Pràctiques Integrades
1er de Matemàtiques

Pràctica 23

curs 2002–03

23 Ajust a les dades. Estadística descriptiva

23.1 Ajust a les dades

Una qüesió que es planteja sovint amb dades experimentals és la següent: es realitza un experiment i es voldria saber si hi ha una relació entre dues de les seves propietats (una relació de proporcionalitat per exemple). És a dir, pot tenir interés descriure a través d'una funció la relació entre dues variables quantitatives X i Y. Posem $Y = f(X)$ on f és una funció que depèn d'uns paràmetres $a_1, ..., a_d$ que haurem de determinar per garantir un *bon* ajust. Naturalment, a més de disposar d'observacions, $(x_1, y_1), ..., (x_n, y_n)$ on $x_i, y_i \in \mathbb{R}$, hem de recorre a un criteri que valori la bondat de l'ajust. El criteri que estudiarem és l'anomenat de *mínims quadrats*.

23.1.1 Llei de Hooke

Començarem plantejant un experiment concret. La llei de Hooke a la física ens diu que hi ha una relació de proporcionalitat entre el pes que es penja d'una molla i la distància que aquesta s'estira. És a dir, si x és un pes i y representa la distància que una molla s'estira en penjar-li aquell pes, existeix una constant k tal que $y = kx$. k s'anomena la constant d'elasticitat de la molla.

L'experiment que simularem consisteix en determinar aquesta constant k per una molla concreta. Per això, s'hauria de mesurar les distàncies que la molla s'estira per a diferents pesos i s'obtindria una taula com la següent

Pes	cm
2	3.2
2.5	4
3	5
3.5	5.07
4	6.5
5	8

Si l'experiment es realitzés en condicions perfectes (mesures de precisió perfecte, sense fregament amb l'aire,...) les dades obtingudes serien de la forma $y_i = kx_i$ i per tant, seria fàcil determinar la constant k. Però com que no podem suposar condicions perfectes, obtenim unes dades que es troben lleugerament distorsionades com es pot observar si es fa l'exercici següent

Exercici 23.1

En uns eixos de coordenades representeu la taula de punts `[Pes,cm]` de l'experiment.

Per tant l'objectiu és trobar la recta de la forma $y = kx$ que millor s'ajusta a les dades obtingudes experimentalment. Per això hem de fixar el criteri: és a dir, què vol dir que una recta s'ajusta millor que una altra al conjunt de dades?

El criteri a fer servir és el següent: buscarem k tal que la suma de les distàncies dels punts obtinguts experimentalment a la recta sigui mínima. És a dir, el valor de k tal que

$$\sum_{i=1}^{6}(y_i - kx_i)^2$$

pren el valor mínim. Observeu que aquesta funció mesura l'error comés en l'experiment respecte les condicions perfectes.

Exercici 23.2

Definiu una funció `d` tal que `d(k)=`$\sum_{\texttt{i=1}}^{\texttt{6}}$`(y_i - kx_i)^2`, busqueu el valor de `k` pel qual `d` pren el valor mínim. En una mateixa gràfica representeu els punts obtinguts a l'experiment i la gràfica de la recta `y=kx` pel valor de `k` obtingut.

Aquest mètode que acabem de descriure per aproximar dades obtingudes experimentalment per funcions és el que es coneix com el mètode dels mínims quadrats.

23.1.2 El mètode dels mínims quadrats

En general, aquest mètode s'utilitza quan es vol analitzar la relació funcional entre un conjunt de dades (y_i, x_i) obtingudes experimentalment. Es suposa que hi ha un conjunt de paràmetres $a_1, \ldots, a_n$ que determina el tipus de funció f que les relaciona però no els paràmetres $a_1, \ldots, a_d$ de la funció.

El criteri de mínims quadrats proposar calcular els paràmetres $a_1, \ldots, a_d$ tals que l'expressió dels errors

$$\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2,$$

és mínima, és a dir, la suma dels desajustos quadràtics entre els valors y_i observats i els valors $f(x_i)$ (que depenen de l'elecció dels a_i) proposats per f.

