

Estadística para la Economía y la Gestión

30 de junio 2016

## Pauta Control 2

Profesor: *Franco Basso*Ayudantes: *V. León, A. Navarrete, M. Rojas*

- P1. Es sabido que la **Regresión Lineal Simple** supone que el fenómeno en estudio sigue una tendencia lineal de la forma

$$y = ax + b,$$

donde  $y$  es la variable dependiente,  $x$  la variable independiente y  $a, b$  son parámetros a determinar. Considerando que  $y_i = ax_i + b + \varepsilon_i$  (la demanda observada  $y_i$ ,  $i = 1, \dots, n$  comete un error  $\varepsilon_i$ ), estos parámetros son escogidos de forma que la suma del cuadrado de los errores  $\varepsilon_i$ , provenientes de la observación  $i$  sea mínima.

Demuestre que bajo lo anterior se tiene que:

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

$$b = \bar{y} - a \bar{x},$$

con  $\bar{x}$  y  $\bar{y}$  promedios muestrales de  $x$  e  $y$  respectivamente.

**Solución:**

Como hay que minimizar la suma de los errores al cuadrado, notando que  $\varepsilon_i = y_i - ax_i - b$ , se tiene que

$$\begin{aligned} \sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n (y_i - ax_i - b)^2 \\ &= \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i \\ &= \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i. \end{aligned}$$

De lo anterior, obtenemos un problema de minimización irrestricto en  $\mathbb{R}^2$ :

$$\min_{a, b \in \mathbb{R}} \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i.$$

Aplicamos las condiciones de primer orden para obtener:

$$\begin{aligned} \frac{\partial}{\partial a} \sum_{i=1}^n \varepsilon_i^2 &= \frac{\partial}{\partial a} \left[ \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i \right] \\ &= 2a \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i y_i + 2b \sum_{i=1}^n x_i = 0. \end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial b} \sum_{i=1}^n \varepsilon_i^2 &= \frac{\partial}{\partial b} \left[ \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i \right] \\ &= 2nb - 2 \sum_{i=1}^n y_i + 2a \sum_{i=1}^n x_i = 0.\end{aligned}$$

De la segunda ecuación, se tiene que

$$\begin{aligned}2nb - 2 \sum_{i=1}^n y_i + 2a \sum_{i=1}^n x_i &= 0 \\ nb - \sum_{i=1}^n y_i + a \sum_{i=1}^n x_i &= 0 \\ nb &= \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i \\ b &= \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i \\ b &= \bar{y} - a\bar{x},\end{aligned}$$

obteniendo la segunda condición. Reemplazando este valor en la primera ecuación, obtenemos:

$$\begin{aligned}a \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i + b \sum_{i=1}^n x_i &= 0 \\ a \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i + (\bar{y} - a\bar{x}) \sum_{i=1}^n x_i &= 0 \\ a \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i + \bar{y} \sum_{i=1}^n x_i - a\bar{x} \sum_{i=1}^n x_i &= 0 \\ a \sum_{i=1}^n x_i^2 - a\bar{x} \sum_{i=1}^n x_i &= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \\ a \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \\ a &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\end{aligned}$$

**P2.** Diversos expertos chilenos han propuesto la existencia de una relación entre el dólar observado y el precio de ciertos commodities claves para la economía chilena. Para analizar dicha relación se han recopilado datos del dólar observado [pesos/USD], precio del cobre [USD/libra] y precio del petróleo [USD/barril] mensuales desde enero 2008 hasta abril 2016. Además se ha incorporado a la base de datos el IMACEC [puntos] durante el mismo periodo. En la tabla a continuación se muestran los valores de dichas series para el año 2016:

| periodo  | imacec | dólar  | precio_cobre | Precio_petroleo |
|----------|--------|--------|--------------|-----------------|
| ene.2016 | 125.3  | 721.95 | 2.0          | 31.5            |
| feb.2016 | 120.5  | 704.08 | 2.1          | 30.4            |
| mar.2016 | 133.6  | 682.07 | 2.2          | 37.8            |
| abr.2016 | 128.1  | 669.93 | 2.2          | 41.0            |

Se ha ajustado la siguiente regresión lineal múltiple:

```
> fit <- lm(dólar~imacec+precio_cobre+Precio_petroleo, data=df)
> summary(fit)

Call:
lm(formula = dólar ~ imacec + precio_cobre + Precio_petroleo,
    data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-63.56 -19.99  -1.61   18.76   71.44

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    610.3650     33.3102   18.324 < 2e-16 ***
imacec           1.7102      0.2610    6.551 2.83e-09 ***
precio_cobre   -59.2196      7.0590   -8.389 4.23e-13 ***
Precio_petroleo -0.9385      0.2069   -4.535 1.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.2 on 96 degrees of freedom
Multiple R-squared:  0.8342,    Adjusted R-squared:  0.829
F-statistic: 161 on 3 and 96 DF, p-value: < 2.2e-16
```

- ¿Qué puede decir de la significancia de las variables explicativas y del modelo en general? ¿Le parece una bondad de ajuste aceptable?
- Calcule los residuos para el año 2016. ¿Qué es la hipótesis de homocedasticidad? Indique un procedimiento para verificarla prácticamente.
- Determine las elasticidades del tipo de cambio con respecto a los precios del cobre, el petróleo y el IMACEC. Interprete su signo y magnitud y compare entre ellas.

