Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Primer Semestre 2017

Curso : Probabilidad y Estadística
Sigla : EYP1113
Profesores : Ricardo Aravena C. y Ricardo Olea O.

Pauta Examen

Problema 1

Cuando los países se van desarrollando, el consumo de energía también aumenta y por esta razón es importante poder predecir el consumo de energía futuro para ir planificando con anticipación como generar dicha energía, como es la construcción de hidroeléctricas o parques eólicos por nombrar algunas formas de generación de electricidad. Para poder predecir, se proponen dos modelos de regresión simple a la demanda eléctrica (en MMWh) de la región Metropolitana 2002-01 a 2015-12, el primero utiliza como regresor el Tiempo y el segundo el PIB nacional.

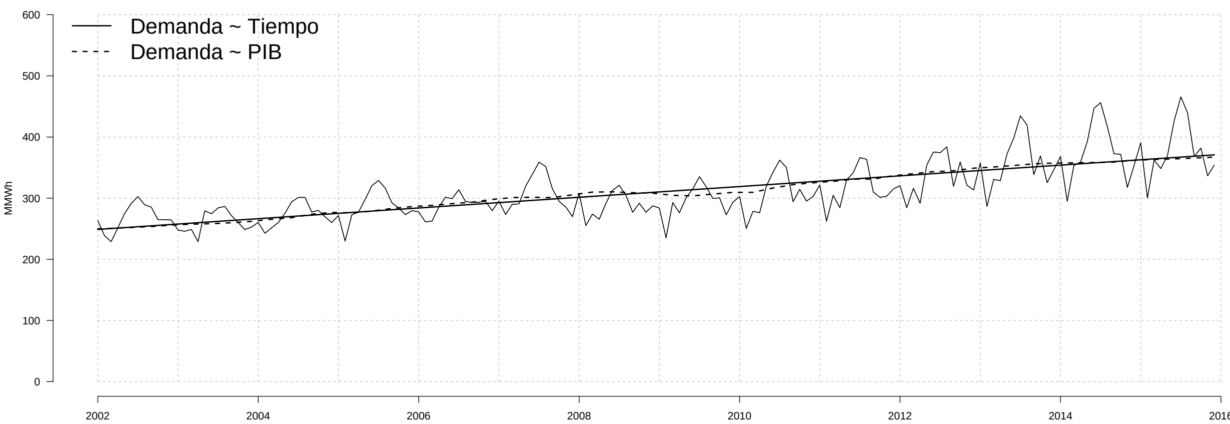

Figura 1: Demanda Electrica Región Metropolitana

```
mean(Demanda) mean(Demanda^2)
     309.8397        98257.34

Demanda ~ Tiempo:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.729e+04  1.222e+03  -14.14   <2e-16 ***
Tiempo       8.759e+00  XXXXXXXXX   XXXXX   XXXXXX

Residual standard error: 31.87 on 166 degrees of freedom
Multiple R-squared:  XXXXXX,Adjusted R-squared:  XXXXXX
F-statistic: 207.3 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
Demanda ~ PIB:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.766e+01  1.467e+01   6.658 3.88e-10 ***
PIB         7.445e-03  5.076e-04  14.668  < 2e-16 ***

Residual standard error: 31.54 on 166 degrees of freedom
Multiple R-squared:  0.5645,Adjusted R-squared:  0.5618
F-statistic: XXXXX on 1 and 166 DF,  p-value: XXXXXXXXX
```

Complete la información faltante e indique cuál de los dos modelos ajusta mejor. Justifique su respuesta.

Respuesta

Tenemos que $n = 168$ y

$$SCT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\,(\overline{Y})^2 = 168 \cdot 98257{,}34 - 168 \cdot (309{,}8397)^2 = 379125{,}7$$

Recordemos que

$$SCT = SCR + SCE \qquad R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \qquad r^2 = 1 - \frac{(n-1)}{(n-2)} \cdot \frac{SCE}{SCT}$$

y que en regresión simple bajo $\mathrm{H}_0 : \beta_1 = 0$ se tiene que $T_{b_1} \sim$ t-Student$(n-2)$ y que el estadístico $F = \frac{SCR/1}{SCE/(n-2)} = T_{b_1}^2 \sim F(1,\, n-2)$.

