MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Problem Sheet 5

Remember: when online, you can access the Statistics 1 data sets from an **R** console by typing

```
load(url("http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/stats1.RData"))
```

*1. As part of a study of the relationship between ‘stress’ and ‘skill’, the stress levels for eight second year student volunteers were assessed and compared with their subject skill levels, as measured by each student’s average mark at the end of their first year.

Summary statistics for the data set are:
$n = 8$ $\sum x_i = 492$ $\sum y_i = 379$ $\sum x_i^2 = 32,894$ $\sum y_i^2 = 20,115$ $\sum y_i x_i = 21,087$
where x_i is the subject skill level for the ith student and y_i is their assessed stress level.

Calculate by hand the least squares estimates of α and β and the equation for the fitted regression line, under the simple linear regression model $\mathrm{E}(Y|x) = \alpha + \beta x$. What broad conclusion can you draw immediately from the fitted model? What stress level would you predict for a student with skill level $x = 60$?

2. The table below shows a data set with five pairs of values $(x_i, y_i), i = 1, \ldots, 5$. It is thought that the data satisfy the simple linear regression model $\mathrm{E}(Y|x) = \alpha + \beta x$, $\mathrm{Var}(Y|x) = \sigma^2$.

x	1	3	4	6	7
y	0	1	2	5	4

(a) Calculate by hand the least squares estimates of α and β.

(b) For $i = 1, \ldots, 5$, calculate by hand the fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ and the residual values $\hat{e}_i = y_i - \hat{y}_i$. Hence calculate by hand an estimate of the common variance σ^2. Also, find the sum of the residuals.

3. The table below shows the average weight (in kg) of piglets in a litter, for seven litters of varying size. The data are contained in the Statistics 1 data frame `pig`, variables `littersize` and `wt` respectively.

Litter size (x)	1	3	5	8	8	9	10
Average weight (y)	1.6	1.5	1.5	1.3	1.4	1.2	1.1

Following the method of §4.7 in the handout, but with the obvious changes to the names used, perform a simple linear regression of average weight on litter size and output the results to the **R** object `piglets`, with the commands:

```
> attach(pig); piglets <- lm(wt ~ littersize)
```

Calculate *by hand* the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ for the simple linear regression model $\mathrm{E}(Y|x) = \alpha + \beta x$ and compare your answers with **R** with the command: `> coef(piglets)`

Draw a scatter plot of the data and add in the fitted regression line, using the commands:

```
> plot(littersize,wt); abline(coef(piglets))
```

Use your fitted regression line to predict the average weight of a piglet in a litter of size 6.

Let x_i denote the litter size for the ith litter and let y_i denote the corresponding average weight for the piglets in that litter. Inspect the fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ and the residual values $\hat{e}_i = y_i - \hat{y}_i$ with the commands:

```
> fitted(piglets); residuals(piglets)
```

Plot the residuals against the litter sizes using the command:

```
> plot(littersize, residuals(piglets))
```

and comment on the fit of the model.

*4. As part of a study to develop an ecologically sustainable policy for regulating the fishing of Dungeness crabs, marine scientists in California needed to be able to 'predict' the size of adult female crabs *before* they moulted (shed their renewable shells) from their size *after* they moulted. (Size here means shell width in mm.; the apparently 'backwards' prediction is relevant because it is the postmoult sizes that are routinely measured in practice, not the premoult sizes). Data on 342 such crabs can be downloaded as the file `crabs.R` in the `downloads` directory on the website, or loaded into **R** if you are online by typing

```
source("http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/crabs.R")
```

The variables of interest in this data frame are `postmoult` and `premoult`.

(a) Let x_i denote the postmoult size of the ith crab and let y_i denote the premoult size. Use **R** to calculate the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ for the simple linear regression model $\mathrm{E}(Y|x) = \alpha + \beta x$. Produce a scatter plot of the data and add in the fitted regression line.

(b) Calculate the fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ and the residual values $\hat{e}_i = y_i - \hat{y}_i$. Plot the residuals against the postmoult sizes, and comment on the fit of the model.

(c) Construct a histogram of the residuals, and comment on its shape.

(d) What would you predict the premoult size to be for a new crab whose postmoult size is 130? By typing `attach(crabs); hist(premoult[postmoult>127&postmoult<133])` you can display the premoult sizes for those crabs in the data set whose postmoult sizes were in the range $(127, 133)$. Compare.

(e) (Of theoretical interest) check that (i) the residuals and fitted values calculated by **R** add up to the original responses $\{y_i\}$, (ii) the sum of the residuals is 0, (iii) if you carry out a regression of the fitted values on the $\{x_i\}$, then the residuals from *this* fit are all 0.

*5. Consider a regression problem where the data values $y_1, \ldots, y_n$ are observed values of response variables $Y_1, \ldots, Y_n$.

In the notes we assume that, for given values $x_1, \ldots, x_n$ of the predictor variable, the Y_i satisfy the simple linear regression model $Y_i = \alpha + \beta x_i + e_i$, where the e_i are i.i.d. $\sim N(0, \sigma^2)$. The least squares estimates of the regression parameter(s) are defined to be the values which minimise the sum of squares of the differences between the observed y_i and the fitted values. For this model, $\mathrm{E}(Y_i \,|\, x_i) = \alpha + \beta x_i$, so the least squares estimates are the values minimising $\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$.

Now consider an alternative model which takes the form $Y_i = \gamma x_i + e_i, \quad i = 1, \ldots, n$, where the e_i satisfy the same assumptions as before but where there is now a single unknown regression parameter γ. This model is sometimes used when it is clear from the problem description that the value of $E(Y)$ must be zero if the corresponding x value is zero.

Derive an expression, in terms of the x_i and y_i values, for the least squares estimate for γ for this new model and suggest, with reasons, an appropriate estimate for σ^2.