Regresión lineal simple

(+ ejemplo y código en R)

J.E. Alcalá
Centro de Estudios e Investigaciones en Comportamiento
Universidad de Guadalajara

30.04.2018

Ejemplo de diapositivas
Propelente

Tabla 1: Datos de propelente

Observación	y	x	Observación	y	x
1	2158.7	15.5	11	2165.2	13
2	1678.15	23.75	12	2399.55	3.75
3	2316	8	13	1779.8	25
4	2061.3	17	14	2336.75	9.75
5	2207.5	5.5	15	1765.3	22
6	1708.3	19	16	2053.5	18
7	1784.7	24	17	2414.4	6
8	2575	2.5	18	2200.5	12.5
9	2357.9	7.5	19	2654.2	2
10	2256.7	11	20	1753.7	21.5

En la tabla 1 se muestran los datos de la resistencia al corte (y) como función de la edad del propelente (x). En un gráfico de dispersión se puede observar qué tipo de relación podría existir entre ambas variables (Figura ??)

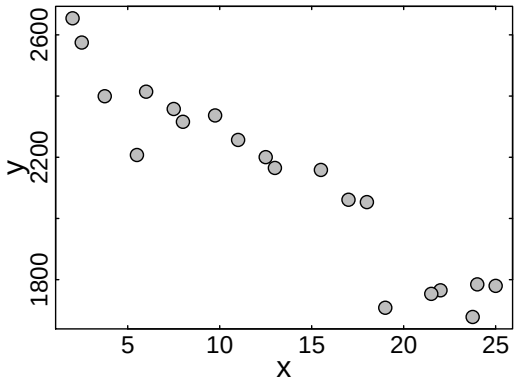

Figura 1: Resistencia al corte como función de la edad del propelente.

Preliminares

La regresión lineal simple es un método que usa la relación estadística entre dos variables cuantitativas en la que una variable, la variable de respuesta (o variable dependiente) puede ser predicha a partir de otra variable (la variable predictora o independiente).

Sean $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ n pares de observaciones tales que y_i es el valor observado (o el valor de la variable de respuesta) de una variable aleatoria Y_i. Asumimos que existen constantes β_0 y $\beta_1 \neq 0$ tales que Y_i sea una combinación lineal de X_i con la siguiente forma

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

Donde $\epsilon_1, \epsilon_2, ..., \epsilon_n \sim N(\mu = 0, \sigma^2)$ y $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, es decir, los errores no están correlacionados entre sí. La ecuación 1 se dice que lineal en los parámetros β_0, β_1, y es regresión simple dado que solo existe una variable X que predice la respuesta en Y.

En su núcleo, la regresión lineal se trata de un problema de aproximación [1, cáp. 8]. Dado que asumimos que $\mathbf{E}[\epsilon] = 0$, el valor esperado del error, la ecuación 1 se reduce a $\beta_0 + \beta_1 x_i$. El problema de la regresión lineal simple se reduce a encontrar la mejor aproximación lineal que minimice la diferencia $y_i - (\beta_0 + \beta_1 x_i)$. La mejor forma de expresar las diferencias es elevándolas al cuadrado:

- Simplemente usando las diferencias los valores negativos y positivos se podrían cancelar.
- Usando valores absolutos no le daría el suficiente peso a valores alejados de la línea (valores atípicos).
- Las diferencias elevadas al cuadrado son diferenciables, lo cuál es bueno si se busca una función que las minimice.

Con base en esto se usa el método de mínimos cuadrados, en los cuáles se intentan obtener β_0 y $\beta_1 \neq 0$ tales que se minimicen la función

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2}$$

con respecto a los parámetros β_0, β_1. Para encontrar tales mínimos, necesitamos las derivadas parciales con respecto a β_0, β_1 que minimicen S

$$\frac{\partial S}{\partial \beta_0} = 0 \text{ y } \frac{\partial S}{\partial \beta_1} = 0$$

esto es,

$$0 = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0) \tag{3}$$

y,

$$0 = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = -2 \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)(x_i) \tag{4}$$

