Pauta Control N°3 MA34B-3 Estadística, Semestre Otoño 2006

Profesor Alexis Peña
Auxiliares Natalia Rodriguez, Diego Díaz

1.- Suponga el siguiente modelo para una regresión lineal: $Y_i = \mu X_i + \varepsilon_i$, denominado *Regresión por el Origen*.

a) Enuncie 3 condiciones que deben cumplir los errores del modelo (ε_i), los parámetros (μ_i) y X_i, Y_i para que el parámetro (*pendiente*) de este modelo se pueda estimar por medio del método de los mínimos cuadrados.
Solución:
Algunos supuestos que pueden poner:

- $\varepsilon_i \sim N(0, \sigma^2)$
- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$
- $X_i \perp X_j, \;\; i \neq j$
- $Y_i \perp Y_j, \;\; i \neq j$

b) Encuentre la estimación de μ mediante el método de los mínimos cuadrados encontrando las ecuaciones normales.
Solución:
Minimizar el error dado por $\sum(\varepsilon_i)^2 = \sum(Y_i - X_i\mu)^2$ o también minimizar $\varepsilon_i = \sum(\hat{Y}_i - X_i\mu)^2$. Así:

$$\frac{\partial \sum \varepsilon^2}{\partial \mu} = \frac{\partial \sum(\hat{Y}_i - X_i\mu)^2}{\partial \mu} = 0$$

Así

$$\frac{\partial \sum(\hat{Y}_i - X_i\mu)^2}{\partial \mu} = 2\sum X_i(Y_i - \mu X_i) = 0$$

Con esto:

$$\hat{\mu} = \frac{\sum(X_i Y_i)}{\sum(X_i)^2}$$

c) Encuentre la esperanza y varianza de $\hat{\mu}$. Es sesgado?.
Solución:

$$E(\hat{\mu}) = \frac{1}{\sum (X_i)^2} E(\sum X_i Y_i) = \frac{1}{\sum (X_i)^2} \sum X_i E(Y_i) = \frac{1}{\sum (X_i)^2} \sum X_i \hat{Y}_i$$

Como $Y_i = \hat{\mu} X_i$ entonces:

$$E(\hat{\mu}) = \frac{\sum X_i^2 \mu}{\sum (X_i)^2} = \mu$$

Luego es insesgado. Ahora la varianza:

$$Var(\hat{\mu}) = \frac{Var(\sum X_i Y_i)}{(\sum X_i^2)^2} = \frac{1}{(\sum X_i^2)^2} Var(X_1 Y_1 + X_2 Y_2 + \ldots + X_n Y_n)$$

$$= \frac{1}{(\sum X_i^2)^2} \sum X_i^2 Var(Y_i) = \frac{\sigma^2 (\sum X_i^2)}{(\sum X_i^2)^2} = \frac{\sigma^2}{\sum X_i^2}$$

d) Encuentre la varianza de una predicción dado que se cumplen los supuestos de a).
Solución

$$Var(\hat{Y}_0 - Y_0) = Var(\hat{Y}_0) + Var(Y_0) = X_0^2 \frac{\sigma^2}{\sum X_i^2} + \sigma^2$$

Hints: Recuerde que para encontrar las ecuaciones normales se debe minimizar la suma de los errores.

2.- La diferencia $Y_0 - \hat{Y}_0$ se denomina *predicción del error*, donde Y_0 es una predicción. Si el modelo ajusta bien los datos - esto ocurre cuando cumple con las suposiciones para los errores - entonces la varianza de la predicción $Var(Y_{pred0})$ se puede calcular como la varianza de la diferencia entre $\hat{Y}_0$ y la futura observación Y_0.

a) Demuestre que dicha varianza es:

$$Var(Y_{pred0}) = \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum {x_i}^2}\right] \sigma^2$$

Solución:

$Var(Y_{\text{pred0}}) = Var(\hat{Y_0} - Y_0) = Var(\hat{Y_0}) + Var(Y_0) = \sigma^2 + Var(\hat{Y_0})$

Usando que $\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$ entonces

$$\hat{Y_0} = \hat{\beta_0} + \hat{\beta_1}X_0 = \bar{Y} - \hat{\beta_1}\bar{X} + \hat{\beta_1}X_0 = \bar{Y} + \hat{\beta_1}(X_0 - \bar{X})$$

Luego aplicando la varianza a esta última ecuación y usando que $Cov(\bar{Y}, \beta_1) = 0$ se tiene:

$$Var(\hat{Y_0}) = Var(\bar{Y}) + (X_0 - \bar{X})^2 Var(\hat{\beta_1})$$

Luego se calcula $Var(\bar{Y})$ sabiendo que $Y_i \perp Y_j \;\; i \neq j$, así

$$Var(\bar{Y}) = \frac{1}{n^2} Var(\sum Y_i) = \frac{1}{n^2} \sum Var(Y_i) = \frac{1}{n^2} \sum \sigma^2 = \frac{\sigma^2}{n}$$

Con lo que la $Var(\hat{Y_0})$ queda:

$$Var(\hat{Y_0}) = \frac{\sigma^2}{n} + (X_0 - \bar{X})^2 Var(\hat{\beta_1})$$

La varianza de β_1 fue vista en clases y es $Var(\hat{\beta_1}) = \frac{\sigma^2}{\sum x_i^2}$ Con lo que se llega a la varianza pedida:

$$Var(Y_{\text{pred0}}) = \sigma^2 + \frac{\sigma^2}{n} + (X_0 - \bar{X})^2 \frac{\sigma^2}{\sum x_i^2}$$

$$= \sigma^2 \Big[1 + \frac{1}{n} + (X_0 - \bar{X})^2 \frac{1}{\sum x_i^2} \Big]$$

