Linear Regression and the Least-squares problem

Aniel Nieves-González

Abril 2013

Variables cuantitativas y relaciones entre estas (Repaso)

Recuerde que informalmente...

- Una variable aleatoria (v.a.) es una asociación (relación) entre los miembros del espacio muestral (conjunto de objetos que nos interesa estudiar) y algún conjunto de números.

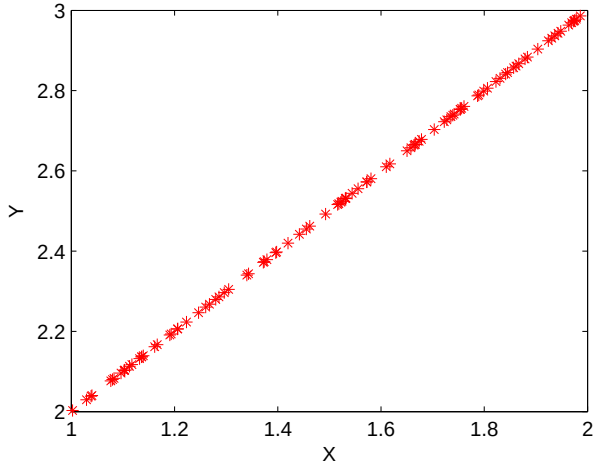
Definition (Response variable)

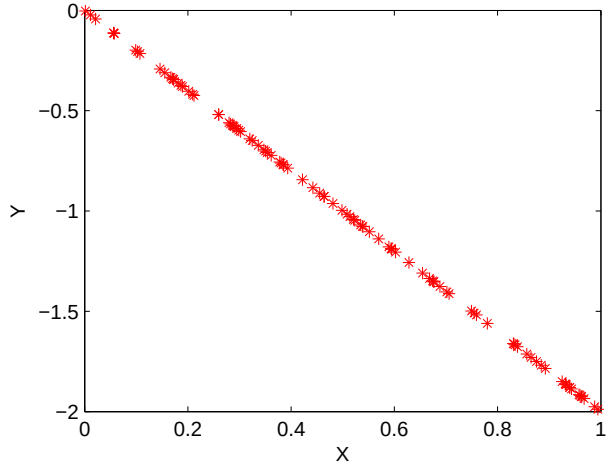
Una v.a. que mide el resultado de un estudio es un response variable. También es llamada **variable dependiente**.

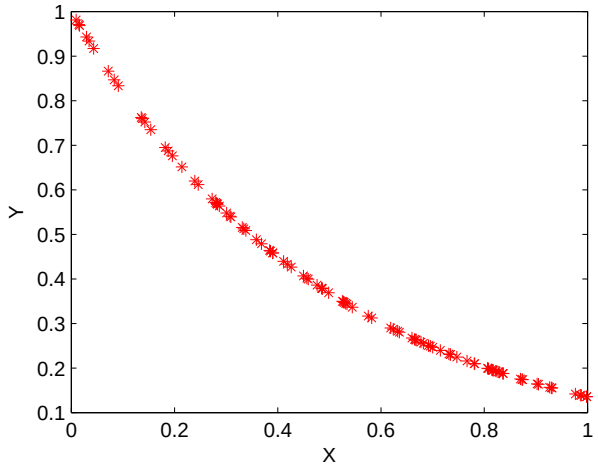
Definition (Explanatory variable)

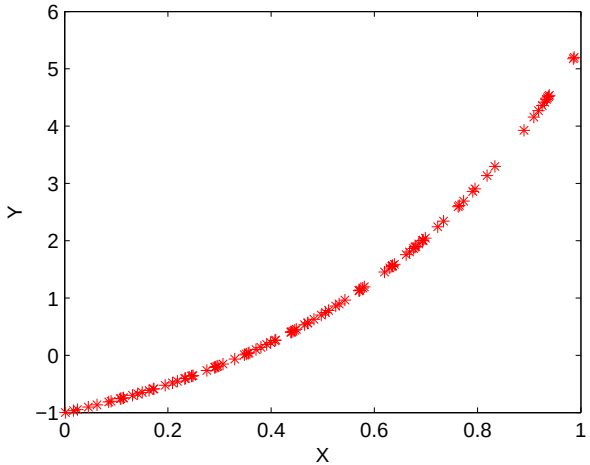
Variable que explica o influencia los cambios en el response variable. También es llamada **variable independiente**.