Las ecuaciones 3 y 4 se simplifican en las llamadas **ecuaciones normales**, donde la palabra **normal** significa **perpendicular**. Dado que estamos estimando a β_0, β_1, denotamos a estos estimadores como $\hat{\beta}_0, \hat{\beta}_1$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{5}$$

y

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{6}$$

La solución al sistema de 5 y 6 es:

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \tag{8}$$

Nótese que la ecuación 7 se puede expresar como $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. También, nótese que dividiendo 9 por n nos da:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{(\sum_{i=1}^{n} x_i^2) - (\bar{x})^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \tag{9}$$

Por 7 sabemos que el valor de $\hat{\beta}_0$ es igual al valor esperado de y_i cuando $x_i = 0$, por lo que se le suele llamar *intercepto* u *ordenada en el origen*, sin embargo, si x_i nunca es 0, este valor no tiene una interpretación intrínseca.

Regresión con datos de propelente

$$\mathrm{Cov}(x,y) = -2163.82 \text{ y } \mathrm{Var}(x) = 58.2$$

Por lo tanto, $\hat{\beta}_1 = -2163.82/58.2 = -37.15$. Además, $\mathbf{E}[\mathbf{y}] = 2131.3$ y $\mathbf{E}[\mathbf{x}] = 13.36$. Por lo tanto, $\hat{\beta}_0 = 2131.13 + 37.15 * 13.36 = 2627.8$. En notación matricial, nuestra ecuación de regresión quedaría:

$$\hat{\mathbf{Y}} = \mathbf{X}\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = \mathbf{X}\begin{bmatrix}2627.8 \\ -37.15\end{bmatrix}$$

Con la cual podríamos obtener los valores de $\hat{\mathbf{Y}}$. Con estos valores podemos ajustar una línea a nuestros datos, calcular el error y medir otros estadísticos.

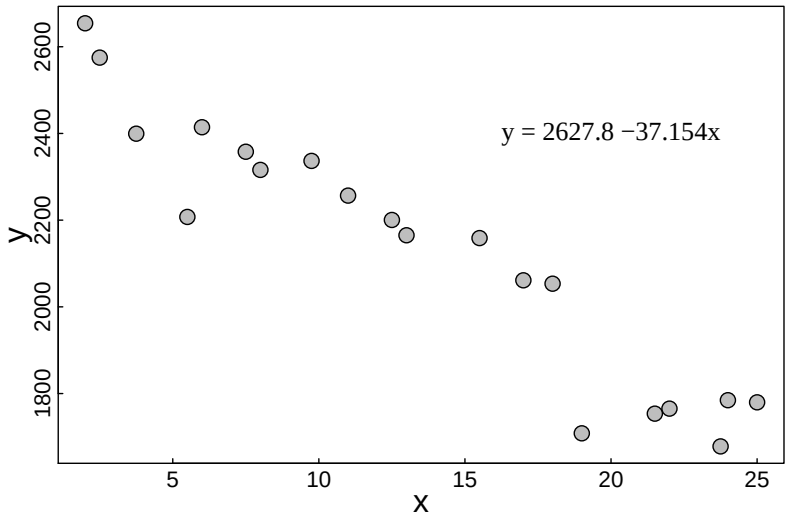

Figura 2: Ajuste de línea de regresión a partir de los parámetros encontrados por el método de mínimos cuadrados. La línea azul representa la recta de mejor ajuste minimizando la distancia entre los puntos grises y los azules. Las líneas verticales representan el error $\epsilon_i = y_i - \hat{y}$

.

