Grados en Informática

Métodos Estadísticos para la Computación.

Control 19 Marzo 2013

APELLIDOS, NOMBRE:
DNI: Titulación: Grupo:

---

1. (**MATLAB: Entregar en folio aparte.**) **Indicad solo las órdenes necesarias para resolver el problema.**

   Dada la tabla bidimensional de frecuencias absolutas:

$X\backslash Y$	0	1	2	3	4
$(-\infty,-4]$	0	0	0	25	12
$(-4,0]$	0	0	7	50	5
$(0,4]$	14	12	6	0	0
$(4,8]$	23	7	2	0	0
$(8,\infty)$	50	1	0	0	0

   Hallar:

   (a) La recta de Y/X y la varianza residual del ajuste lineal.
   (b) Ajustar una función de la forma: $\mathbf{y = \frac{10}{a+be^x}}$ y la varianza residual correspondiente.
   (c) Hallar la media y varianza de la variable $Y/(X > 0)$.

   (0.75+1+0.75=2.5 Puntos)

2. Se ha medido el consumo eléctrico (en miles de Kwh) y el costo (en miles de euros) de la climatización del Centro Comercial "Patio Andaluz", resultando los siguientes valores:

	2009		2010		2011		2012	
	I	*II*	*I*	*II*	*I*	*II*	*I*	*II*
Consumo	235	189	255	175	267	199	273	201
Costo	42.6	34.5	46.4	32.0	48.6	36.1	49.6	36.7
IPC	2.234	2.254	2.280	2.307	2.330	2.584	2.794	2.825

   donde se han considerado los periodos semestrales: I={Octubre(año anterior) a Marzo} y II={Abril a Septiembre}.

   Calcular:

   (a) Calcular los índices de variación estacional para el consumo y **desestacionalizar** la serie original. Interpreta los resultados e indica el año y semestre con consumo anormalmente grande si lo hubiese.
   (b) Agrupa los valores del costo por año y calcula los índices simples (base 2009) y de cadena.
   (c) Si consideramos el IPC dado en la tabla. Deflaciona la serie temporal costo.
   NOTA: Téngase en cuenta que la base del IPC es anterior al año 2009.

   (1.5+0.75+0.75=3 Puntos)

3. La tabla bidimensional siguiente indica los puntos obtenidos por el equipo "X" al enfrentarse con los otros 19 de una liga.

$D\backslash F$	0	1	3
0	3	1	0
1	2	4	1
3	5	3	0

   (a) Calcular las rectas de regresión F/D y D/F. El coeficiente de correlación lineal.
   (b) Estimar, mediante el ajuste lineal, el resultado dentro (D) conocido que fuera empató (F=1).
   (c) Ajustar una función de la forma $F = \frac{K}{D+1}$
   (d) ¿Qué ajuste resulta mejor? Este último o el lineal. Justificar la respuesta.
   (e) ¿Que variable tiene mayor dispersión relativa? La F o la D.

   (1.25+0.5+1+1+0.75=4.5 Puntos)

SOLUCIONES:

Problema 1:

```
x=[-6 -6 -2 -2 -2  2  2 2  6 6 6 10 10]
y=[ 3  4  2  3  4  0  1 2  0 1 2  0  1]
n=[25 12  7 50  5 14 12 6 23 7 2 50  1]
N=sum(n)
M=[N sum(n.*x);sum(n.*x) sum(n.*x.^2)]
B=[sum(n.*y);sum(n.*x.*y)]
sol=M\B
a=sol(1), b=sol(2)
medY=sum(n.*y)/N
Vy=sum(n.*y.^2)/N-medY^2
sy=sqrt(Vy)
medX=sum(n.*x)/N
Vx=sum(n.*x.^2)/N-medX^2
sx=sqrt(Vx)
cov=sum(n.*x.*y)/N-medX*medY
r=cov/(sx*sy)
Vr_recta=(1-r^2)*Vy
disp('Apartado b')
% Hago el cambio Y=10/y, X=exp(x) y queda Y=a+bX
Y=10./y,X=exp(x)
M2=[N sum(n.*X);sum(n.*X) sum(n.*X.^2)]
B2=[sum(n.*Y);sum(n.*X.*Y)]
sol2=M2\B2
a2=sol2(1),b2=sol2(2)
yest=10./(a2+b2.*exp(x))
res=y-yest
Vr2=sum(n.*res.^2)/N-(sum(n.*res)/N)^2
disp('Apartado 1-c')
% Construyo la variable agrupando las 3 últimas filas
ycond=[0 1 2 3 4]
nn=[14+23+50 12+7+1 6+2 0 0]
NN=sum(nn)
media=sum(nn.*ycond)/NN
varianza=sum(nn.*ycond.^2)/NN-media^2
```

NOTA: Para los datos de la tabla, la inversión $Y = \frac{1}{y}$ produce un valor $1/0$ y el ajuste no podría realizarse. Eso se observa al ejecutar, cosa que no hay que realizar.