El primer pas sempre consisteix en representar gràficament en uns eixos de coordenades els punts obtinguts experimentalment com ja heu fet en la secció anterior. D'aquest tipus de diagrama se'n diu un diagrama de dispersió. Aquest tipus de gràfics es pot fer amb una comanda `plot` amb opció `style=point` però al paquet **stats** hi ha una funció dedicada específicament a aquest tipus de gràfics. Observeu el seu funcionament en el següent exemple,

Exemple 23.1

Donades les observacions següents:

$$(0.70, 0.035), (0.76, 0.025), (0.37, -0.18), (0.82, 0.045)$$
$$(0.29, -0.16), (0.56, -0.058), (0.42, -0.11), (0.47, -0.085),$$

el següent grup de comandes realitza la representació gràfica de les dades anteriors en un gràfic de dispersió:

```
> with(stats[statplots]):
> X:=[0.70,0.76,0.37,0.82,0.29,0.56,0.42,0.47];
> Y:=[0.035,0.025,-0.18,0.045,-0.16,-0.058,-0.11,-0.085];
> scatterplot(X,Y, symbol=circle);
```

Noteu que amb `scatterplot` podem guardar en dues llistes independents els resultats de les mesures `X` i `Y` per a fer el diagrama de dispersió mentre que amb un `plot` cal fer una llista amb el valors de la forma `[[x1,y1],...]`.

Un cop fet el diagrama, el següent pas consisteix en analitzar la forma del núvol de punts obtinguts per determinar quin tipus de funció s'hi ajustaria millor.

Exercici 23.3

Quin tipus de funció creus que podríem triar per aconseguir un bon ajust a les dades de l'exemple anterior?

Exercici 23.4

Donades les observacions:

$$(-1,5),(-0.4,2.5),(0.1,2.1),(0.8,3.98)$$
$$(1.2,6.5),(1.5,8.8),(2.1,15.2)$$

Representeu les dades en un gràfic de dispersió. En aquest cas, quina mena de funció creus que podríem triar per ajustar a les dades?

En l'exemple 23.1 semblava adequat proposar una funció lineal $y = f(x) = a + bx$ (una recta). Ara bé, en l'exercici 23.4 sembla més adequat triar $y = f(x) = a + bx + cx^2$ (polinomi de grau 2). En ambdós casos es tracta de funcions que son lineals en els paràmetres.

Un cop determinat el tipus de funció, podem calcular els paràmetres pels quals la suma dels quadrats dels errors de mesura és mínim. En el paquet **stats**[1] hi ha la comanda `leastsquare` que calcula els paràmetres que minimitzen aquests errors. Com a paràmetres necessita el nom de les variables, el tipus de funció i les llistes de dades experimentals.

En el exemple següent es veu com utilitzar aquesta comanda pel cas d'una recta. L'aplicarem a les dades obtingudes per l'experiment de la molla. Anomeneu `Pes` i `Cm` les dades obtingudes en l'experiment de la molla.

[1] També hi ha una funció semblant en el paquet **CurveFitting**

Exemple 23.2

```
> with(stats):
> fit[leastsquare[[x,y]]]([Pes,Cm]);
```

Exercici 23.5

Compareu aquest resultat amb l'obtingut a la secció 23.1.1.

Exercici 23.6

Calculeu els paràmetres de la recta que millor s'ajusta a les dades de l'exemple 23.1. En una mateixa gràfica representeu el diagrama de dispersió i la recta obtinguda.

En el cas d'ajustar altra tipus de corbes, com és el cas de l'exercici anterior on un polinomi de grau 2 semblava més adequat, la sintaxi és

```
> with(stats):
> fit[leastsquare[[x,y],y=a*x^2+b*x+c,{a,b,c}]]([X,Y]);
```

Exercici 23.7

Representeu en una mateixa gràfica el diagrama de dispersió i la funció quadràtica obtinguda.

Exercici 23.8

El punt d'ebullició d'una mescla d'ethanol i aigua depèn de la proporció d'ethanol a la mescla. S'ha realitzat un experiment on hem obtingut les dades següents:

```
Temp=[100, 95.5, 89.0, 86.7, 85.3, 84.1, 82.7, 82.3, 81.5, 80.7, 79.8, 79.7, 79.3,
                                                              78.74, 78.41, 78.15]
Prop=[0.0, 0.0190, 0.0721, 0.0966, 0.1238, 0.1661, 0.2337, 0.2608, 0.3273, 0.3965,
                                    0.5079, 0.5198, 0.5732, 0.6763,0.7472, 0.8943]
```

Si sabem que la funció que relaciona el punt d'ebullició `Temp` de la mescla i la proporció `Prop` d'ethanol ha de ser de la forma $y = ae^{-14x} + bx + c$ on y és la temperatura d'ebullició i x la proporció d'ethanol a la mescla. Calculeu les constants a, b, c per l'ethanol.