Después de obtener el modelo de nuestros datos, debemos preguntarnos lo siguiente:

- ¿Qué tan bueno es el ajuste de esta ecuación a los datos (bondad de ajuste)?
- ¿Es el modelo un buen predictor? Es decir, ¿podríamos obtener valores confiables de y con nuevos valores de x?
- ¿Se violan los supuestos básicos de la regresión? Tales como la varianza constante, ϵ_i no correlacionados entre sí, y $\epsilon_i \sim N(\mu = 0, \sigma^2)$

Para respondernos lo anterior necesitamos un estimador de σ^2, que denotamos por $\hat{\sigma}^2$ y estimar un intervalo de confianza. $\hat{\sigma}^2 = SS_{res}/(n-2)$ es la suma de cuadrados residuales promediada entre los grados de libertad (n-2 parámetros, que lo vuelven un estimador robusto).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{10}$$

A partir de lo cual obtenemos $\hat{\sigma}^2 = 166253.7/(20-2) = 9236.31$ y el error estándar de la estimación de la regresión, $\hat{\sigma} = \sqrt{9236.3} = 96.105$

Las varianzas de los estimadores se obtienen con

$$\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}} \tag{11}$$

y

$$\text{Var}(\hat{\beta}_0) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \tag{12}$$

Así tenemos:
$\text{Var}(\hat{\beta}_1) = \frac{9236.31}{1106.56} = 8.34$ y $\hat{\sigma}_{\hat{\beta}_1} = 2.88$;
$\text{Var}(\hat{\beta}_0) = 9236.31\left(\frac{1}{20} + \frac{13.36^2}{1106.56}\right) = 1952.2$ y $\hat{\sigma}_{\hat{\beta}_0} = 44.18$.

Estos valores calculados nos servirán para probar si los valores de los parámetros son iguales a una constante determinada. Generalmente, para la pendiente y el intercepto, o sus estimados $\hat{\beta}_0 \text{y} \hat{\beta}_1$ se desea saber si son significativamente diferentes de 0 (si la pendiente es 0, implicaría que no hay relación lineal entre las variables, y si el intercepto es 0).

$$\frac{H_0 : \beta_0 = 0,\ H_a : \beta_0 \neq 0 \text{ para } \beta_0}{H_0 : \beta_1 = 0,\ H_a : \beta_1 \neq 0 \text{ para } \beta_1}$$

La forma general de probar la hipótesis concerniente a la diferencia $\beta_j = \beta_{H_0}$ es:

$$t_{\beta_j} = \frac{\hat{\beta}_j - \beta_{H_0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_j}}} \tag{13}$$

A partir de lo cual obtenemos:

$$t_{\beta_0} = 59.48 \text{ y}$$
$$t_{\beta_1} = -12.86$$

El valor teórico para la distribución de t-Student es $t_{\alpha/2,n-2} = t_{0.05/2,18=2.75}$. Sabemos que si al nivel de significancia de $\alpha = 0.05$ si $|t_{\beta_j}| > t_{0.05/2,18=2.75}$ podemos rechazar H_0. Dado que $59.48 > 2.75$ para β_0 y $|-12.86| > 2.75$ para β_1 en ambos casos decimos que son estadísticamente diferentes de 0 a un nivel de significancia de 0.05. Podemos computar el valor p, que es una *densidad*

[1] de probabilidad condicional de la siguiente forma $p(t_{\beta_j} > t_{\alpha/2,n-2}|H_0)$, con lo que obtenemos: $p = 1.64e-10$ para t_{β_1} y $p = 4.063559e-22$ para t_{β_0}.

Ahora calculamos la F de Fisher para probar la significancia del modelo para ajustar los datos. Recordemos que el estadístico de Fisher se define como

$F_0 = MS_R/MS_{res}$, donde MS_i hace referencia a la suma cuadrática media

$SS_R = \hat{\beta}_1 S_{xy}$, $SS_{res} = SS_T - \hat{\beta}_1 S_{xy}$ con gradis de libertad de 1 y $n-2$ respectivamente. A partir de esto, tenemos una F de Fisher de $1527334.9/9244.59 = 165$. Sabemos que la F es significativa si $F_0 = F_{\alpha,n-2}$, cuyo valor es $F_{0.05,18} = 4.14$, por lo tanto $165.21 > 4.14$ y concluimos que el modelo es significativo.

Por último, calculamos el coeficiente de determinación, que nos dice la varianza explicada por el modelo. Se calcula de la siguiente manera:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T} \tag{14}$$

Con lo cual obtenemos $R^2 = 1527334.95/1693737.6 = 0.901$, es decir, nuestro modelo de regresión explica el $90.18\,\%$ de la varianza.

