Probabilidad y Estadística

Clase 9

Gonzalo Blanco

2023

Regresión lineal por mínimos cuadrados

1. Introducción

Puede existir una relación entre dos variables x e y, y que nos interese estudiar la naturaleza de la misma. La relación matemática determinística (es decir, sin componente aleatoria) más simple entre estas dos variables es **la relación lineal dada por** $y = \beta_0 + \beta_1 x$. La misma determina una línea recta con pendiente β_1 y ordenada al origen β_0 para los pares (x,y).

Si las dos variables **no están determinísticamente relacionadas**, entonces para un valor fijo de x, el valor de y será aleatorio. Más generalmente, la variable cuyo valor fija el experimentador será denotada por x y se llamará **variable independiente, pronosticadora o variable explicativa**. Con x fija, la segunda variable será aleatoria; esta variable aleatoria y su valor observado se designan Y e y, respectivamente, y se la conoce como **variable dependiente o de respuesta**.

Normalmente se realizarán varias observaciones para varios escenarios (o valores) de la variable independiente x. Sean $x_1, x_2, ..., x_n$ los valores de la variable independiente para la que se realizan las observaciones y sean Y_i e y_i, respectivamente, la variable aleatoria y el valor observado de la misma asociado a x_i. Los datos disponibles se componen entonces de pares (x_i, y_i) a partir de los cuales querremos encontrar la vinculación (en este caso lineal) que existe entre las variables x e Y.

Un primer paso para hacer una análisis de regresión lineal entre dos variables es hacer una gráfica de estos pares de puntos para cerciorarnos, al menos a primera instancia, de que la relación entre ambas es lineal o tiende a serlo (Fig. 1).

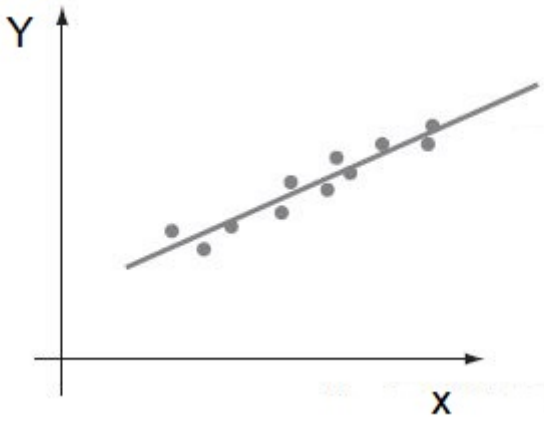

Figura 1: Gráfico de los pares de puntos (x,y) y una recta que representa su tendencia lineal.

2. Modelo Probabilístico Lineal

Para el modelo determinístico $y = \beta_0 + \beta_1 x$ el valor observado de y es una función lineal de x, es decir, determinamos el valor exacto de y, y este no tiene ningúna componente aleatoria. La generalización apropiada de esto a un modelo probabilístico supone que *el valor esperado de Y es una función lineal de x*, pero que con x fija, la variable aleatoria Y difiere de su valor esperado en una cantidad aleatoria.

Existen parámetros β_0, β_1 y σ^2 tal que para cualquier valor fijo de la variable independiente x, la variable dependiente Y está relacionada con x por el modelo

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

La cantidad ϵ en la ecuación del modelo es una variable aleatoria, que se supone está normalmente distribuida con $E(\epsilon) = 0$ y $V(\epsilon) = \sigma^2$.

La variable aleatoria ϵ se conoce como **término de error aleatorio o desviación aleatoria** del modelo. Sin ϵ, estaríamos diciendo que Y se vinculará con x de forma perfectamente lineal, es decir que cualquier valor de Y puede ser obtenido de forma exacta a partir del modelo $y = \beta_0 + \beta_1 x$. A esta recta la llamamos *recta de regresión verdadera*. La inclusión del error aleatorio ϵ provoca que los pares (y_i, x_i) puedan estar por encima o por debajo de esta recta. Los pares $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ se dispersarán alrededor de de la línea de regresión verdadera como se observa en la Figura 2.