Problema 2-a:

Año	*Per.*	X	$\bar{X}_2$	$\bar{\bar{X}}_2 = T$	X/TC	Desest.
	I	235		–	–	202.2366
2009			212			
	II	189		217.0	0.8710	225.5384
			222			
	I	255		218.5	1.1670	219.4482
2010			215			
	II	175		218.0	0.8028	208.8318
			221			
	I	267		227.0	1.1762	229.7752
2011			233			
	II	199		234.5	0.8486	237.4716
			236			
	I	273		236.5	1.1543	234.9387
2012			237			
	II	201		–	–	239.8583

$VE(I) = \frac{1.167+1.1762+1.1543}{3} \approx 1.1659$

$VE(II) = \frac{0.871+0.8028+0.8486}{3} \approx 0.8408$

Como la suma $VE(I) + VE(II) = 2.0067 \neq 2$ necesitamos ajustar los índices: $I_{VE}(I) = \frac{1.1659*2}{2.0067} \approx 1.162$ $I_{VE}(II) = \frac{0.8380*2}{2.0067} \approx 0.838$

Se interpreta como que en el primer semestre se consume un 16.2% más que la media anual y en el segundo un 16.2% menos.

No existe semestre con consumo anormalmente grande. El lugar para verlo más correctamente sería la serie destendenciada y desestacionalizada (no calculada).

2-b: Agrupamos los datos:

$Año$	X	I_S	I_{Cad}
2009	77.14	1.0000	–
2010	78.40	1.0163	1.0163
2011	84.70	1.0980	1.0804
2312	86.32	1.1190	1.0191

2-c:

	I	II	I	II	I	II	I	II
Costo	42.6	34.5	46.4	32.0	48.6	36.1	49.6	36.7
IPC	2.234	2.254	2.280	2.307	2.330	2.584	2.794	2.825
$IPC_{BaseI-2009}$	1.0000	1.0090	1.0206	1.0327	1.0430	1.1567	1.2507	1.2645
Costo Deflacionado	42.6000	34.1939	45.4639	30.9874	46.5976	31.2103	39.6587	29.0222

Donde $IPC_{BaseI-2009}$ se ha calculado dividiendo IPC por 2.234 (valor del IPC en el período I de 2009). Por último el "Costo deflacionado" es la fila "Costo" dividida término a término por la $IPC_{BaseI-2009}$.

Problema 3-a:

Calculamos los parámetros necesarios para las ecuaciones normales:

$D\backslash F$	0	1	3	$n_{i.}$	$D_i n_{i.}$	$D_i^2 n_{i.}$
0	3	1	0	4	0	0
1	2	4	1	7	7	7
3	5	3	0	8	24	72
$n_{.j}$	10	8	1	$N=19$	$A=31$	$C=79$
$F_j n_{.j}$	0	8	3	$B=11$		
$F_j^2 n_{.j}$	0	8	9	$D=17$		

$$\begin{cases} N=\sum_i\sum_j n_{ij}=19 \\ A=\sum_i D_i n_{i.}=31 \\ B=\sum_j F_j n_{.j}=11 \\ C=\sum_i D_i^2 n_{i.}=79 \\ D=\sum_j F_j^2 n_{.j}=17 \\ E=\sum_i\sum_j n_{i,j} D_i F_j = \\ \quad =4(1)(1)+(3+1)(1)(3)=16 \end{cases}$$

Para la recta F/D:

$$\left.\begin{array}{rll} \sum_j F_j n_{.j} & = aN & +b\sum_i D_i n_{i.} \\ \sum_i\sum_j n_{ij} D_i F_j & = a\sum_i D_i n_{i.} & +b\sum_i D_i^2 n_{i.} \end{array}\right\} \Rightarrow \left.\begin{array}{l} 11=19a+31b \\ 16=31a+79b \end{array}\right\} \Rightarrow \left\{\begin{array}{c} a=0.6907 \\ b=-0.0685 \end{array}\right\} \Rightarrow$$