Por último, podemos calcular los intervalos de confianza para nuestra regresión. Estos se calculan mediante $\hat{y} \pm t_{\beta_1} * \mathrm{Var}(\hat{\beta}_1)$. Con esto, podemos construir un gráfico como el de la Figura 3.

[1]Es decir, es la integral de la función de densidad de $t - Student$ en la región de confianza, y por lo tanto es un área.

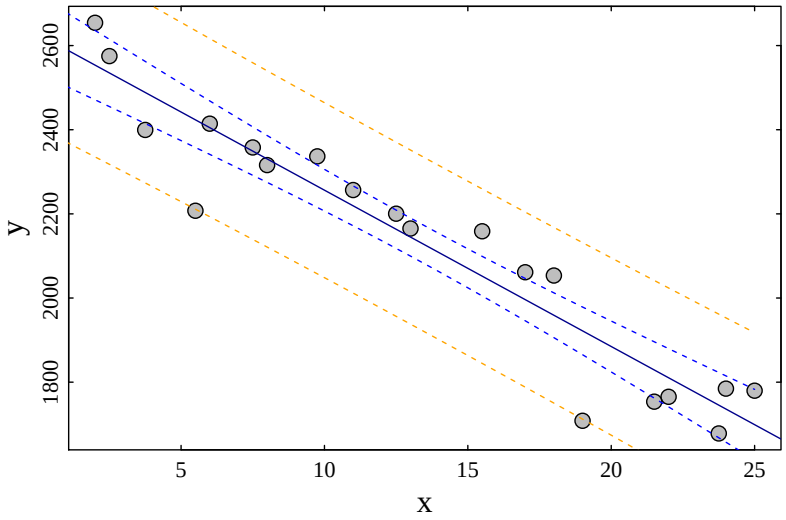

Figura 3: Regresión mostrando los intervalos de confianza (líneas azules) y de predicción (líneas naranjas)

.

Apéndice

Supuestos de la regresión lineal

1. $\mathbf{E}[\epsilon_\mathbf{i}] = \mathbf{0}$ para todo $i = 1, 2, ..., n$ o equivalentemente, $\mathbf{E}[\mathbf{y_i}] = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ para todo $i = 1, 2, ..., n$ o equivalentemente, $\text{Var}[y_i] = \sigma^2$.
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$, o equivalentemente, $\text{Cov}(y_i, y_j) = 0$.

El primer supuesto implica que y_i depende solo de x_i y el resto de la variación es aleatoria. El segundo supuesto implica que la varianza de ϵ o y no depende de los valores de x_i. Este supuesto se conoce comúnmente como supuesto de *homocedasticidad* o de *varianza homogénea*. El tercer supuesto implica que que los i-ésimos ϵ (o y) no están correlacionados unos con otros.