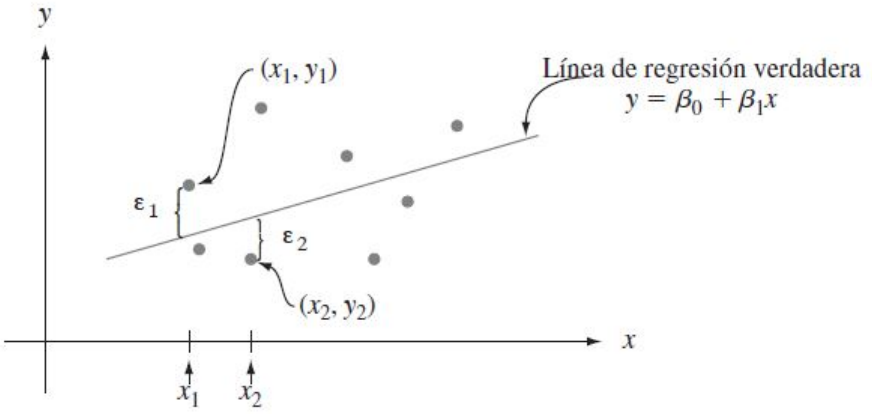

Figura 2: Pares de datos (x_i, y_i) y la recta de regresión verdadera.

Para aclarar un poco posibles dudas, el modelo planteado en (1) se entiende mejor con la ayuda de la siguiente notación. Sea x_0 un valor particular de la variable independiente x

$$\mu_{Y,x_0} = \textit{ el valor esperado de } Y \textit{ cuando } x = x_0$$

.

$$\sigma^2_{Y,x_0} = \textit{ la varianza de } Y \textit{ cuando } x = x_0$$

.

La notación alternativa sería, a partir de probabilidad condicional, $E[Y/x_0] = \mu_{Y,x_0}$ y $V(Y/x_0) = \sigma^2_{Y,x_0}$.

Entonces si tenemos un conjunto de pares observados (x_i, y_i) y obtenemos un modelo probabilístico lineal para x e Y, entonces para un dado valor x_0 el valor esperado de Y es $E[Y/x_0] = \mu_{Y,x_0} = \beta_0 + \beta_1 x_0$.

$$E[Y/x_0] = \mu_{Y,x_0} = E[\beta_0 + \beta_1 x + \epsilon] = \beta_0 + \beta_1 x_0 + E[\epsilon] = \beta_0 + \beta_1 x_0$$
$$V(Y/x_0) = \sigma^2_{Y,x_0} = V(\beta_0 + \beta_1 x_0 + \epsilon) = V(\beta_0 + \beta_1 x_0) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Por lo tanto Y es la suma de una constante $\beta_0 + \beta_1 x$ y una variable aleatoria ϵ normalmente distribuida, así que por si misma Y tiene una distribución normal. Estas propiedades se ilustran en la Figura 3.

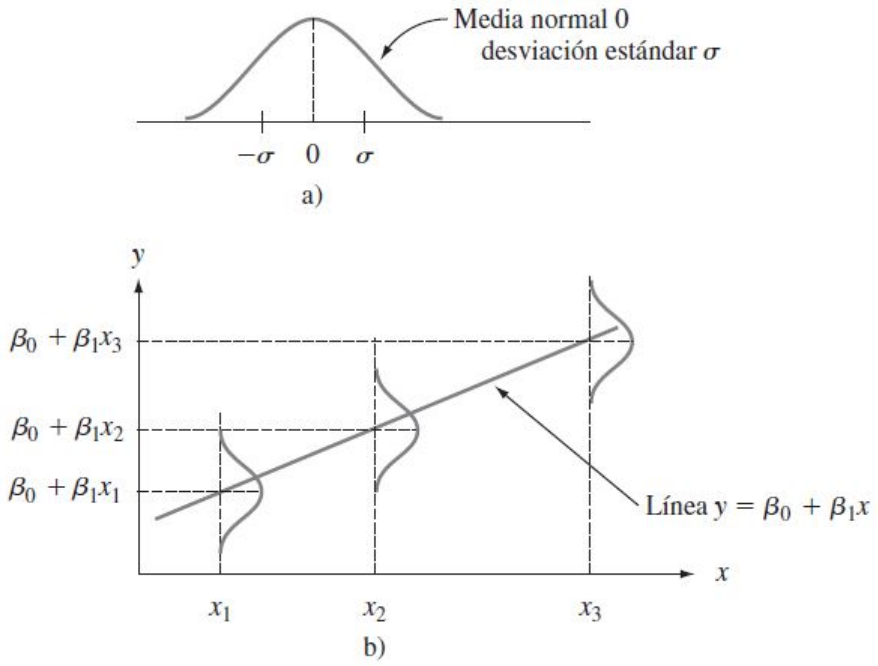

Figura 3: a) Distribución de ϵ. b) Distribución de Y para distintos valores de x.