Además $\sqrt{SCE/(n-2)} =$ Residual standard error.

```
Demanda ~ Tiempo:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.729e+04  1.222e+03  -14.14   <2e-16 ***
Tiempo       8.759e+00  0.6083519   14.40   <2.2e-16 ***

Residual standard error: 31.87 on 166 degrees of freedom
Multiple R-squared:  0.5553,Adjusted R-squared:  0.5526
F-statistic: 207.3 on 1 and 166 DF,  p-value: < 2.2e-16
```

[3.0 Ptos.]

```
Demanda ~ PIB:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.766e+01  1.467e+01   6.658 3.88e-10 ***
PIB         7.445e-03  5.076e-04  14.668  < 2e-16 ***

Residual standard error: 31.54 on 166 degrees of freedom
Multiple R-squared:  0.5645,Adjusted R-squared:  0.5618
F-statistic: 215.2 on 1 and 166 DF,  p-value: < 2e-16
```

[1.2 Ptos.]

El modelo que explica la demanda de energía por PIB presenta un mayor R^2 y r^2, un estadístico T y F más alejado de cero y por ende un menor valor-p.

[1.8 Ptos.]

+ 1 Punto Base

Problema 2

Un especialista modela, a través de una regresión múltiple, la tasa semanal de enfermedades respiratorias (ER) en función de tres contaminantes: PM10 (micro-particulado de 10 μg), O_3 (Ozono) y CO_2 (dióxido de carbono), considerando además la temperatura promedio y la humedad de la semana respectiva. Selecciona al azar una muestra de 34 semanas de los dos últimos años, ajustando diversos modelos, como se muestra a continuación:

Modelo	Variables Incluidad	R2
1	CO2	27.4
2	O3,TEMP	31.5
3	O3, CO2, TEMP	34.4
4	O3, CO2, HUM	41.3
5	PM10, O3, CO2, TEMP	49.2
6	PM10, O3, CO2, TEMP, HUM	54.7

La desviación estándar de ER es igual a 0,45.

(a) Determine si existe un aporte significativo del PM10 para explicar linealmente la tasa semana de ER.

(b) Determine si existe un aporte significativo del CO2 para explicar linealmente la tasa semana de ER.

(c) ¿Qué aporte conjunto es más significativo: {PM10, HUM} o {PM10, TEMP}?

(d) Para una de las dos posibles clases jerárquicas y mediante el método forward proponga un modelo de regresión múltiple para explicar linealmente la tasa semanal de ER.

Respuesta

Tenemos que

$$SCT = 0{,}45^2 \cdot (34 - 1) = 6{,}6825$$

Con este resultado y los R^2 podemos agregar a cada modelo su suma de cuadrado de errores

$$SCE = (1 - R^2) \cdot SCT$$

Modelo	Variables Incluidad	R2	SCE
0	NULO		6.682500
1	CO2	27.4	4.851495
2	O3,TEMP	31.5	4.577513
3	O3, CO2, TEMP	34.4	4.383720
4	O3, CO2, HUM	41.3	3.922627
5	PM10, O3, CO2, TEMP	49.2	3.394710
6	PM10, O3, CO2, TEMP, HUM	54.7.	3.027172

(a) .

SOLO se puede evaluar comparando los Mod5 vs mod3:

H0: M3 vs M1: M5, lo que equivale del mod5 imponer H0: B1=0 vs H1: B1<> 0 **[0,3 ptos]**

Test F = [(4,3837 – 3,3947) / 1] / [3,3947 / (34-5)] = 8,44 **[0,5 ptos]**

Como Fc=8,44 > F(0,95; 1; 29) = 4,18 se rechaza H0, **[0,3 ptos]**

es decir PM10 aporta en presencia O3, CO2, TEMP **[0,4 ptos]**

(b) .