- Dos v.a. están **asociadas positivamente** si valores sobre el prom. de una acompañan a valores sobre el promedio de otra, y valores por debajo del prom. tienden a ocurrir juntos.
- D es sobre el pr io de otra, y vi









Sección 2.1, 2.2 (Repaso)

El concepto de **correlación** mide cuantitativamente la dirección y fuerza de una relación *lineal* entre dos variables.

Definition (Correlación de la muestra)

Suponga q. tiene n datos de las variables X y Y. Denotamos dichos datos como $\{x_1, \ldots, x_n\}$ y $\{y_1, \ldots, y_n\}$ respectivamente. Sea $\bar{X}$ y $s_x$la media aritmética y la desviación estándar para los datos de X. Análogamente $\bar{Y}$ y s_y para con Y. La **correlación** de la muestra, denotada como r, entre X y Y es

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{s_x} \right) \left(\frac{y_i - \bar{Y}}{s_y} \right)$$

Sección 2.1, 2.2 (Repaso)

Observe:

- r no hace distinción entre var. dependente e independiente.
- r solo aplica a variables cuantitativas.
- r no tiene unidades ("dimensionless").
- $r > 0 \Rightarrow$ relación lineal positiva
- $r < 0 \Rightarrow$ relación lineal negativa
- $|r| < 1$ (i.e. $-1 \leq r \leq 1$).
- $r \approx 0 \Rightarrow$ relación lineal débil.
- $r \approx -1$ o $r \approx 1$ implica relación lineal fuerte.
- Por último: r mide fuerza de relaciones **lineales** solamente. Y al igual que $\bar{X}$ no es resistente a outliers.

Ejemplo 1: Calculando r

Suponga los siguientes datos y calcule r.

rapidez (speed)	20	30	40	50	60
MPG	24	28	30	28	24

Sea $X \equiv$ 'rapidez en mph', $Y \equiv$ 'MPG', y $n = 5$.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{20+30+40+50+60}{5} = \frac{200}{5} = 40$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{(-20)^2 + (-10)^2 + (0) + (-10)^2 + (10)^2}{4}$$

$$= \frac{400+100+0+100+400}{4} = \frac{1000}{4} = 250$$

$$\Rightarrow s_x = 15.81$$

Análogamente:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{134}{5} = 26.8$$

Recuerde que el modelo ANOVA es:

$$x_{ij} = \mu_i + \epsilon_{ij}$$

Ahora estudiaremos otro modelo usado para estudiar nuestro sistema de interés. Dos v.a. pueden estar relacionadas linealmente o no linealmente. El primero es el tipo de relación entre v.a. más simple.

Definition (Simple Linear Regression model)