23.2 Estadística descriptiva

El paquet **stats** no conté únicament les funcions per a determinar ajusts per mínims quadrats o per a fer gràfics de conjunts de dades, la part bàsica d'aquest paquet és un recull de subpaquets cadascun adequat per a diferents tipus d'anàlisis estadístics de dades. En aquesta pràctica farem un cop d'ull als subpaquets **describe**, **statsplots** i **transform**.

Exemple 23.3

Les dades que es vulguin analitzar estaran guardades en una llista sobre la que aplicarem les diferents funcions del paquet. Si volem estudiar estadísticament els resultats del llançament d'un dau de sis cares cinc cops i els resultats obtinguts són $1, 3, 4, 3, 2$ posarem

```
> with(stats):
> data:=[1,3,4,3, 2];
```

i podrem aplicar a la llista `data` les diferents funcions.

En cas que l'ordre de les dades sigui irrellevant podem incorporar a la llista la funció `Weight(valor,pes)` on el paràmetre `valor` indica un dels valors de la llista i `pes` és el nombre de vegades que el valor apareix a la llista (com si poséssim `valor$pes`).

```
> [1,4,6, Weight(3,2), 5$3];
```

Maple ens permet importar un fitxer de dades ja existent utilitzant `importdata("fitxer",n)` on `fitxer` és l'adreça del fitxer de dades (que serà un fitxer text on les dades estan organitzades per columnes) i `n` és el número de columnes del fitxer. Maple importa el fitxer com una seqüència de n llistes de dades, una per cada columna. En l'exemple següent importem un fitxer que es diu `"tempsvida.dat"`, i està en el directori `"c:/Mis documentos"` del nostre ordinador, on hi ha guardat el temps de vida (en hores) d'una mostra de llums fluorescents[2].

```
> llums:=importdata("c:/Mis documentos/tempsvida.dat",1);
```

ara tenim en `llums` les dades recollides del que han durat aquests llums.

23.2.1 Les comandes de `describe`

Habitualment l'anàlisi de les dades $x_1, \ldots, x_n$ obtingudes en un experiment s'inicia amb un estudi descriptiu. En aquest punt és el subpaquet **describe**, el que ens proporciona les funcions necessàries com són, per exemple, `count` (que compta el nombre de valors n que apareixen en la llista de dades), `mean` (que calcula la mitjana $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ aritmètica de les dades introduïdes), `mode` (que dóna com a resultat el valor que apareix més cops dins del conjunt de dades), `variance` (que dóna una mesura de la dispersió dels valors de les dades calculant la mitjana aritmètica dels quadrats de les diferències de cada valor amb la mitjana de les dades, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$) i moltes altres que podreu consultar a l'ajuda de Maple a partir de `help(describe);`

[2] El podeu baixar des de `http://mat.uab.es/gguasp/Lamp/tempsvida.dat`

Exemple 23.4

```
> with(stats):
> data:=[1,1,3,Weight(2,4)];
> describe[count](data);
> describe[mean](data);
> describe[mode](data);
> describe[variance](data);
```

Noteu que, de fet, `describe` és una `table` de Maple en la que cada un dels seus elements és una funció diferent que es pot aplicar al conjunt de dades `data`.

Podem agrupar un seguit d'aquestes comandes per formar una llista d'operacions que s'apliquen conjuntament sobre la mateixa llista de dades i així obtenir el conjunt dels valors significatius en un sol resultat.

Exemple 23.5

```
> descriptiva:=[describe[mean],describe[median],describe[variance]];
> descriptiva(data);
```

Exercici 23.9

Apliqueu la comanda `descriptiva` a les dades del temps de vida dels llums fluorescents guardats en `llums`.

23.2.2 Representació gràfica de les dades

Dins el paquet **stats** l'apartat que conté les funcions per a fer representacions gràfiques és `statplots`

23.2.3 Gràfics `boxplot`

Les representacions gràfiques resumeixen alguns aspectes de les dades. Un tipus de gràfic adequat per la representació de dades quantitatives és l'anomenat gràfic BoxPlot (diagrama de caixes) del que en teniu un exemple a continuació.