Aporte de CO_2 → comparar Mod1 vs nulo y comparar mod3 vs mod2

Ho: mod nulo vs H1: mod 1 **[0,2 ptos]**

Test F = (6,6825 – 4,8515) / (4,8515/32) = 12,077 **[0,3 ptos]**

Como Fc = 12,077 > F(0,96; 1;32) = 4,15 se rechaza H0. CO2 aporta "solo" **[0,3 ptos]**

Ho: Mod2 vs H1: Mod3 **[0,2 ptos]**

Test F = (4,5775-4,3837) / (4,3837/ 30) = 1,32 **[0,3 ptos]**

Como Fc=1,32 < F(0,95; 1;30)=4,17 NO se rechaza Ho.

CO2 no aporta cuando =3 y TEMP esta en el modelo. **[0,3 ptos]**

(c) .

{PM10,HUM} → Mod6 vs mod3 {PM10,TEMP} → mod6 vs mod4

1er caso: Ho: Mod3 vs H1: Mod6 <-> Ho: B1=B5=0 vs h1: uno distinto de 0 **[0,2 ptos]**

Test F = [(4,3837-3,0272) / 2] / [3,0272/(34-6)] = 6,27 **[0,2 ptos]**

Como Fc=6,27 > F(0,95; 2; 28) = 3,34 se rechaza Ho. **[0,2 ptos]**

2do caso Ho: Mod4 vs H1: Mod6 <-> Ho: B1=B4=0 vs h1: uno distinto de 0 **[0,2 ptos]**

Test F = [(3,9226-3,0272)/2] / [3,0272/(34-6)] = 4,14 **[0,2 ptos]**

Como Fc=4,14 > F(0,95; 2; 28) = 3,34 se rechaza Ho. **[0,2 ptos]**

*** PERO el aporte mas importante es de {PM10, HUM} una mayor SC **[0,3 ptos]**

(d) .

Hay dos clases:

M1={mod0,mod1,mod3,mod5,mod6} o bien M2={Mod0,Mod2,Mod3,Mod5,Mod6}

Elijo Clase M2={Mod0,Mod2,Mod3,Mod5,Mod6} → hacia adelante **[0,5 ptos]**

1er paso H0: Modelo Nulo vs H1: Mod2

Test F = [(6,6825-4,5775)/2] / [4,5775/31] = 7,12 > F(0,95;2;31) = 3,30 se rechaza H0.. se itera

[0,5 ptos]

2d paso H0: mod2 vs H1: mod3 … hecho en b) NO se rechaza H0.. FIN.. **[0,3 ptos]**

el mejor modelo es Mod2 **[0,2 ptos]**

ALTERNATIVAMENTE

ELIJO clase M1={mod0, mod1, mod3, mod5, mod6} **[0,5 ptos]**

1er Ho: Modelo nulo vs H1: mod1 .. hecho en b) 1era parte.. se rechaza Ho.. se itera **[0,2 ptos]**

2do Ho: mod1 vs H1: mod3 **[0,3 ptos]**

Test F = [(4,8515 – 4,3837) / 2] / [4,3837/(34-4)] =1,6 < F(0,95; 2; 30) = 3,32 **[0,4 ptos]**

NO se rechaza Ho→ FIN .. el mejor modelo es Mod1. **[0,1 ptos]**

+ 1 Punto Base

Problema 3

Como han sabido, diversos partidos de futbol se definen en lanzamientos penales. Usted, interesado en determinar si el comportamiento observado en este tipo de campeonato (selecciones) también se da en los partidos a nivel de club, revisa las estadísticas y determina que en un total de 125 partidos a nivel de selección, en 42 se tuvo que llegar a la instancia de los penales. Mientras, que una muestra de 275 definiciones a nivel de equipo muestra que en 72 de ellos se definió por penales.

(a) ¿Los datos anteriores permiten afirmar que es mas probable una definición a penales entre selecciones que entre clubes? Especifique hipótesis, test, valor-p y conclusión para un nivel de significancia del 5 %.

Por otra parte, Bravo se lució atajando penales, Usted constata que en los 42 partidos de mundiales que se han definidos por penales, solo han sido atajados 21 penales (de más de 150 que efectuaron), mientras que en los 72 partidos de otras competencias han sido atajados 58 penales (de más de 250 que se efectuaron).

(b) ¿Para qué valor de significancia existe evidencia que permita afirmar que la tasa de penales atajados difiere según tipo de competencia? Asuma que el número de penales atajados en las definiciones se comporta como una variable aleatorio con distribución Poisson.