El parámetro de varianza σ^2 determina el grado al cual cada curva normal se dispersa respecto a su valor esperado. Cuando σ^2 es pequeña, un punto observado (x_i, y_i) estará bastante cerca de la recta de regresión verdadera, mientras que si la varianza σ^2 es grande entonces dicho punto puede ser que se encuentre más alejado de dicha recta.

Ejemplo: Supongamos que el modelo de regresión lineal con una recta de regresión verdadera dada por $y = 65 - 1.2x$ y $\sigma = 8$, describe la relación entre el esfuerzo aplicado x y el tiempo de falla y. Entonces para un valor fijo x_0 de esfuerzo, el tiempo de falla sigue una distribución normal con valor medio de $65 - 1.2x_0$ y desvío estándar $\sigma = 8$. Para este caso, en la población compuesta por todos los pares (x, y), la magnitud de una desviación típica respecto a la recta de regresión verdadera es aproximadamente 8.

Si tomamos a $x_0 = 20$, entonces Y tiene un valor esperado dado por $E[Y/20] = \mu_{Y,20} = 65 - 1.2(20) = 41$. Supongamos que me interesa calcular la probabilidad de que el tiempo de falla sea mayor a 50 si el esfuerzo aplicado es de 20, es decir

$$P(Y > 50/x = 20) = P(Z > \frac{50-41}{8}) = 1 - P(Z < \frac{50-41}{8}) = 1 - \Phi(1.13) = 0.1292$$

Podría calcular este tipo de probabilidades para cualquier valor fijo x_0 ya que conozco la distribución que sigue Y para dichos casos. Esto se observa en la figura 4.

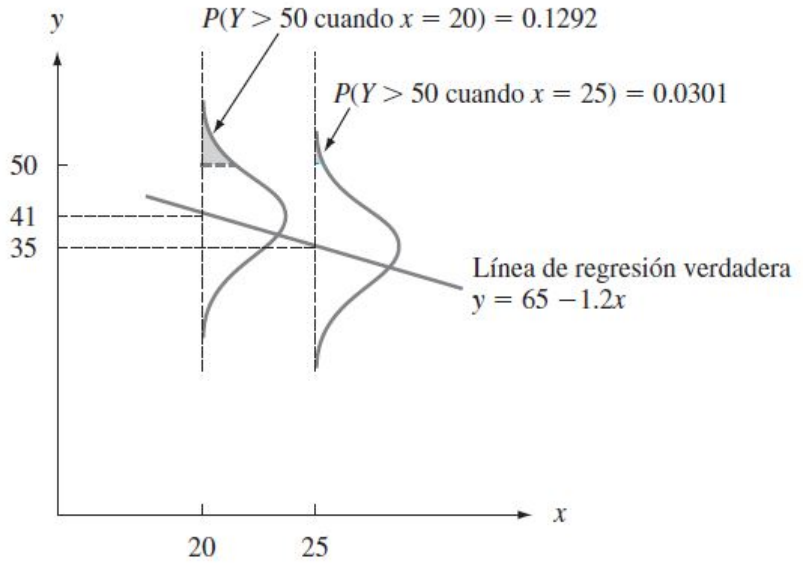

Figura 4: Probabilidades basadas en el modelo de regresión lineal simple.

Notar que en los cálculos previos estamos suponiendo que $P(Y/x = x_o)$ sigue una distribución Gaussiana, podría suceder que la distribución del error no sea Gaussiana (podría ser uniforme, o cualquier otra), entonces en ese caso la fpd que regiría mis estimaciones no sería la Gaussian sino la que gobierna al error aleatorio.

