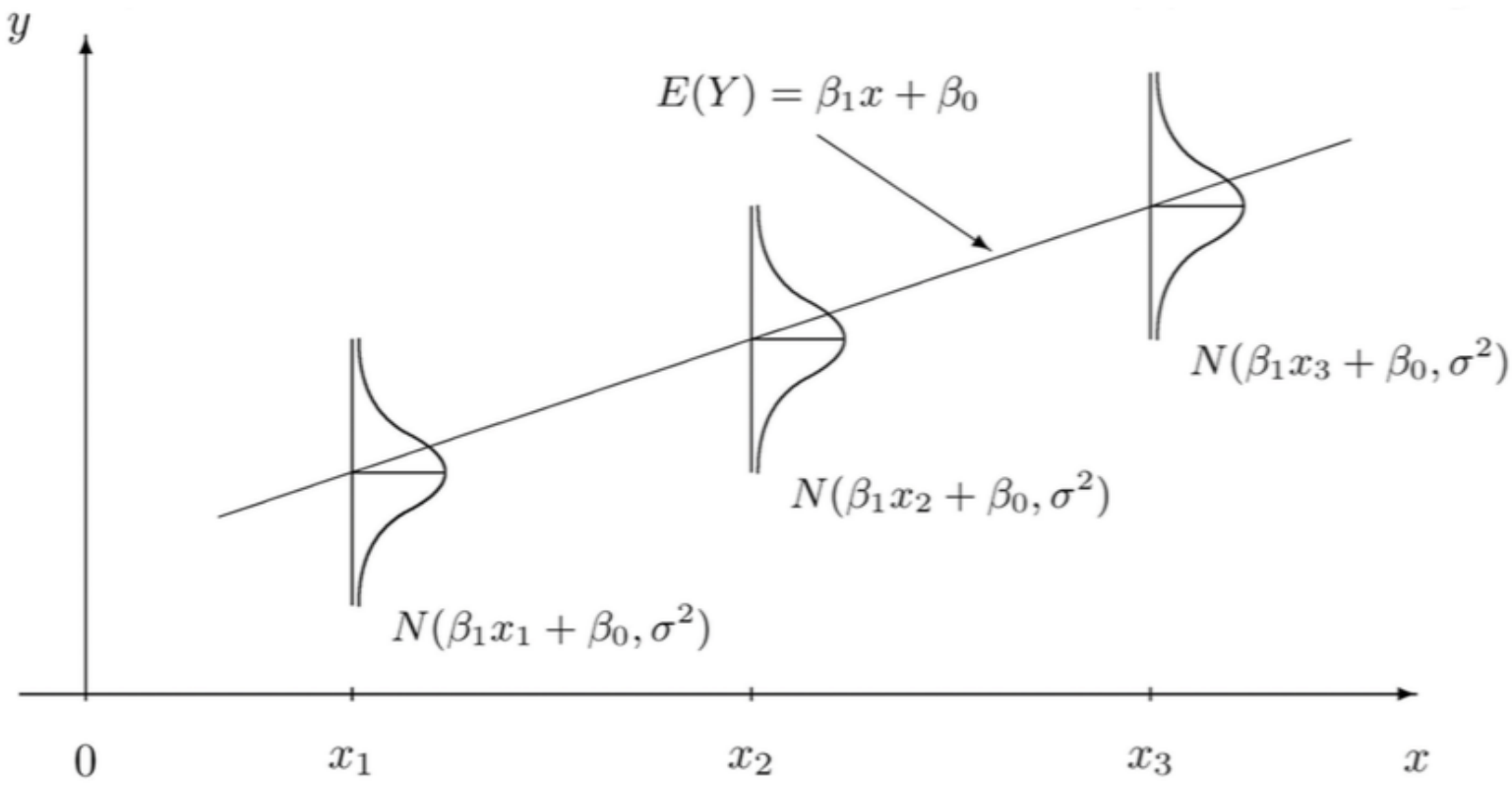
$$Y|X_1 \sim Normal\left(\beta_0 + \beta_1 X_1\ ,\ \sigma^2\right)$$