### Solución:

- Todas las variables del modelo son significativas por cuanto el p-valor del t-test es menor a 5% (incluso 1%). Recordemos que este t-test realiza el test de hipótesis  $H_0 : \beta_k = 0$  vs  $H_1 \neq 0$  donde  $\beta_k$  es el coeficiente asociado a la covariable  $k$ .
- Los residuos se definen  $\epsilon_i = y_i - \hat{y}_i$  donde  $y_i$  es el valor real del dólar en el periodo  $i$  (en la tabla) y  $\hat{y}_i$  es el valor predicho (a estimar utilizando el modelo).

| Periodo      | Predicción | Residuos |
|--------------|------------|----------|
| Enero 2016   | 675.22     | 46.73    |
| Febrero 2016 | 664.45     | 39.64    |
| Marzo 2016   | 670.44     | 11.63    |
| Abril 2016   | 660.63     | 9.30     |

La hipótesis de homocedasticidad indica que la varianza de los errores es constante  $E(\epsilon_i^2|X) = \sigma^2$ . Para series de tiempo, como en este caso, se puede medir correlación temporal de los errores mediante el test de Durbin-Watson.

- (c) En los modelos lineales los coeficientes de las variables pueden interpretarse como elasticidades. Así las elasticidades del tipo de cambio con respecto al precio del cobre, precio el petróleo y el IMACEC son -59.2196, -0.9385 y 1.7102 respectivamente. Esto quiere decir, por ejemplo, que el aumento de un USD en el precio del cobre implica una disminución de casi 60 USD en el precio del dolar observado. El signo de esta cifra tiene sentido pues como Chile es un gran exportador de cobre si el precio de este aumenta, entonces entran a Chile a una gran cantidad mayor de la divisa estadounidense empujando el precio a la baja. Las elasticidades pueden también calcularse de manera porcentual.

**P3.** Sea  $Y_1, Y_2, \dots, Y_n$  una secuencia de variables aleatorias normales e independientes. Es decir que la densidad marginal de cada una de ellas está dada por:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}$$

Además se sabe que  $V(Y_i) = \sigma^2$  y  $E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$  donde  $\beta_0, \beta_1$  y  $X_i$  son valores fijos (no aleatorios). Obtenga el estimador máximo verosímil para  $\beta_0$  y  $\beta_1$  en función de  $Y_i$  y  $X_i$ .

### **Solución:**

La verosimilitud viene dada por

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}}$$

$$\ln L(\beta_0, \beta_1) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_0} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

Así,

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \bar{X} \hat{\beta}_1 \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \end{aligned}$$

- P4. Se busca analizar la cantidad de premios (variable num awards) que obtienen los alumnos de cierto colegio en función de su puntaje en el examen final en matemáticas (variable math) con un puntaje entre cero y cien. Para ello se realiza una regresión de Poisson utilizando una base de 200 alumnos.

```
> fit <- glm(num_awards~math, data=df, family=poisson)
> summary(fit)

Call:
glm(formula = num_awards ~ math, family = poisson, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1853  -0.9070  -0.6001   0.3246   2.9529

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.333532    0.591261  -9.021  <2e-16 ***
math         0.086166    0.009679   8.902  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 204.02  on 198  degrees of freedom
AIC: 384.08

Number of Fisher Scoring iterations: 6
```

- (a) ¿Le parece que la nota del examen final de matemáticas influye en la cantidad de premios? Justifique.  
 (b) ¿Qué es el AIC? ¿Para qué sirve? Calcule el logaritmo de la verosimilitud.  
 (c) ¿Cuál es la probabilidad que un alumno que tiene 25 puntos en matemáticas no gane ningún premio?  
 Responda la misma pregunta para un alumno que tiene 50 y 75 puntos?

### Solución:

- (a) El test de hipótesis asociado a la única variable tiene un p-valor menor que 5% por lo que la nota del examen final de matemáticas pareciese si influir en la cantidad de premios de los alumnos.  
 (b) El AIC es una medida que permite comparar modelos de regresiones de Poisson. Está definida por  $AIC = 2k - 2 \ln L$  donde  $k$  es la cantidad de parámetros a estimar y  $L$  es la verosimilitud por lo que  $\ln L = \frac{2k - AIC}{2} = -190,04$   
 (c) La regresión de Poisson es de la forma  $P(Y_i = y) = \frac{\lambda_i^y \exp(-\lambda_i)}{y!}$  con  $\lambda_i = \lambda_0 \exp(\beta X_i)$ . De la tabla podemos ver que  $\ln \lambda_0 = -5.333532$  y  $\beta_{math} = 0.086166$ . Con esto si un alumno  $i$  obtiene  $X_i = 25$  puntos entonces  $\lambda_i = \exp(-5.333532) \cdot \exp(0.086166 \cdot 25) = 0.041611$ . Nos piden  $P(Y_i = 0) = \frac{\lambda_i^0 \exp(-\lambda_i)}{0!} = 0.959242$ . Analogamente si  $X_i = 50$  y  $X_i = 75$  las probabilidades son 0.698574 y 0.045397 respectivamente.