3. Estimación de parámetros del modelo

Supondremos entonces que las variables x e y están relacionadas de acuerdo con el modelo de regresión lineal simple. Nosotros nunca conoceremos los valores de β_0, β_1 y σ^2, pero tendremos disponible una muestra de datos compuesta por n pares (x_i, y_i). A partir de los cuales los datos del modelo de regresión y la recta de regresión lineal verdadera pueden ser estimados. Siempre supondremos que cada una de estas observaciones (x_i, y_i) se obtuvieron de forma independientes unas de otras. Entonces y_i es el valor observado de la variable aleatoria Y_i , donde $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, y las n desviaciones ϵ_i son independientes, por lo tanto las variables aleatorias $Y_1, Y_2, ..., Y_n$ también lo serán.

De acuerdo al modelo los puntos observados estarán distribuidos de forma aleatoria alrededor de la recta de regresión verdadera. Dicha recta será aquella que se ajuste mejor a a los datos observados. Esta idea es la que motiva el principio del método de mínimos cuadrados.

3.1. Principio de mínimos cuadrados aplicado al modelo de regresión lineal simple

El método de mínimos cuadrados consiste en determinar los parámetros β_0 y β_1 de manera tal que se minimice la suma de los cuadrados de las desviaciones verticales entre las observaciones y los valores predichos por la recta de regresión a estimar.

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))$$

Es decir, buscamos una estimación puntual para β_0 y β_1, que denotaremos como $\hat{\beta}_0$ y $\hat{\beta}_1$ llamadas **estimaciones de mínimos cuadrados**, que son aquellos valores para dichos parámetros que minimizan $L(\beta_0, \beta_1)$. A partir de estas estimaciones puntuales obtendremos la **recta de regresión estimada** que viene dada por $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (Fig. 5).

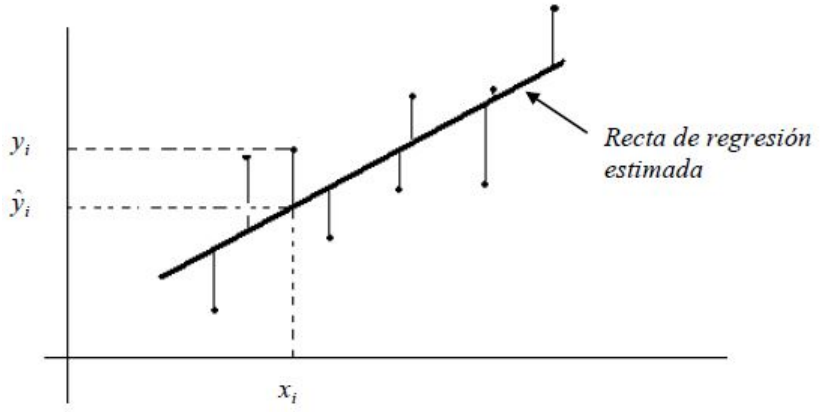

Figura 5: Recta de regresión estimada a partir de las observaciones.

Para econtrarlos entonces queremos minimizar $L(\beta_0, \beta_1)$, para esto calcularemos las derivadas parciales de esta función respecto de los parámetros de interés y las igualaremos a 0.

$$\frac{d(L(\beta_0, \beta_1)}{d\beta_0} = 0$$
$$\frac{d(L(\beta_0, \beta_1)}{d\beta_1} = 0$$

Tendremos un sistema de dos ecuaciones con dos incógnitas, desarrollando lo escrito arriba tengo que

$$\frac{d(L(\beta_0, \beta_1)}{d\beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{d(L(\beta_0, \beta_1)}{d\beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Cancelando el factor -2 en ambas ecuaciones y reordenando, se obtiene el siguiente sistema de ecuaciones conocido como **ecuaciones normales de mínimos cuadrados**

$$n\beta_0 + \left(\sum x_i\right)\beta_1 = \sum y_i$$
$$\left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 = \sum x_i y_i$$

Las solución de estas ecuaciones dan como resultado **las estimaciones puntuales de** β_0 **y** β_1.

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

La recta de regresión obtenida a partir del método de mínimos cuadrados no deberá ser utilizada para predecir un valor de y más allá del rango de los datos, es decir, si nuestra recta fue estimada a partir de pares ordenados (x_i, y_i) tal que el rango de x $\in$ [a,b], entonces no sería correcto predecir algun valor de y para valores de x fuera de este rango, ya que solo conocemos la relación entre las dos variables en el rango de x [a,b].