Respuesta

(a) .

Hipótesis: Ho: Ps = Pc vs H1: Ps > Pc **[0.5 ptos]**

Selección: 42 de 125 → Ps= 0,336 Clubes: 72 de 275 → Pc=0,262 **[0.5 ptos]**

Pcomun = (42+72) de (125+275) → P = 0,285 **[0.5 ptos]**

Test z = 1,5196 **[0.5 ptos]**

→ valor-p = P(Z > 1,52) = 0,064 **[0.5 ptos]**

Como valor-p no es menor que alfa, no hay evidencia para rechazar H0. Es decir, no se puede afirmar que sea más probable una definición a penales entre selecciones que entre clubes.

[0.5 ptos]

(b) .

DOS PROCESOS POISSON INDEPENDIENTES. – COMPARACION DE TASAS

Ho: L1 = L2 vs H1: L1 <> L2 **[0.5 ptos]**

Estimaciones L1 = 21/42 = 0,5 L2 = 58/72 = 0,806 , **[0.5 ptos]**

Con L-Comun = (21+58)/(42+72) = 0,693 **[0.5 ptos]**

Test z = (0,5 – 0,806) / (raíz(0,693 x (1/42+1/72)) = -1,893 **[0.5 ptos]**

Valor-p = P(Z < -1,893) = 0,029 **[0.5 ptos]**

Por tanto, para todo alfa > 0,029 existe evidencia que permite afirmar que la tasa de penales atajados difiere según tipo de competencia.

[0.5 ptos]

+ 1 Punto Base

Problema 4

El mejor ajuste de regresión lineal simple para la demanda eléctrica en la Región Metropolitana visto en el problema 1, entregó los siguientes residuos

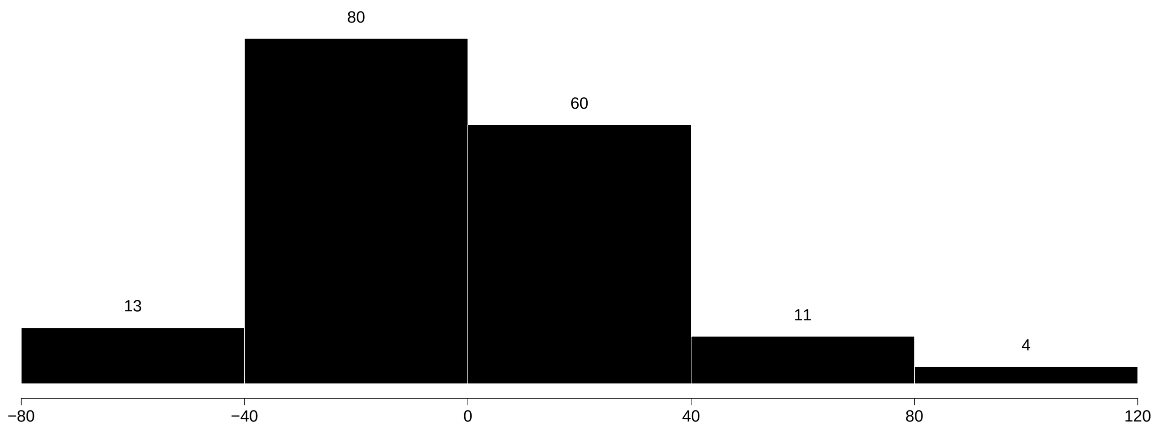

Figura 2: Histograma de frecuencia para residuos mejor ajuste demanda eléctrica región metropolitana

```
   Min. 1st Qu.  Median
-70.700 -21.420  -3.266
```

¿Entre una distribución Normal o Log-Normal, cuál ajusta mejor?

Respuesta

Definamos como Z a los residuos del mejor modelo del problema 1.

