Probabilidad y Estadística - 300MAE005
Profesor Daniel Enrique González Gómez

TALLER 11.1 REGRESION LINEAL

Con el fin de estimar la relación que exite entre el nivel de ingresos de una persona y el nivel de consumo , un investigador recolectó la siguiente información:

id	1	2	3	4	5	6
Ing	24.3	12.5	31.2	28.0	35.1	10.5
Con	16.2	8.5	15.0	17.0	24.2	11.2

id	7	8	9	10	11	12
Ing	23.2	10.0	835	15.9	14.7	15.0
Con	15.0	7.1	3.5	11.5	10.7	9.2

Donde :
Id: identificador del hogar
Ing : Ingresos familiares
Con : Consumo en viveres

Realice un analisis de regresión que permita estimar la relación entre el consumo (Con) y los ingresos familiares (Ing).

id	x	y	x2	y2	xy
1	23.4	16.2	590.49	262.44	393.66
2	12.5	8.5	156.25	72.25	106.25
3	31.2	15.0	973.44	225.00	468.00
4	28.0	17.0	784.00	289.00	476.00
5	35.1	24.2	1232.01	585.64	849.42
6	10.5	11.2	110.25	125.44	117.60
7	23.2	15.0	538.24	225.00	348.00
8	10.0	7.1	100.00	50.41	71.00
9	8.5	3.5	72.25	12.25	29.75
10	15.9	11.5	252.81	132.25	182.85
11	14.7	10.7	216.09	114.49	157.29
12	15.0	9.2	225.00	84.64	138.00
suma	228.90	149.10	5250.83	2178.81	3337.82

Resumen de formulas

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

$$\widehat{E[Y_i|X_i]} = \widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i = b_0 + b_1 x_i \tag{2}$$

$$e_i = \widehat{u_i} = y_i - \widehat{y_i} \tag{3}$$

$$SCE = \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{4}$$

Método de Mínimos Cuadrados Ordinarios (MCO)

Objetivo : Encontrar los valores de β_0 y β_1 que minimecen la Suma de Cuadrado de los Erroeres (SCE).

Ecuaciones normales:

$$\frac{\partial SCE}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{5}$$

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

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{6}$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_1^2 = \sum_{i=1}^{n} x_i y_i$$

Solución :

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\frac{149.10}{12} - 0.55817\frac{228.90}{12} = 1.77788$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}$$

$$S_{xx} = 5280.83 - \frac{(228.90)^2}{12} = 884.5625$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$S_{yy} = 2178.81 - \frac{(149.0)^2}{12} = 326.2425$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{n}$$

$$S_{xy} = 3337.82 - \frac{228.90 \times 149.10}{12} = 493.7375$$

Suma de Cuadrados de los Errores (Residuales)

$$SCE = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = S_{yy} - b_1 S_{xy}$$

$$SCE = 326.2425 - 0.55817 \times 493.7375 = 50.65304$$

$$s^2 = \frac{SCE}{n-2} = \frac{Syy - b_1 S_{xy}}{n-2} =$$

$$IC_{\beta_1} : b_1 t_{v=n-2;\alpha/2} \frac{s}{\sqrt{S_{xx}}}$$

$$T_0 = \frac{b_1 - \beta_{1o}}{s/\sqrt{S_{xx}}}$$

Pruebas de hipotesis individuales
Prueba de hipotesis individual para el intercepto

$H_0 : \beta_0 = 0$
$H_a : \beta_0 \neq 0$
Estadístico de prueba

$$t = \frac{b_0 - 0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}} = \frac{1.77788 - 0}{\sqrt{5.065}\sqrt{\frac{5250.83}{12 \times 884.5625}}} = 1.123$$

Valor-p

```
> 2*pt(-1.123,11)
[1] 0.2853501
```

Prueba de hipotesis individual para la pendiente

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
Estadistico de prueba:

$$t = \frac{b_1 - 0}{\frac{s}{\sqrt{S_{xx}}}} = \frac{0.55817}{\sqrt{5.0652/884.5625}} = 7.376$$

```
> 2*pt(-7.376,11)
0.0000
```

Intervalo de confianza para la media $\mu_{Y|x_0}$

$$\widehat{y_0} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$v = n - k$ grados de libertad
Intervalo de confianza para predicción para y_0

$$\widehat{y_0} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$v = n - k$ grados de libertad

Códigos en R

```
x=c(24.3,12.5,31.2,28,35.1,10.5,23.2,
    10,8.5,15.9,14.7,15)
y=c(16.2,8.5,15,17,24.2,11.2,15,7.1,
    3.5,11.5,10.7,9.2)
regresion=lm(y ~ x)
summary(regresion)
plot(x,y, xlab = "Ingresos", ylab = "Consumo")
abline(regresion)
```

ANOVA (Análisis de Varianza)

Fuentes de variación	grados de libertad	Suma de Cuadrados	Cuadrados Medios	F	Valor-p
Regresión	$k-1$	$b_1 S_{xy}$	$\frac{SCReg}{(k-1)}$	$\frac{CMReg}{CMRes}$	$1 - pf$
	$2-1$	275.590	275.590	$\frac{275.590}{5.0652} = 54.408$	0.0000
Residuales	$n-k$	$S_{yy} - b_1 S_{xy}$	$\frac{SCRes}{(n-k)}$		
	$12-1$	50.652	5.0652		
Total	$n-1$	S_{yy}			
	$12-1$	326.2425			