1

Auxiliar Estadística (The last One)

Prof: Rodrigo Assar (Sección 3) (Otoño - 2008)
P. Aux: Andrés Iturriaga, Victor Riquelme

Tema: Regresión lineal.

Problema 1 El siguiente problema está relacionado a la producción de Biomasa de una planta (masa total de materia viva seca de una planta) y la radiación solar. A continuación, la tabla de datos:

Obs	Radiación solar (X)	Biomasa (Y)
1	28,7	16.6
2	68,4	48.1
3	120,7	121,7
4	217,2	219,6
5	313,5	374,5
6	419,1	570,8
7	535,9	648,2
8	641,5	758,6

(a) Calcule $\hat{\beta}_0$ y $\hat{\beta}_1$ para la regresión lineal de Biomasa de una planta en función de la radiación solar (X). Escriba la ecuación de regresión e interprete.

Sol: La ecuación de regresión es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0,\sigma^2) \quad , \quad i=1,\dots,n \qquad (n=8)$$

Por el método de mínimos cuadrados, encontramos $\hat{\beta}_0$ y $\hat{\beta}_1$:

Sea $$f(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

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

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \mapsto \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

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

$$\underbrace{\sum_{i=1}^{n} x_i y_i}_{xy} - \beta_0 \underbrace{\sum_{i=1}^{n} x_i}_{xe} - \beta_1 \underbrace{\sum_{i=1}^{n} x_i^2}_{xx} = 0$$

Tenemos que $xy - \hat\beta_0 xe - \hat\beta_1 x \cdot x = 0$; $\hat\beta_0 = \bar y - \hat\beta_1 \bar x$ $\quad e^t = (1, \ldots, 1)$

$$xy - (\bar y - \hat\beta_1 \bar x)\, xe - \hat\beta_1 xx = 0$$
$$xy - \bar y\, xe + \hat\beta_1 \bar x\, xe - \hat\beta_1 xx = 0$$
$$xy - \bar y\, xe = \hat\beta_1 (xx - \bar x\, xe)$$
$$\hat\beta_1 = \frac{xy - \bar y\, xe}{xx - \bar x\, xe} = \frac{\frac{1}{n} xy - \bar y \bar x}{\frac{1}{n} xx - \bar x \bar x} = \frac{cov_{x,y}}{S_x^2}$$

$$\boxed{\hat\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar x \bar y}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar x^2}}$$

$$\boxed{\hat\beta_0 = \bar y - \frac{cov_{x,y}}{S_x^2}\,\bar x}$$

Numericamente, con los datos dados, se tiene que

$\hat\beta_1 = 1{,}269$; $\hat\beta_0 = -27{,}677$ $\quad$ Si se aumenta 1 en Radiación, aumenta 1,269 en Biomasa.

(b) Construya intervalos de confianza de un 95% para β_0 y β_1. Interprete los resultados.

Dem Recordemos que los estimadores son insesgados

$$\mathbb{E}(\hat\beta_i) = \beta_i$$

Obs: Como $\varepsilon_i \sim N(0, \sigma^2)$, $\hat\beta_i \sim N(\beta_i, Var(\hat\beta_i))$ $\qquad (Var(\hat\beta) = \sigma^2 (X^tX)^{-1})$

$$Var(\hat\beta_0) = \frac{\sum x_i^2\, \sigma^2}{n\sum x_i^2 - n^2 \bar x^2} \quad ; \quad Var(\hat\beta_1) = \frac{n\, \sigma^2}{n\sum x_i^2 - n^2 \bar x^2}$$

$$Var(\hat\beta_0) = \frac{1}{n}\, \frac{\frac{1}{n}\sum x_i^2\, \sigma^2}{S_x^2} \quad ; \quad Var(\hat\beta_1) = \frac{1}{n}\, \frac{\sigma^2}{S_x^2}$$

En otra forma: $\hat\beta_i \sim N(\beta_i, \sigma^2 (X^tX)^{-1}_{ii})$.

$$\Rightarrow \frac{\hat\beta_i - \beta_i}{\sigma \sqrt{(X^tX)^{-1}_{ii}}} \sim N(0,1)$$

Un estimador para σ^2 es (del método de max. verosimilitud)

$$\hat\sigma^2 = \frac{1}{n}\sum_{j=1}^n (y_j - \hat\beta_0 - \hat\beta_1 x_j)^2 \qquad \left(= \frac{1}{n}(y - \hat y)^t (y - \hat y)\right)$$

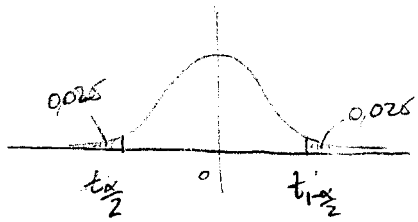
Est. sesgado $\hat\sigma^2$.