Notar que $\sum x_i^2 = \sum (X_i - \bar{X})^2$

b) A partir de a) encuentre una expresión para la media de q futuras observaciones del valor X. (ELIMINADA)

3.- Un grupo de biólogos investiga en Coquimbo el ciclo reproductivo de un gastrópodo de la zona. Buscan relacionar la cantidad de huevos (cápsulas ovígeras; Y_i) en la hembras con la longitud de sus conchas (X_i). Para ello se han investigado grupos de estos moluscos en las costas de la región y obtenido una tabla. El modelo propuesto para dicha relación es: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Suponiendo que se cumplen los supuestos para una *regresión lineal simple* dichos científicos mediante el uso de *r-project* han obtenido el siguiente output:

```
Coefficients:
            |Estimate| Std. Error| t value| Pr(>|t|)
(Intercept) | 3.46312|    0.21119|   16.40| 5.19e-08 ***
NC          | 0.79656|    0.00577|  138.06| 2.79e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2505 on 9 degrees of freedom
Multiple R-Squared: 0.9995,     Adjusted R-squared: 0.9995
F-statistic: 1.906e+04 on 1 and 9 DF,  p-value: 2.789e-16
```

De acuerdo a estos datos, responda:

a) Indique cuánto valen $\hat{\beta}_0$ y $\hat{\beta}_1$ y escriba la ecuación del modelo.
Solución
$\hat{\beta}_0$ = intercept = itercept = 3,46312 y $\hat{\beta}_1$ = pendiente = NC = 0,79656 Luego el modelo queda

$$Y_i = 3{,}46312 + 0{,}79656 X_i + \varepsilon_i$$

O también

$$\hat{Y}_i = 3{,}46312 + 0{,}79656 X_i$$

b) Encuentre un IC para β_0 y β_1 con un 95 % de confianza.

Grados de libertad = n - p = 11 - 2 = 9. Luego:

$$IC[0{,}025,\ \ 9,\ \ \beta_0] = \pm\tau_{(0{,}025,\ 9)} S(\hat{\beta}_0) + \beta_0$$

Viendo la tabla de datos entregada se tiene que: $\hat{\beta}_0 = 3{,}46312$ y $S(\hat{\beta}_0) = 0{,}21119$ luego:

$$\pm\tau_{(0{,}025,\ 9)} S(\hat{\beta}_0) + \beta_0 = (2{,}262)(0{,}21119) \pm 3{,}46312$$

$$= [2{,}985\ ,\ 3{,}940]$$

$$IC[0{,}025\ \ 9\ \ \beta_1] = \pm\tau_{(0{,}025\ \ 9)} S(\hat{\beta}_1) + \beta_1$$

Viendo la tabla de datos entregada se tiene que: $\hat{\beta}_1 = 0{,}79656$ y $S(\hat{\beta}_1) = 0{,}00577$ luego:

$$= \pm(2{,}262)(0{,}00577) + 0{,}79656$$

$$= [0{,}7835\ ,\ 0{,}809]$$

c) Haga un test de significancia para saber si el modelo pudiese ser de la forma $Y_i = \beta_1 X_i + \varepsilon_i$.
Solución de 4 formas distintas (cualquiera es igualmente válida y equivalente, pueden usar cualquiera)

1) Por la tabla estadística t-student y usando el pivote de una normal/chi cuadrado, se tiene:

$$H_0 : \ \ \beta_0 = 0$$
$$H_1 : \ \ \beta_1 \neq 0$$
$$\tau_{obs} = \frac{\beta_0 - 0}{S(\hat{\beta_0})} = 16{,}398 \quad \tau_{\text{teórico}} = 2{,}262$$

Luego $\tau_{\text{obs}} > \tau$teórico $\Rightarrow$ Rechazo hipótesis nula.

2) Por los datos de la tabla entregada. La tabla entrega un t value $=$ 16,40 y la probabilidad de cometer error tipo I es $pr(> |t|) = 5{,}19e-08$ con significancia de 0, con lo que se rechaza la hipótesis nula pues $\alpha_{\text{observado}} << \alpha_{\text{teórico}}$.

3) Usando la estadística de Fisher con los datos que entrega el programa. Debe usarse que:

$$F_{(GLss_{reg},\ GLss_{res})} \sim \frac{SS_{reg}}{SS_{res}}$$

Viendo los datos entregados por el programa se ve que la estadística de Fisher observada es $F_{1,9} = 1906$ y el teórico es (viendo una tabla F fisher) $F_{1,9} = 5{,}12$. Luego $F_{\text{observado}} >> F_{\text{teórico}}$ Con lo que se rechaza la hipótesis nula.

4) También puede verse con el p-valor (mínima probabilidad de cometer error tipo I) de los datos entregados, $p-value = 2{,}789e-16 \sim 0$ Luego se rechaza la hipótesis nula.

d) Si al ingresar los datos al *r-project* uste d obtiene el siguiente output (usando el modelo de la parte c), indique cuál modelo ajusta mejor los datos (¿con intercepto o sin intercepto?) y porqué.

Solución:
De acuerdo a c) se prefiere un modelo con intercepto. También puede compararse R^2_{SI} y R^2_{CI}. El R^2_{CI} es más cercano a 1 que el R^2_{SI}, por lo tanto se prefiere el modelo con intercepto.

```
Coefficients:
   Estimate| Std. Error| t value| Pr(>|t|)
NC  0.88491|    0.01088|   81.34| 1.93e-15 ***
---
Residual standard error: 1.321 on 10 degrees of freedom
Multiple R-Squared: 0.9985,     Adjusted R-squared: 0.9983
F-statistic:  6616 on 1 and 10 DF,  p-value: 1.928e-15
```