3.2. Errores y coeficiente de determinación

Los valores ajustados o predichos $\hat{y_1}, \hat{y_2}, ...\hat{y_n}$ se obtienen sustituyendo sucesivamente $x_1, x_2, ..., x_n$ en la ecuación de la recta de regresión estimada $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Se definen los **residuos**, e_i, como la diferencia entre los valores observados y_i y los valores predichos $\hat{y_i}$

$$e_i = y_i - \hat{y_i}$$

Si prestamos atención, lo que hicimos para encontrar las estimaciones puntuales para β_0 y β_1 fue minimizar la suma de los cuadrados de los residuos, que denotaré como SCE

$$SCE = \sum(e_i)^2 = \sum(y_i - \hat{y_i})^2$$

Se puede obtener un buen estimador (insesgado, de varianza mínima y consistente) para σ^2 dado por

$$\hat{\sigma^2} = \frac{SCE}{n-2} = \frac{\sum(y_i - \hat{y_i})^2}{n-2}$$

El $n-2$ del denominador aparece para garantizar que el estimador sea insesgado (similar a lo que pasa con el estimador de la varianza muestral S^2).

Una forma sencilla para el cálculo de SCE es rescribirlo, llegando a

$$SCE = \sum(y_i)^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

.

A partir de esto, podemos calcular el **coeficiente de determinación** que se interpreta como la proporción de los cambios de y que pueden ser explicados por el modelo de regresión lineal (atribuida a la relación lineal entre x e y), y viene dado por

$$r^2 = 1 - \frac{SCE}{STC}$$

$$0 \leq r^2 \leq 1$$

Siendo STC = *la suma total de las desviaciones al cuadrado respecto al valor medio de las observaciones*

$$STC = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

Mientras más alto sea el valor de r^2, más exitoso es el modelo de regresión lineal simple al explicar la variación de y. Cuando se realiza un análisis de regresión este coeficiente es una parte importante de los resultados, si r^2 es pequeño, entonces el analista deseará buscar un modelo alternativo, como por ejemplo un modelo no lineal o un modelo de regresión múltiple que tiene encuenta más de una variable independiente.

Si por ejemplo calculo el coeficiente de determinación y obtengo $r^2 = 0.97$, significa que el 97 % de la variación de y puede ser explicado por la relación lineal aproximada que tiene esta variable con x. La figura 6 muestra tres series de datos donde en a) $r^2 = 1$, es decir el comportamiento de y puede ser perfectamente explicado a partir de su relación lineal con x, y en b) y en c) el valor de r^2 será menor, ya que parte del comportamiento de y no puede ser explicado mediante una relación lineal con x.

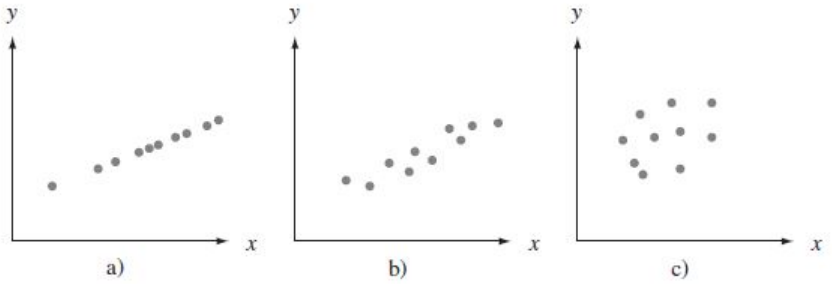

Figura 6: Utilización del modelo lineal para explicar la variación de y: a) datos con los cuales toda la variación es explicada; b) datos con los cuales gran parte de la variación es explicada; c) datos con los cuales poca variación es explicada.

Esto es solo la punta del iceberg en lo que se refiere a regresión lineal, tener en cuenta que es posible hacer un analisis mucho más profundo de cada uno de los estimadores y parámetros descriptos, por ejemplo para $\hat{\beta}_0$ y $\hat{\beta}_1$ se les puede aplicar lo visto en intervalos de confianza, lo mismo para $\hat{y}$.