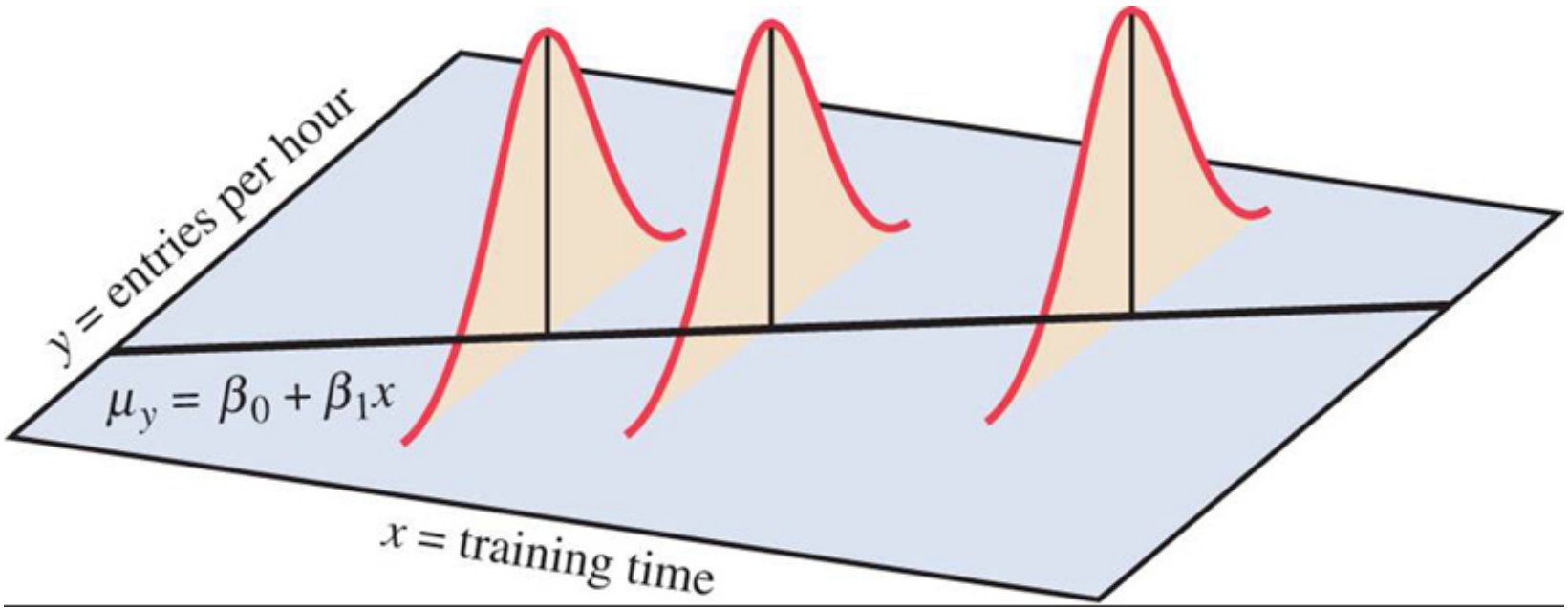
Considere n observaciones independientes de la variable independiente X y de la variable dependiente Y, donde $Y \sim N(\mu_y, \sigma)$:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$$

El **simple linear regression statistical model** es

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

donde los parámetros del modelo son el intercepto en y (β_0), la pendiente (β_1) de la linea de regresión de la población, y la variabilidad en y alrededor de esta linea (σ). ϵ_i es el error aleatorio para la i-ésima observación. Los ϵ_i son independientes y se distribuyen $N(0, \sigma)$. Note que $\mu_y = \beta_0 + \beta_1 X$ es la respuesta media (mean response) cuando $x = x_i$.


y = entries per hour
$\mu_y = \beta_0 + \beta_1 x$
x = training time

- La meta es obtener de la muestra estimados de β_0, β_1, σ. A dichos estimados le llamaremos b_0, b_1, y s_{YX} respectivamente.
- El término de "regresión" lo acuñó Francis Galton en el s. 19 para describir un fenomeno biológico. El fenomeno consistía en que en promedio la altura de los decendientes tiende a regresar ("regress") hacia el promedio (regression toward the mean). O sea que la altura de los decendientes de padres altos es menor que la altura de los padres.

Least-squares regression (regresión por cuadrados mínimos)

En esencia el problema de cuadrados mínimos (o least squares (LS) problem o LS regression) es un problema de minimización, el cual se puede escribir como:

$$\min_{\mathbf{p}} \sum_{i=1}^{n} (F(x_i, \mathbf{p}) - y_i)^2$$

donde:

- $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ son los datos (note las dos variables X y Y)
- $\mathbf{p}$ son los parámetros (constantes del modelo matemático).
- $F(x_i, \mathbf{p})$ es el modelo matemático. Note que F depende de los datos y de los parámetros.
- En este caso el modelo matemático representará a la relación entre las dos variables X y Y.
- En LSP se buscan los parámetros del modelo que minimizan la distancia vertical (a lo largo de eje de y) entre modelo y datos.