Exemple 23.6

```
> Xdata := [4.535, 4.029, 5.407, 1.605, 5.757, 3.527,
> 7.890,8.159, 6.092, 13.442, 2.845, 5.172, 3.277, 8.810, 3.657, 7.226,3.851,
> 2.162, 2.668, 4.692];
> statplots[boxplot](Xdata);
```

la línia central de la caixa representa la mediana (el valor central de les dades), les línies superior i inferior representen el primer i el tercer quartil i les línies que surten cap dalt i cap baix s'estenen fins a 3/2 del rang de valors entre quartils (els quartils d'una llista de dades són els tres valors que divideixen el conjunt total de dades, un cop s'han ordenat de menor a major, en quatre parts amb el mateix nombre de dades en cada un d'ells).

Noteu que Maple situa el diagrama de caixa sobre la posició zero de l'eix horitzontal. Si volem que el desplaci podem utilitzar l'opció `shift=n`, on `n` és el número d'unitats que volem que es desplaci. A més podem indicar l'amplada de la caixa mitjançant l'opció `width=n`. Si a la comanda `boxplot` introduïu més d'un fitxer de dades

```
> statplots[boxplot](Xdata1,Xdata2);
```

obtindreu els diagrames de caixa corresponents en un sol gràfic. Aquesta opció és útil quan es vol comparar el comportament de més d'un conjunt de dades.

23.2.4 Histogrames

La distribució de les freqüències dels valors observats podem representar-la amb un histograma de freqüències com en l'exemple següent:

Exemple 23.7

```
> statplots[histogram](Xdata, numbars=5, area= 100);
> statplots[histogram](Xdata, numbars=5, area= count, color=cyan);
```

Observeu que l'opció `numbars=5` fa que es distribueixen les dades en cinc grups, que cada una de les barres té la mateixa amplada i una altura proporcional al nombre de dades que cauen dins el grup corresponent i que el valor numèric d'aquesta altura ens dóna els percentatge de dades en cada grup quan posem l'opció `area=100` i el nombre total de dades quan l'opció és `area=count`. Noteu també que s'hi poden afegir opcions gràfiques com per exemple `color=cyan` que farà que el color del dibuix sigui `cyan`.

Exercici 23.10

Importeu el fitxer `data3.dat` que podreu baixar de `http://mat.uab.es/gguasp/Lamp/data3.dat`. Aquest fitxer conté tres columnes que corresponen al temps de vida en hores de tres tipus de fluorescents segons el seu cebador: ràpid, de pre-escalfament i instantani. Anomeneu els tres tipus de dades com `instantani`, `escalfament` i `rapid` (recordeu com s'utilitza `importdata`).

- Calculeu el número de fluorescents de cada tipus estudiats.
- Per cada classe de fluorescent calculeu la mitjana, moda i variància del seu temps de vida.
- Feu un histograma de 10 barres per cada tipus de fluorescent en que les columnes representin la freqüència de casos.
- En un mateix gràfic dibuixeu els corresponents diagrama de caixa i analitzeu el resultat.

23.2.5 Les comandes de `transform`

Finalment analitzarem algunes de les comandes del subpaquet **transform**. En particular veurem com ordenar un conjunt de dades i com comptar el número de vegades que apareix un mateix valor.

Per als exemples següents utilitzarem la llista `sim` que simula el llançament d'una moneda 20 cops obtinguda fent que Maple generi 20 números aleatoris enters entre 0(=cara) i 1(=creu) i guardant els resultats en `sim`.

```
> moneda:=rand(2);
> sim:=[seq(moneda(),i=1..20)];
```

La comanda `transform[statsort](data)` serveix per ordenar un conjunt de dades (`data`) en ordre creixent.

Exemple 23.8

```
> ord:=transform[statsort](sim);
```

La comanda `transform[tally](data)` agrupa totes les dades que prenen el mateix valor mitjançant l'expressió `Weight`. D'aquesta manera podem observar el número de vegades que apareix cada valor.

Exemple 23.9

```
> agrup:=transform[tally](sim);
```

Finalment, i un cop hem agrupat les dades, la comanda `transform[frequency](data)` retorna una llista amb la freqüència de cada valor de la llista (número de vegades que es repeteix cada valor).

Exemple 23.10

```
> transform[frequency](sim);
```

Exercici 23.11

Simuleu l'experiment que consisteix en llançar un dau 100 vegades (recordeu la comanda `rand`). Calculeu el número de vegades que ha sortit cada cara del dau i feu un gràfic on quedin reflexats aquestes dades.