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

- β_0 es el valor esperado de Y dado un valor nulo de la variable X
- β_1 es el aumento en Y,dado un aumento de 1 en X

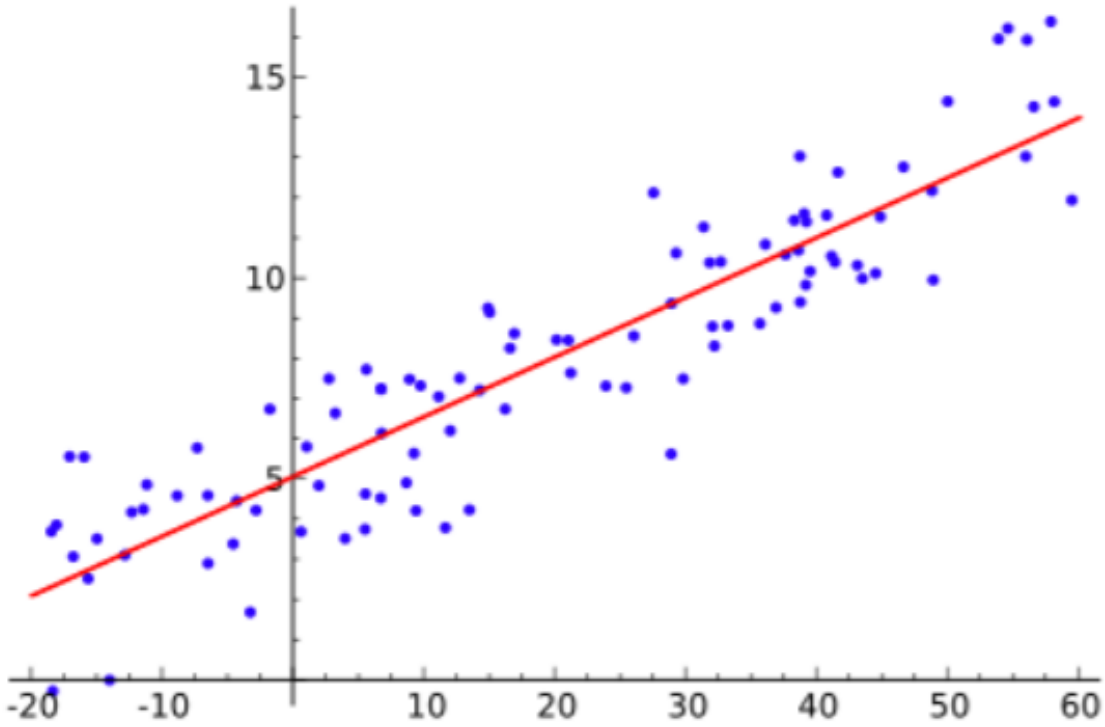
Planteamiento

Gráficamente se tendría la siguiente situación:



Planteamiento

Gráficamente se tendría la siguiente situación:



Outline

1. Introducción
2. Regresión Lineal Simple
   - Planteamiento
   - Estimacion
3. Ejemplo

Estimación

En la práctica, sucede que el modelo es desconocido, esto es, no tenemos conocimiento de los coeficientes β_0 y β_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:

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

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

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.

Estimación

Con lo anterior se puede escribir el estimador de la pendiente como:

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

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$$

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

Es decir, la pendiente está relacionada con el nivel de asociación dado por la covarianza entre las variables.

Outline

1. Introducción
2. Regresión Lineal Simple
   - Planteamiento
   - Estimacion
3. Ejemplo

Ejemplo

Una compañia dedicada a la fabricación de computadores quiere evaluar si las ganancias mensuales en millones de dólares están afectadas por un gasto mensual en publicidad. Para esto, se tienen los siguientes datos:

Y: Ingresos mensuales (millones)	65	89	46	34	76	96	24	35	88	90
X: Gasto en publicidad (millones)	3	13	6	2	6	12	2	3	11	13

Encuentre la recta de regresión que relaciona los ingresos mensuales con el gasto en publicidad