UNIVERSIDAD SIMÓN BOLÍVAR
DEPARTAMENTO DE CÓMPUTO CIENTÍFICO Y ESTADÍSTICA
CÁTEDRA: ESTADISTICA PARA INGENIEROS (CO-3321)

Laboratorio de Regresión Lineal Simple.

Las técnicas de regresión lineal simple busca establecer una relación entre una variable de respuesta o variable dependiente y, y una variable explicativa, predictoria o independiente x

La ecuación de regresión lineal simple tiene la forma:

$$y = \beta_0 + \beta_1 x$$

Métodos de mínimos cuadrados:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

donde ε es la componente de error o residuo aleatorio; $\varepsilon \sim N(0, \sigma^2)$.

De aquí se obtiene:

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

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde:

Estos estadísticos $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados de los parámetros β_0 y β_1, respectivamente, con varianzas:

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Para la i-ésima observación, el valor predicho por el modelo es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$

El residuo correspondiente a la i-ésima observación es:

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

Por definición, la suma de los residuales es cero, pero las varianzas no son uniformes. para tomar esto en cuenta, definimos los residuales estandarizados:

$$e_{s_i} = \frac{e_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Estos residuos estandarizados no suman cero, pero tienen la misma varianza.

Coeficiente de correlación muestral: Es un estimador del coeficiente de correlación poblacional ρ y viene dado por:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$-1 \leq R \leq 1$

Nota:

- $R = 0$ implica ausencia de correlación lineal entre x e y.
- $R > 0$ implica correlación lineal positiva entre x e y.
- $R < 0$ implica correlación negativa entre x e y.

Coeficiente de determinación: Representa la proporción de la variación total en y que es explicada por x. Si R^2 está cerca de 1, entonces x explica una gran parte de la variación en y.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$0 \leq R^2 \leq 1$

Inferencias respecto a los parámetros: Para hacer pruebas de hipótesis o intervalos de confianza, el estadístico de prueba es:

- Para β_0:

$$T = \frac{\hat{\beta}_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t_{n-2}$$

donde:

$$s^2 = \frac{1}{n-2}\sum_{i=1} n(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

El error estándar es:

$$e.e(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Se tiene:

1. Para una prueba de hipótesis bilateral:
   $H_0 : \beta_0 = \beta_{00}$ contra $H_1 : \beta_0 \neq \beta_{00}$
   La región de rechazo es: $RR = (-\infty,\ -t_{n-2;\alpha/2}) \cup (t_{n-2;\alpha/2},\ \infty)$
   El p-valor es $2 - 2P(T \leq |t_{obs}|)$

2. Para una prueba de hipótesis unilateral derecha:
   $H_0 : \beta_0 = \beta_{00}$ contra $H_1 : \beta_0 > \beta_{00}$
   La región de rechazo es: $RR = (t_{n-2;\alpha}, \infty)$
   El p-valor es $1 - P(T \leq t_{obs})$
3. Para una prueba de hipótesis unilateral izquierda:
   $H_0 : \beta_0 = \beta_{00}$ contra $H_1 : \beta_0 < \beta_{00}$
   La región de rechazo es: $RR = (-\infty, -t_{n-2;\alpha})$
   El p-valor es $P(T \leq t_{obs})$

- Para β_1:

$$T = \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{s} \sim t_{n-2}$$

El error estándar es:

$$e.e(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Se tiene:

1. Para una prueba de hipótesis bilateral:
   $H_0 : \beta_1 = \beta_{10}$ contra $H_1 : \beta_1 \neq \beta_{10}$
   La región de rechazo es: $RR = (-\infty, -t_{n-2;\alpha/2}) \cup (t_{n-2;\alpha/2}, \infty)$
   El p-valor es $2 - 2P(T \leq |t_{obs}|)$
2. Para una prueba de hipótesis unilateral derecha:
   $H_0 : \beta_1 = \beta_{10}$ contra $H_1 : \beta_1 > \beta_{10}$
   La región de rechazo es: $RR = (t_{n-2;\alpha}, \infty)$
   El p-valor es $1 - P(T \leq t_{obs})$
3. Para una prueba de hipótesis unilateral izquierda:
   $H_0 : \beta_1 = \beta_{10}$ contra $H_1 : \beta_1 < \beta_{10}$
   La región de rechazo es: $RR = (-\infty, -t_{n-2;\alpha})$
   El p-valor es $P(T \leq t_{obs})$

Análisis de Varianza:

Tenemos el modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ y las hipótesis:
$H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$
El estadístico de prueba bajo H_0 es:

$$F = \frac{R^2(n-2)}{1-R^2} \sim F_{1,\,n-2}$$

La región de rechazo es $RR = (f_{1,n-2;\alpha},\, \infty)$ y el p-valor es: $1 - P(F \leq f_{obs})$

Ejemplo:

Ajustar una recta de mínimos cuadrados a los datos que se presentan a continuación los cuales representan al precio unitario de un determinado producto en 5 supermercados distintos y el número de unidades vendidas del producto.

Precio unitario	277.1	279.3	281.4	283.2	284.8
Número de unidades vendidas	52	51	50	49	48

Solución:

El número de unidades vendidas es la variable dependiente (y) y el precio unitario es la variable independiente (x).