$\text{H}_0 : Z \sim \text{Normal}(\mu,\, \sigma) \quad \text{vs} \quad \text{H}_a : Z \not\sim \text{Normal}(\mu,\, \sigma)$

Si z_p es el percentil $p \times 100\,\%$, entonces

$$P(Z \leq z_p) = \Phi\left[\frac{z_p - \mu}{\sigma}\right] = p \to z_p = \mu + \sigma\,\Phi^{-1}(p) \quad \textbf{[0.2 Ptos.]}$$

es decir,

$$\textbf{[0.2 Ptos.]} \quad z_{0{,}25} = \mu - \sigma\, 0{,}6745 \qquad \text{y} \qquad z_{0{,}50} = \mu \quad \textbf{[0.2 Ptos.]}$$

Despejando se tiene que

$$\textbf{[0.2 Ptos.]} \quad \mu = -3{,}266 \qquad \text{y} \qquad \sigma = 26{,}91475 \quad \textbf{[0.2 Ptos.]}$$

Al calcular los valores esperados para cada intervalo, tenemos que el 5to tiene una esperanza menor a 5.

	Observado	Probabilidad	Esperado
[-80, -40)	13	0.086153597	14.4738043
[-40, 0)	80	0.462137947	77.6391751
[0, +40)	60	0.397738276	66.8200304
[+40, +80)	11	0.052981772	8.9009378
[+80,+120]	4	0.000988407	0.1660524
Total	168	1.000000000	168.0000000

[1.0 Ptos.]

Dado lo anterior, se procede a colapsar los últimos dos intervalos y rehacer la tabla

	Observado	Probabilidad	Esperado	(O-E)^2/E
[-80, -40)	13	0.08615360	14.47380	0.15007106
[-40, 0)	80	0.46213795	77.63918	0.07178714
[0, +40)	60	0.39773828	66.82003	0.69609090
[+40,+120]	15	0.05397018	9.06699	3.88228128
Total	168	1.00000000	168.00000	4.80023038

[0.5 Ptos.]

El estadístico de prueba $X^2 = 4{,}8$ y como distribuye $\chi^2(4-1-2)$ se tiene que

$$2{,}5\,\% < \text{valor-p} < 5\,\% \quad \textbf{[0.5 Ptos.]}$$

$\text{H}_0 : Z \sim \text{Log-Normal}(\lambda,\, \zeta)$ trasladad en α vs $\text{H}_a : Z \nsim \text{Log-Normal}(\lambda,\, \zeta)$ trasladad en α

Si z_p es el percentil $p \times 100\,\%$, entonces

$$P(Z \leq z_p) = P(Z - \alpha \leq z_p - \alpha) = \Phi\left[\frac{\ln(z_p - \alpha) - \lambda}{\zeta}\right] = p \to z_p = e^{\lambda + \zeta\,\Phi^{-1}(p)} + \alpha \quad \textbf{[0.2 Ptos.]}$$

es decir,

$$z_{0,00} = \alpha \text{ (mínimo valor posible)} \quad \textbf{[0.2 Ptos.]}$$
$$z_{0,25} = e^{\lambda - \zeta\, 0{,}6745} + \alpha \quad \textbf{[0.2 Ptos.]}$$
$$z_{0,50} = e^{\lambda} + \alpha \quad \textbf{[0.2 Ptos.]}$$

Despejando se tiene que

$$\alpha = -70{,}7; \quad \textbf{[0.2 Ptos.]} \qquad \lambda = 4{,}211149; \quad \textbf{[0.2 Ptos.]} \qquad \zeta = 0{,}464983 \quad \textbf{[0.2 Ptos.]}$$

Al calcular los valores esperados para cada intervalo, tenemos que

	Observado	Probabilidad	Esperado	(O-E)^2/E
[-80, -40)	13	0.04529522	7.609596	3.8183957
[-40, 0)	80	0.49521375	83.195910	0.1227686
[0, +40)	60	0.31628049	53.135123	0.8869188
[+40, +80)	11	0.10134181	17.025424	2.1324424
[+80,+120]	4	0.04186873	7.033946	1.3086296
Total	168	1.00000000	168.000000	8.2691551

[1.0 Ptos.]

El estadístico de prueba $X^2 = 8{,}27$ y como distribuye $\chi^2(5-1-3)$ se tiene que

$$\text{valor-p} < 0{,}5\,\% \quad \textbf{[0.2 Ptos.]}$$

Aunque ambos modelos son rechazados al $5\,\%$ de significancia, un mejor ajusta se logra con el modelo Normal, por tener un mayor valor-p. **[0.4 Ptos.]**

+ 1 Punto Base