Least-squares regression

En este caso F (el modelo) será simple linear regression model, esto es,

$$F(x_i, b_0, b_1) = b_0 + b_1 x_i \quad \text{una función lineal.}$$

$F(x_i, b_0, b_1)$ puede renombrarse como $\hat{y}_i$.
Usando técnicas de cálculo multivariable (optimización) se obtienen las *ecuaciones normales*. Las mismas se resuelven para los parámetros (b_0 y b_1 en este caso). Observe que:

- Para la regresión por LS es importante distinguir entre var. independiente y dependiente.
- La pendiente (slope) del LS regression es

$$b_1 = r\frac{s_y}{s_x} = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{X}}{s_x}\right)\left(\frac{y_i - \bar{Y}}{s_y}\right)\right)\frac{s_y}{s_x}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}$$

y el intercepto en y (y-intercept) es

$$b_0 = \bar{Y} - b_1\bar{X}$$

Una medida de la variabilidad de alrededor de la linea de regresión es el standard error de la regresión. Este es un estimador de σ. El mismo se define como

$$s_{XY} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

donde $\hat{y}_i = b_0 + b_1 x_i$. Este es un "unbias estimator" de σ.

Suposiciones para regresión e inferencia

Vimos el planteamiento de simple linear regression model para estudiar la relación entre dos v.a. y como se encuentran los parámetros de la linea resolviendo el LSP. Ahora estudiaremos las condiciones que se deben cumplir si queremos hacer inferencia con la regresión lineal.

Suposiciones para regresión e inferencia

Las cuatro suposisiones importantes en la regresión lineal son:

1. *Normality*: se requiere que los valores de Y esten normalmente distribuidos para cada valor de X. (El análisis de regresión es robusto para violaciones de esta suposición).
2. *Homoscedasticity (Homocedastisidad)*: se requiere la variación alrededor de la linea de regresión sea constante para todos los valores de X (σ constante).
3. *Independence of error* (la muestra es un SRS): se requiere que el error residual $(y_i - \hat{y}_i)$ sea independiente para cada valor de x_i.
4. *Linealiadad*: se requiere que la relación entre las variables sea lineal. (Dos variables pueden estar relacionadas en forma no lineal y el coef. de correlación lineal puede ser cero.)

Intervalos de confianza (IC) y test de significancia (TS)

Los IC para la pendiente e intercepto en y son de la forma:

$$\text{estimate} \pm t^* \text{SE}_{\text{estimate}}$$

donde t^* es el valor crítico de la dist. t.
IC para β_1 y β_0 están basados en las dist. de b_1 y b_0:

- Cuando el modelo de regresión lineal es cierto b_0, b_1 se dist. normalmente.
- $\mu_{b_0} = \beta_0$ y $\mu_{b_1} = \beta_1$. Esto es, b_0 y b_1 son estimadores sin sesgo de β_0 y β_1.
- σ_{b_0} y σ_{b_1} son multiplos de σ.

Intervalos de confianza (IC) y test de significancia (TS)

Definition (Inferencia para la pendiente)

Un IC de nivel C para pendiente β_1 de la linea de regresión de la población es:

$$b_1 \pm t^* \mathrm{SE}_{b_1}$$

Donde el area bajo la curva de densidad de $t(n-2)$ y sobre el intervalo $[-t^*, t^*]$ es igual a C. El Margen de error (MOE) es $t^* \mathrm{SE}_{b_1}$ y el standard error de b_1 (SE_{b_1}) es:

$$\mathrm{SE}_{b_1} = \frac{s_{YX}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}}$$

Intervalos de confianza (IC) y test de significancia (TS)

Definition (Inferencia para la pendiente)

Para el test de hipótesis $H_0 : \beta_1 = 0$ (no relación lineal entre X y Y) se calcula la estadítica t

$$t = \frac{b_1}{\text{SE}_{b_1}}$$

En términos de una v.a. T que se distribuye con $t(n-2)$, el P-value para probar H_0 en contra de H_a es:

$$\begin{aligned} H_a : \beta_1 > 0 \quad &\text{es} \quad P(T \geq t) \\ H_a : \beta_1 < 0 \quad &\text{es} \quad P(T \leq t) \\ H_a : \beta_1 \neq 0 \quad &\text{es} \quad P(T \geq t) + P(T \leq -t) = 2P(T > |t|) \end{aligned}$$