$> x < -c(277.1, 279.3, 281.4, 283.2, 284.8)$
$> y < -c(52, 51, 50, 49, 48)$
$> ajuste < -lm(y \sim x)$
$> summary(ajuste)$

Muestra los siguientes resultados:
Residuales: -0.09446, 0.04047, 0.05239, -0.12221
$\hat{\beta}_0 = 195.0440$
$\hat{\beta}_1 = -0.5159$
$e.e(\hat{\beta}_0) = 5.5398$
$e.e(\hat{\beta}_1) = 0.0197$
$t.obs = 35.21$ (para β_0)
$t.obs = -26.18$ (para β_1)
$p - valor = 0.0000504$ (para β_0)

$p - valor = 0.000122$ (para β_1)
0.1205 (error estándar residual)
$R^2 = 0.9956$
$f_{obs} = 685.6$ con 1 y 7 grados de libertad.
$p - valor = 0.0001222$ (para la prueba F)

El modelo lineal ajustado es $y = 195.0440 - 0.5159$
El coeficiente de determinación indica que el 99.56% de la variación en el nmero de unidades vendidas del producto, se debe a diferencias en el precio del producto.

Por otra parte, si queremos hallar el coeficiente de determinación y estimar por intervalos de confianza del 95% los coeficiente de regresión, hacemos:

$> n < -length(x)$
$> coef < -ajuste\$coefficients$
$> B0 < -coef[1]$
$> B1 < -coef[2]$
$> xbarra < -mean(x)$
$> ybarra < -mean(y)$
$> x.barra < -rep(xbarra, n)$
$> y.barra < -rep(ybarra, n)$
$> SSxx < -sum((x - x.barra) \wedge 2)$
$> SSyy < -sum((y - ybarra) \wedge 2)$
$> SSxy < -sum((x - x.barra) * (y - y.barra))$
$> b0 < -rep(B0, n)$
$> b1 < -rep(B1, n)$
$> S.cuadrado < -sum((y - b0 - b1 * x) \wedge 2)/(n - 2)$
$> R < -SSxy/sqrt(SSxx * SSyy)$
$R.cuadrado < -R \wedge 2$

Es el coeficiente de determinación.

Intervalo de confianza para β_0:

$> alpha < -0.05$
$> t.alphamedio < -qt(1 - alpha/2, n - 2)$
$> lim.inf1 < -B0 - t.alphamedio * sqrt(S.cuadrado * (1/n + xbarra/SSxx))$

$> lim.sup1 < -B0+t.alphamedio*sqrt(S.cuadrado*(1/n+xbarra/SSxx))$
$> Intervalo.B0 < -c(lim.inf1, lim.sup1)$
$> Intervalo.B0$

Intervalo de confianza para β_1:

$> alpha < -0.05$
$> t.alphamedio < -qt(1 - alpha/2, n - 2)$
$> lim.inf2 < -B1 - t.alphamedio * sqrt(S.cuadrado/SSxx)$
$> lim.sup2 < -B1 + t.alphamedio * sqrt(S.cuadrado/SSxx)$
$> Intervalo.B1 < -c(lim.inf2, lim.sup2)$
$> Intervalo.B1$

Prueba de hipótesis unilateral derecha para β_1 al nivel de significación del 5%:

$H_0 : \beta_1 \leq 0$, contra $H_1 : \beta_1 > 0$
$> alpha < -0.05$
$> t.obs < -B1 * sqrt(SSxx/S.cuadrado)$
$> t.obs$
$> t.alpha < -qt(1 - alpha, n - 2)$
$> t.alpha$
$> p.valor < -1 - pt(t.obs, n - 2)$
$> p.valor$

Gráficas de los datos:

1. Diagrama de dispersión de los datos con recta de regresión:

   $> plot(x, y, \text{main='Diagrama de dispersión',xlab='Periodo (días)',ylab='Talla(cm.)'})$

   $> abline(lm(y \sim x))$

2. Matriz de correlación:

   $> M = cbind(x, y)$

   $> pairs(M, main = \text{'Matriz de correlación'})$

3. Histograma de residuales:

   $> hist(resid(ajuste), main =' \text{Histograma de residuales}', xlab =' Residuales', ylab =$ 'Frecuencia absoluta')

4. Gráfica de normalidad de los residuos:

   $> qqnorm(resid(ajuste), main =$ 'Gráfica de normalidad de los residuos')

   $> qqline(resid(ajuste))$

   Nota:

   Cuando el modelo es correcto, los residuos estandarizados caen entre -2 y 2.

5. Gráfica de independencia de los residuos:

   $> plot(ajuste\$fitted.values, resid(ajuste))$

   $> lines(lowess(ajuste\$fitted.values, resid(ajuste)))$

   $> abline(h = 0)$

Predicción:

Queremos predecir el número de unidades vendidas del producto cuando este cuesta Bs.280.

$> x < -280$
$> y < -B0 + B1 * x$
$> y$
[1]50.59842

Entonces, para un producto que cuesta Bs 280 se venderán 51 unidades de este.

Transformación para linealidad:

Una de las hipótesis habituales en análisis de regresión es que el modelo que describe los datos es lineal. A partir de consideraciones teóricas o de un examen del diagrama de dispersión de y contra x, la relación puede aparecer

no lineal. En estos casos, es posible hacer una transformación que produzca un modelo lineal.

- $y = \alpha x^{\beta}$, se transforma en: $y' = log(y),\ x' = log(x)$, y su forma lineal es $y' = log(\alpha) + \beta x'$
- $y = \alpha x^{\beta x}$, se transforma en: $y' = ln(y)$, y su forma lineal es $y' = ln(\alpha) + \beta x$
- $\alpha + \beta log(x)$, se transforma en: $x' = log(x)$, y su forma lineal es $y' = \alpha + \beta x'$
- $y = \frac{x}{\alpha x - \beta}$, se transforma en: $y' = \frac{1}{y}$, $x' = \frac{1}{x}$, y su forma lineal es $y' = \alpha - \beta x'$
- $y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$, se transforma en: $y' = ln(\frac{y}{1-y})$, y su forma lineal es $y' = \alpha - \beta x'$