

## Auxiliar 14 MA34B-03

### Regresión Lineal

Se busca describir un conjunto de variables  $X_1, \dots, X_p$  llamadas variables explicativas o exógenas que influyen sobre otra variable llamada variable a explicar o endógena ( $y$ ), mediante una relación lineal del tipo

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 \dots + \beta_p x_i^p + \epsilon_i.$$

El modelo lineal obtenido de una muestra de tamaño  $n$ , normalmente no es exacto, por lo que debe considerarse el error  $\epsilon_i$  asociado al modelo para la observación  $i$ . Por otro lado, se busca minimizar los errores con el criterio de los mínimos cuadrados  $\min \sum \epsilon_i^2 = \epsilon' \epsilon$

Matricialmente tenemos que  $y = X\beta + \epsilon$  y con el criterio anterior nos queda que los coeficientes que minimizan el error al cuadrado son de la forma

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

Para el caso de un modelo simple  $y = \beta_0 + \beta_1 x$ , la estimación de los coeficientes viene

dada por: 
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

### Propiedades de $\hat{\beta}$

- El estimador  $\hat{\beta}$  es insesgado
- El estimador  $\hat{\beta}$  es consistente
- El estimador  $\hat{\beta}$  tiene mínima varianza
- La estimación de  $\hat{\sigma}^2$  obtenida por el método de máxima verosimilitud es sesgada.

- $$\hat{\sigma}^2 = \frac{\sum \epsilon_i^2}{n - r} \quad \text{con } r = \text{rango de la matriz } X = \text{nro de variables } x + 1$$

- $$\text{Var}(\hat{\beta}_j) = \hat{\sigma}^2 (X^t X)^{-1}_{jj}$$

## Calidad del modelo

Los residuos  $\varepsilon_i$  dan la calidad del ajuste para cada observación. Un índice que evita el problema de que  $\varepsilon_i$  dependa de cada observación es

$$\frac{\sum \varepsilon_i^2}{\sigma^2} \rightarrow \chi_{n-r}^2 \Rightarrow \frac{n-r}{\sigma^2} \hat{\sigma}^2 \rightarrow \chi_{n-r}^2$$

## Coefficiente de correlación múltiple

Compara la varianza explicada con la varianza total.

$R = \sqrt{R^2}$  = coeficiente de correlación lineal entre el verdadero valor con lo que estamos estimando.

$$R = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- Si  $R=0$ , el modelo es la media muestral de los valores de  $y$ .
- Si  $R$  cercano a 1, el modelo es bueno siendo los valores observados cercanos a los estimados.
- Si  $R=1$ , existe un modelo lineal que permite escribir las observaciones  $y_i$  como combinación lineal de las variables explicativas.

## Test de hipótesis

*Por un lado vimos como calcular los coeficientes del modelo y que tan cercano a una recta es con el coeficiente de correlación  $R$ . Sin embargo, para decidir si una variable aporta o no al modelo de manera significativa estadísticamente, es decir, si el coeficiente  $B$  es distinto de 0, debemos hacer un test de hipótesis global y luego general.*

## Test global

$$H_0 : \beta_1 = \beta_2 = \dots \beta_p = 0$$

Para decidir si aceptamos o rechazamos  $H_0$ , se construye el estadístico  $F$  que sigue una distribución de Fisher.

$$F = \frac{R^2 / (r-1)}{(1-R^2) / (n-r)} \rightarrow F_{(r-1)(n-r)}$$

Buscamos C tal que  $P(F_{(r-1)(n-r)} > C) = \alpha$ , Si  $F > C \Rightarrow$  rechazamos  $H_0$  ( lo que indica que existe al menos una variable que aporta al modelo)

### Test local

Con el test global probamos si existe al menos una variable que debería ir en el modelo, pero no sabemos cual. Por tanto para cada coeficiente debemos hacer el siguiente test

$$H_0 : \beta_j = 0$$

$$\text{Pero } \hat{\beta}_j \rightarrow N(\beta_j, \sigma^2 (x^t x)^{-1}_{jj}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \rightarrow t_{n-r}$$

Buscamos  $P(|t_{n-r}| > Q) \leq \alpha$ ,

Si  $t_{n-r} > Q$  para  $\alpha = 0.05$  o si  $P(|t_{n-r}| > Q) \leq 0.05$  rechazamos  $H_0$  ( lo que indica que existe al menos una variable que aporta al modelo)

### Problema 1

Consideramos un modelo de regresión lineal simple :  $y = \beta_0 + \beta_1 x + \varepsilon$

1.1) Dé la expresión de los estimadores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  de los mínimos cuadrados para  $\beta_0$  y  $\beta_1$

1.2) Se hace el cambio de variable **z=10x**. Obtenga los estimadores del nuevo modelo.

$y = \gamma_0 + \gamma_1 z + \varepsilon$  en función de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ . Compare ambos modelos.

1.3) Dé el estadístico del test  $H_0 : \beta_1 = 0$  y explique que pasa cuando se rechaza.