$\tilde\sigma^2 = \frac{n}{n-(p+1)}\hat\sigma^2$ insesgado

$$\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)}{n}\sum_{j=1}^{n}\frac{(y_j-\hat{\beta}_0-\hat{\beta}_1 x_j)^2}{\sigma^2} \sim \chi^2_{n-2}$$

Entonces, $$T = \frac{\dfrac{\hat{\beta}_i-\beta_i}{\sigma\sqrt{(X^tX)^{-1}_{ii}}}}{\sqrt{\dfrac{(n-2)\,\hat{\sigma}^2}{\sigma^2\,(n-2)}}} = \frac{\hat{\beta}_i-\beta_i}{\underbrace{\hat{\sigma}\sqrt{(X^tX)^{-1}_{ii}}}_{\hat{\sigma}_i}} \sim t_{n-2} \qquad i=0,1$$

t-student de $(n-2)$ grados de libertad



$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} < \frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}_i} < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i < \hat{\beta}_i-\beta_i < t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i\right) = 1-\alpha$$

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i-\hat{\beta}_i < -\beta_i < t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i-\hat{\beta}_i\right) = 1-\alpha$$

$$\mathbb{P}\left(\hat{\beta}_i-t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i < \beta_i < \hat{\beta}_i+t_{1-\frac{\alpha}{2}}\,\hat{\sigma}_i\right) = 1-\alpha$$

Para la t-student con $(n-2)=(8-2)=6$ grados de libertad, tenemos que $(\alpha=0{,}05)$

$$t_{1-0{,}025} = 2{,}447$$

Ahora, $$\hat{\sigma}_0 = \hat{\sigma}\sqrt{(X^tX)^{-1}_{00}} = \sqrt{\frac{(Y-\hat{Y})^t(Y-\hat{Y})}{n}}\sqrt{\frac{1}{n}\,\frac{\frac{1}{n}\sum x_i^2}{S_x^2}} = 23{,}533\cdot 0{,}61$$

$$\hat{\sigma}_1 = \hat{\sigma}\sqrt{(X^tX)^{-1}_{11}} = \underbrace{\sqrt{\frac{(Y-\hat{Y})^t(Y-\hat{Y})}{n}}}_{23{,}533}\sqrt{\frac{1}{n}\,\frac{1}{S_x^2}} = 23{,}533\cdot 0{,}002$$

$\hat{\sigma}_0 = 14{,}3$; $\hat{\sigma}_1 = 0{,}04$. //

Los intervalos de confianza quedan

para β_0: $(-27{,}677-2{,}447\cdot 14{,}3;\ -27{,}677+2{,}447\cdot 14{,}3) = (-62{,}669;\ 7{,}315)$

para β_1: $(1{,}269-2{,}447\cdot 0{,}04;\ 1{,}269+2{,}447\cdot 0{,}04) = (1{,}171;\ 1{,}367)$

(c) Pruebe la significancia de los parametros: considere

(i) * $H_0: \beta_1 = \beta_2 = 0$ vs H_1: alguno no es nulo.

(ii) * $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

(iii) * $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$.

Sol. Primero, para contrastar $H_0: \beta_i = 0$ vs $\beta_i \neq 0$, usamos la t-student

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i} \sim t_6 \quad ; \text{bajo } H_0,\ \beta_i = 0. \qquad \#\text{ grados libertad} = n - (p+1)$$

(*) $\dfrac{\hat{\beta}_0}{\hat{\sigma}_0} = -1,835 \rightsquigarrow \mathbb{P}(|t_6| > 1,835) > \mathbb{P}(|t_6| > 2,447) = 0,05$ No rechazo $\beta_0 = 0$

(*) $\dfrac{\hat{\beta}_1}{\hat{\sigma}_1} = 31,725 \rightsquigarrow \mathbb{P}(|t_6| > 31,725) < \mathbb{P}(|t_6| > 2,447) = 0,05$ Rechazo $\beta_1 = 0$

Ahora, para el modelo global, comparamos los residuos aportados por la regresión

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^{n} \underbrace{(y_i - \hat{y}_i)^2}_{\varepsilon_i^2} / (n-(p+1))} \overset{\text{Bajo } H_0}{\sim} F_{(p+1),\, n-(p+1)}$$

→ varianza explicada / gl

→ varianza residual / gl

$$= \frac{R^2 / p}{(1-R^2)/(n-(p+1))}, \quad \text{con } R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$ coef. de correlación múltiple.

$F = 264,58$

$\mathbb{P}(F_{2,6} > 5,14) = 0,05$

$\Rightarrow \mathbb{P}(F_{2,6} > 264,58) \approx 0,00$

Entonces, algún coeficiente es significativo ($\neq 0$).