Código de R

```r
# Regresion lineal - Metodos estadisticos
# Set de problema de propelente

y = c(2158.7,1678.15,2316,2061.3,2207.5,1708.3,1784.7,
  2575,2357.9,2256.7,2165.2,2399.55,1779.8,2336.75,
  1765.3,2053.5,2414.4,2200.5,2654.2,1753.7)
x = c(15.5,23.75,8,17,5.5,19,24,2.5,7.5,11,13,3.75,
  25,9.75,22,18,6,12.5,2,21.5)

# Comandos y parametros para guardar graficos
pdf("xy_plot_t11.pdf",width = 4,height = 3)
par(tcl = 2, mgp=c(1, 0.1, 0),mar=c(2,2,1,1))
plot(x,y, pch = 21,col = "black", bg = "grey",tck=0.01,
cex =1.3, cex.lab=1.3,cex.axis=1.1)
# Marcas de los ejes derecha y arriba
axis(side = 3,tck=0.01,labels = FALSE)
axis(side = 4,tck=0.01,labels = FALSE)
dev.off()

 beta_0 y beta_1 con ecuacion normal
 # Primero necesitamos la matriz de diseno, X
 # Con 1s en una columna y x en la segunda

 X = matrix(c(rep(1,length(x)),x), ncol = 2)

 # la multiplicacion de X'X da como resultado la matriz
 # |   n       Sigma x_i|
 # |Sigma x_i  Sigma x^2|

 t(X)  %*% X

 # La ecuacion normal en forma matricial es
 # betas = inv(X'X)X'Y : la inversa del producto matricial de X
   y X
 # por el producto matricial de X y y

 betas = solve(t(X)  %*% X)  %*% (t(X)  %*% y)

 betas

 # lm arroja el mismo resultado

 coef(lm_model <- lm(y ~ x,data = df_reg))

 # Valores ajustados con modelo de ecuacion normal
```

```r

fitted_y = betas[1] + betas[2] * x

pdf("reg_fitted.pdf",width=6,height = 4)
par(tcl = 2, mgp=c(1, 0.1, 0),mar=c(2,2,1,1))
plot(x,y,pch = 21,col = "black", bg = "grey",
tck=0.01,cex.lab=1.4,cex.axis=1, cex = 1.5)
lines(x,fitted_y)
points(x,fitted_y,pch = 21, col = "blue",bg = "lightblue")
segments(x0=x,x1=x,y0=y,y1=fitted_y, col = "darkblue")
text(x=20,y=2400,substitute(
paste("y = ",beta_0," ", beta_1, "x"),
list(beta_0 = signif(betas[1],5),beta_1 = signif(betas[2],5))
),family = "serif", cex = 1.2)
dev.off()

# Intervalos de confianza y prediccion

# conjunto de x nuevo
new_x = 1:max(x)

pdf("reg_ci_pi.pdf",width=6,height = 4, family = "serif")
par(tcl = 2, mgp=c(1.5, 0.2, 0),mar=c(2.5,2.5,1,1))
plot(x,y,pch = 21,col = "black", bg = "grey",
tck=0.01,cex.lab=1.4,cex.axis=1, cex = 1.5)
abline(lm_model, col="darkblue")
axis(side = 3,tck=0.01,labels = FALSE)
axis(side = 4,tck=0.01,labels = FALSE)
conf_interval <- predict(lm_model, newdata=data.frame(x=new_x),
    interval="confidence",
level = 0.95)
lines(new_x, conf_interval[,2], col="blue", lty=2)
lines(new_x, conf_interval[,3], col="blue", lty=2)

pred_interval <- predict(lm_model, newdata=data.frame(x=new_x),
    interval="prediction",
level = 0.95)
lines(new_x, pred_interval[,2], col="orange", lty=2)
lines(new_x, pred_interval[,3], col="orange", lty=2)
dev.off()

# Manualmente se pueden calcular todos los valores de t y p-
  value
# t value

```

```r
SSres = sum((y - fitted_y)^2)

# La varianza del error es SSres/(n-2)
n = length(y)
sigma_err  = SSres/(n - 2)

# La varianza del beta_1 (slope) es

var_slope = sigma_err/Sxx

# del intercepto

var_intercept = sigma_err*(1/n + (mean(x)^2)/Sxx)

# El valor t del beta_j es dado por :
# (beta_j - beta_H0)/sqrt(var_beta)

# para slope beta_H0 = 0, la hipotesis nula

t_slope = (betas[2] - 0)/sqrt(var_slope)

# Para intercept

t_intercept = (betas[1] - 0)/sqrt(var_intercept)
# p value para ambos usando pt(t,n)

# p de slope, multiplicar por dos para hacerla two.tailed
# lower.tail = F y tomar valor absoluto de t_slope, dado que
# estamos calculando la probabilidad de que t_calculado < T de
  distribucion

2*pt(abs(t_slope), df=n-2, lower.tail = F)

2*pt(abs(t_intercept), df=n-2, lower.tail = F)

# comparar con calculado por funciones

summary(lm_model)
```

Referencias

[1] R. L. Burden, J. D. Faires, and A. Burden. *Numerical analysis. 2015.* Cengage Learning, 9 edition, 2015.