**Sol.:**

1.1) Los estimadores de mínimos cuadrados de un modelo de regresión lineal simples son:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

1.2) Los coeficientes del nuevo modelo son

$$\hat{\gamma}_1 = \frac{\sum z_i y_i - n\bar{z}\bar{y}}{\sum z_i^2 - n\bar{z}^2} = \frac{\sum 10x_i y_i - n10\bar{x}\bar{y}}{\sum 100x_i^2 - n100\bar{x}^2} = \frac{10(\sum x_i y_i - n\bar{x}\bar{y})}{100(\sum x_i^2 - n\bar{x}^2)} = \frac{\hat{\beta}_1}{10}$$

$$\hat{\gamma}_0 = \bar{y} - \hat{\gamma}_1 \bar{z} = \bar{y} - \frac{\hat{\beta}_1}{10} * 10\bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0.$$

Comparando los modelos

$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 z = \hat{\beta}_0 + \frac{\hat{\beta}_1}{10} * 10x = \hat{\beta}_0 + \hat{\beta}_1 x = y$ , es decir, ambos modelos tiene la misma predicción.

1.3) El estadístico del test  $H_0: \beta_1 = 0$  es  $\frac{\hat{\beta}_1}{\hat{\sigma}_1}$  que sigue una t- student a n-2 grados de libertad bajo la hipótesis nula. Rechazar la hipótesis nula indica que hay una cierta influencia de la variable x sobre la variable y. Se podría usar después de estimar el modelo para hacer predicciones.

## **Problema 2**

Si se tiene el modelo  $Y = X\beta + \varepsilon$  donde se cumplen los supuestos de Gauss-Markov:

2.1) Dé la expresión de  $\hat{\beta}_{MCO}$ .

2.2) Encuentre las propiedades estadísticas de MCO.

**Sol.:**

2.1)

$$\hat{\beta}_{MCO} = (X'X)^{-1} X'Y = \beta + (X'X)^{-1} X'\varepsilon$$

2.2)

$$E(\hat{\beta} / X) = \beta \Rightarrow \text{insesgado}$$

$$V(\hat{\beta} / X) = \sigma^2 (X'X)^{-1} \quad \textbf{(Desarrollado en Auxiliar)}$$

### Problema 3

Considere el modelo lineal :  $y = X\beta + \varepsilon$ .

3.1 Dé los supuestos usuales que se hacen para calcular el estimador de máxima verosimilitud del vector de parámetros  $\beta$ .

3.2 Se desea explicar la **esperanza de vida** de mujeres adultas en Chile con un modelo lineal a partir de las variables explicativas **gasto en salud**, **calorías consumidas** y **tasa de alfabetización**. A partir de 20 observaciones, se efectúa la regresión de la esperanza de vida (de media 67.11 años y desviación típica de 5.52) sobre las tres otras variables; obteniéndose los resultados siguientes:

Variable	coeficiente	Des. típica	T-student	P( X >T)
Constante	29.603	5.282	5.600	0.0001
Gasto salud	0.959	0.417	2.302	0.0336
Calorías	0.161	0.054	2.980	0.0085
Alfabetización	0.217	0.064	3.350	0.0041

F-Fisher para las tres variables: 19.714 con  $P(|X|>F)=0.0001$

Interprete los T-Student y la F-Fisher. Dé los grados de libertad de cada distribución.

3.3 Dé el estimador de máxima verosimilitud y un estimador insesgado de  $\sigma^2$ , la varianza de los errores del modelo.

3.4 Construye un intervalo de confianza de nivel 95% para el parámetro  $\beta_{alfabet}$  asociado a la variable **tasa de alfabetización**. Interprete y discuta la significación de la tasa de alfabetización para explicar la esperanza de vida.

**Sol.:**

3.1 Se supone que  $\varepsilon_i \sim N(0, \sigma^2)$  y los  $\varepsilon_i$  son independientes entre si.

3.2 El p-valor de la F de Fisher es casi nulo: el modelo aporta más que el modelo constante (sin las 3 variables explicativas). Los p-valores de los T-student son todos muy pequeños, lo que indica que los parámetros son significativamente no nulos y entonces que las variables explicativas del modelo tienen efectos sobre la esperanza de vida de las mujeres.

Los grados de libertad de la F de Fisher son: 3 y 16 y los de las T-student son 16.

3.3 Si los  $\hat{\varepsilon}_i^2$  son los estimadores de los errores (residuos),  $\hat{\sigma}^2 = \frac{1}{n} \sum_i \hat{\varepsilon}_i^2$  es el estimador

de máxima verosimilitud y  $\tilde{\sigma}^2 = \frac{1}{n-4} \sum_i \hat{\varepsilon}_i^2$  es un estimador insesgado.

$$3.4 [\hat{\beta}_{alfabet} - t_{16}^{\alpha/2} \hat{\sigma}_{alfabet}, \hat{\beta}_{alfabet} + t_{16}^{\alpha/2} \hat{\sigma}_{alfabet}] = [0.217 - 2.12 * 0.064, 0.217 + 2.12 * 0.064]$$

$[\hat{\beta}_{alfabet} - t_{16}^{\alpha/2} \hat{\sigma}_{alfabet}, \hat{\beta}_{alfabet} + t_{16}^{\alpha/2} \hat{\sigma}_{alfabet}] = [0.0813, 0.3527]$ . Se observa que el 0 no pertenece al intervalo de confianza, lo que hace pensar que la variable alfabetización tiene un efecto sobre la esperanza de vida de las mujeres.