Intervalos de confianza (IC) y test de significancia (TS)

- Los IC y TS para el intercepto en y son análogos a los de la pendiente (cambie subscrito en las expresiones, de 1 a 0) y use la siguiente expresión para el SE:

$$\mathrm{SE}_{b_0} = s_{XY}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2}}$$

- *Regression fallacy*: Es suponer que "regression toward the mean" es consecuencia de algún efecto sistemático. O sea es la falla de no considerar fluctuaciones naturales del sistema bajo consideración.

Inferencia y correlación

El concepto de correlación de la muestra puede generalizarse a la correlación de la población (ρ).

- Este será un parámetro de la población y el estimador correspondiente es r (correlación de la muestra).
- Este es un número con las propiedades y la interpretación de r.
- La pendiente de la linea de regresión de la población (β_1) nos da info. de ρ.
- De hecho, cuando β_1 es cero, positiva, o negativa entonces támbien lo es ρ.

Inferencia y correlación (no se embotelle este slide...)

Si X y Y son dos v.a. la correlación ρ se define como

$$\begin{aligned}\rho &= \text{Cov}\left(\frac{X - E[X]}{\sigma_X}, \frac{Y - E[Y]}{\sigma_Y}\right) \\ &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}\end{aligned}$$

donde $E[X]$ y $E[Y]$ denotan μ_X y μ_Y respectivamente, $\text{Cov}(X, Y)$ es la covarianza de X y Y (i.e. $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$), y σ_X, σ_Y son la desv. estándar de X y Y respectivamente.
Note, $\text{Cov}(X, X)$ es igual a...

Inferencia y correlación

Conociendo r: ¿que podemos inferir sobre ρ?

Definition (Test for zero population correlation)

Considere el test de hipótesis

$$H_0 : \rho = 0 \quad \text{no correlación}$$
$$H_a : \rho \neq 0 \quad \text{hay correlación}$$

Puede definirse una estadística t usando a r:

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

y usar la dist. t con $n-2$ grados de libertad para calcular P-value.

Note que para probar H_0 se toma $\rho = 0$ en la def. anterior de la estadistica t.

Usando la linea de regresión

La idea de la regresión lineal es estudiar la relación (si alguna) entre dos v.a. y hacer predicciones (cuidado con extrapolación). Considere el siguiente ejemplo:

Ejemplo:

Considere el "stock market". Algúnos piensan que el comportamiento del stock market a principios del año predice lo que ocurriá el resto del año. Ahora, suponga que X es la variable independiente y Y la dependiente donde:

$X \equiv$ taza de cambio en el stock market index en enero

$Y \equiv$ taza de cambio en el stock market index en el resto del año

Los datos en los últimos 38 años se resumen como sigue: $\bar{X} = 0.0175$, $\bar{Y} = 0.0907$, $s_x = 0.0536$, $s_y = 0.1535$, y $r = 0.596$

1. Calcule la linea de regresión:

$$b_1 = r\frac{s_y}{s_x} = (0.596)\frac{0.1535}{0.0536} = 1.7068$$

$$b_0 = \bar{Y} - b_1\bar{X} = 0.0907 - (1.7068)(0.0175) = 0.0608$$

$$\Rightarrow \hat{y}_i = 1.7068x_i + 0.0608$$

2. Suponga que sus datos para X fluctuan entre 0.005 y 0.03. Si queremos predecir el valor de de $\hat{Y}$ para $x = 0.012$, que hacemos...

$\hat{Y}(0.012) = 1.7068(0.012) + 0.0608 = 0.0813$

Usando la linea de regresión

Ahora, si en lugar de querer predecir el valor de Y para un posible valor en particular de X (ejemplo anterior), queremos estimar el "mean response" de Y (i.e. μ_y) entonces necesitamos construir un IC para β_0 y β_1.