Recta de F/D: $\mathbf{F = 0.6907 - 0.0685D}$

Para la recta D/F:

$$\left.\begin{array}{rll} \sum_i D_i n_{i.} & = a'N & +b'\sum_j F_j n_{.j} \\ \sum_i\sum_j n_{ij} D_i F_j & = a'\sum_i F_j n_{.j} & +b'\sum_j F_j^2 n_{.j} \end{array}\right\} \Rightarrow \left.\begin{array}{l} 31=19a'+11b' \\ 16=11a'+17b' \end{array}\right\} \Rightarrow \left\{\begin{array}{c} a'=1.7376 \\ b'=-0.1832 \end{array}\right\} \Rightarrow$$

Recta de D/F: $\mathbf{D = 1.7376 - 0.1832F}$

$$r=\sqrt{bb'}=-\sqrt{(0.0685)(0.1832)}\approx -0.1120$$

3-b: Debemos usar la recta de D/F (no la de F/D): D=1.7376-0.1832(1)=1.5545

3-c: De la expresión: $F=\frac{K}{D+1}$ podemos deducir que se trata de expresar la función en una base con un único elemento $B=\{\frac{1}{D+1}\}$.

De la condición de óptimo:

$$\left\langle e, \frac{1}{D+1}\right\rangle = 0 \Rightarrow \left\langle F-\frac{K}{D+1}, \frac{1}{D+1}\right\rangle = 0 \Rightarrow \sum_i \frac{F_i}{D_i+1} = K\sum_i \frac{1}{(D_i+1)^2}$$

D_i	F_i	n_i	$n_i F_i$	$X_i=\frac{1}{D_i+1}$	$n_i F_i X_i$	$n_i X_i$	$n_i X_i^2$	$F_i^{est}=\frac{0.84}{X_i}$	e_i	$n_i e_i$	$n_i e_i^2$	$n_i F_i^2$
0	0	3	0	1.00	0.00	0	3	0.84	−0.84	−2.52	2.1168	0
0	1	1	1	1.00	1.00	0	1	0.84	0.16	0.16	0.0256	1
1	0	2	0	0.50	0.00	2	0.5	0.42	−0.42	−0.84	0.3528	0
1	1	4	4	0.50	2.00	4	1	0.42	0.58	2.32	1.3456	4
1	3	1	3	0.50	1.50	1	0.25	0.42	2.58	2.58	6.6564	9
3	0	5	0	0.25	0.00	15	0.3125	0.21	−0.21	−1.05	0.2205	0
3	1	3	3	0.25	0.75	9	0.1875	0.21	0.79	2.37	1.8723	3
		19	11		5.25		6.25			3.02	12.59	17

Por lo que: $\mathbf{K} = \frac{\mathbf{5.25}}{\mathbf{6.25}} = \mathbf{0.84}$ y la función ajustada es: $\mathbf{F} = \frac{\mathbf{0.84}}{\mathbf{D+1}}$

3-d: Para compararlos calculamos la varianza residual de ambos.

$$V_F = \frac{\sum_i n_i F_i^2}{N} = \frac{17}{19} - \left(\frac{11}{19}\right)^2 \approx 0.5596$$

$$V_r(lineal) = (1 - r^2)V_F = (1 - 0.112^2)0.5596 = 0.5525$$

$$V_r(hiperbolico) = \frac{12.59}{19} - \left(\frac{3.02}{19}\right)^2 = 0.6374$$

En resumen: Ambos ajustes son malos resultando algo mejor (menos malo) el lineal.

Hemos calculado las columnas necesarias de la tabla, entre ellas: $F_i^{est} = \frac{0.84}{D_i+1}$, $e_i = F_i - F_i^{est}$.

3-e: Para poder comparar dispersiones entre poblaciones diferentes, se emplea el coeficiente de variación.

$\bar{F} = \frac{\sum_i n_i F_i}{19} = \frac{11}{19} = 0.5789$, $\bar{D} = \frac{\sum_i n_i D_i}{19} = \frac{31}{19} = 1.6316$, $var(F) = 0.5596 \Rightarrow \sigma_F = \sqrt{0.5596} = 0.7480$, $var(D) = \frac{79}{19} - (1.6316)^2 = 1.4958 \Rightarrow \sigma_D = 1.2230$

$$\mathbf{CV_D} = \frac{\mathbf{1.2230}}{\mathbf{1.6316}} = \mathbf{0.7496}, \qquad \mathbf{CV_F} = \frac{\mathbf{0.7480}}{\mathbf{0.5789}} = \mathbf{1.2921}$$

Luego tiene mayor dispersión relativa la variable F, a pesar de tener menor varianza (desviación típica).