- Var. Residual: $\frac{1}{n} \sum \varepsilon_i^2 = \frac{1}{n} \sum (y_i - \hat{y}_i)^2 \to SSE$ $(n-(p+1) = n-r)$ SCTotal = SSE + SSR
- Var. Explicada: $\frac{1}{n} \sum (\hat{y}_i - \bar{y})^2$
- Var. Total: $\frac{1}{n} \sum (y_i - \bar{y})^2 \to SSR$ $(p = r-1)$

Problema 2 El ministerio de educación quiere estudiar de qué depende el gasto anual en educación de un hogar, para ello, recolecta información en 100 hogares y plantea el modelo lineal $\mathbb{E}(y)=\beta_0+\beta_1x^1+\beta_2x^2+\beta_3x^3+\beta_4x^4$ (1)

con x^1 = ingreso del hogar, x^2 = # hijos ; x^3 = talla del jefe de hogar, x^4 = # personas en casa.

(i) Complete los resultados de la regresión lineal (1) dados en las tablas 1 y 2. Interprete los resultados.

TABLA 1

Variable	Estimación		Desviación típica		t-student	p-valor	
cte	20,387	:	20,384	=	1,000	0,319	→ No Rechazo $\beta=0$
Ingreso	0,189	:	$\boxed{0,021}$	=	9,242	0,000	→ Rechazo $\beta=0$
Nº hijos	17,379	:	2,978	=	5,836	0,000	→ Rechazo $\beta=0$
Talla jefe	8,869	:	6,176	=	$\boxed{1,436}$	0,154	→ No Rechazo $\beta=0$
Nº personas	$\boxed{0,187}$		0,107		1,749	0,083	→ No Rechazo $\beta=0$

TABLA 2

Fuente	G. Libertad	Suma Cuadrados	F	p-valor
Regresión	$\boxed{4}$	129489,083	38,165	0,000
Residuos	95	80539,635		
Total	99	$\boxed{210028,718}$		

→ Hay algún coeficiente significativo

$R = 0,786$; $\hat{\sigma}_{\text{insesgada}} = 29,12$ → El modelo es bueno (el R es alto)

$$F=\frac{SC\ Reg/GL\ Reg}{SC\ Residuos/GL\ Residuos}$$

Se pueden sacar del modelo x^3 y x^4.

(ii) Se plantea un modelo con el ingreso y el # de hijos:

$$\mathbb{E}(y)=\beta_0+\beta_1x^1+\beta_2x^2 \quad (2)$$

Se propone resolver el test H_0: modelo (2) vs H_1: modelo (1). Para ello se resuelve el modelo (2) obteniendose las tablas (3) y (4). Comente el cambio en la suma de los cuadrados de los residuos de los modelos (2) y (1).

Var	Est.	Des. Std	t-St	P-valor
cte	54,035	8,575	6,301	0,000
Ingreso	0,196	0,019	10,019	0,000
Nº hijos	17,804	2,969	5,995	0,000

Fuente	GL.	SC.	F	p-valor
Regresión	2	125292,851	71,71	0,000
Residuos	97	84735,867		
Total	99	210028,718		

$R = 0,772$, $\hat{\sigma}_{\text{insesgado}} = 29,56$

Sol El cambio es pequeño entre los R. Este modelo explica así el gasto anual (en relación con el modelo anterior.

En resumen

(*) Test t-student: decide si el coeficiente β_j es significativo; para ello se plantea el test $\frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim t_{n-(p+1)}$ bajo $H_0: \beta_j = 0$.

p-valor $< 0{,}05$ rechaza H_0 (el coef. es significativo en el modelo, o sea, la variable correspondiente debe ser considerada)

(*) Test Fisher: decide si el modelo es explicativo (si alguno de los β_j es distinto de 0)

plantea el test $F = \frac{R^2/p}{(1-R^2)/n-(p+1)} = \frac{\text{Suma Cuadrados Regresión}/p}{\text{Suma Cuadrados Residuos}/n-(p+1)} \sim F_{p,n-(p+1)}$

(bajo H_0: todos $\beta_j = 0$)

p-valor $< 0{,}05$ rechaza H_0 (alguna variable explica a la variable Y)

(*) R^2 cercano a 1 $\Rightarrow$ hay un buen ajuste del modelo hacia la variable explicada.
R^2 cercano a 0 $\Rightarrow$ modelo malo

(*) Modelo $Y = X\beta + \varepsilon$ con $\varepsilon \sim N_n(0, \sigma^2)$, σ^2 desconocida

$$\hat{\beta} = (X^tX)^{-1}X^ty \quad ; \quad \hat{y} = \mathbb{E}(y) = X(X^tX)^{-1}X^ty$$

$$\mathbb{E}(\hat{\beta}) = \beta \quad ; \quad \mathbb{V}ar(\hat{\beta}) = \sigma^2 (X^tX)^{-1}$$

Est. Max. Vero: $\hat{\sigma}^2 = \frac{1}{n}(y-\hat{y})^t(y-\hat{y})$ sesgado ; insesgado: $\tilde{\sigma}^2 = \frac{n}{n-(p+1)}\hat{\sigma}^2$