MATH11400 Statistics 1 2010–11

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

Solution Sheet 5

1. Summary statistics for the data set are:
$n = 8 \quad \sum x_i = 492 \quad \sum y_i = 379 \quad \sum x_i^2 = 32,894 \quad \sum y_i^2 = 20,115 \quad \sum y_i x_i = 21,087$
giving $\bar{x} = 61.5,\ \bar{y} = 47.375,\ ss_{xx} = 2636,\ ss_{xy} = -2221.5$ and $ss_{yy} = 2159.875$.
Thus the least squares estimates are

$$\hat{\beta} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{ss_{xy}}{ss_{xx}} = -0.842754 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 99.204$$

giving the fitted regression line $\quad y = \hat{\alpha} + \hat{\beta}x = 99.20 - 0.84x.$

Since the coefficient of x is negative, we can immediately conclude that the model predicts that the assessed stress level (y) will on average decrease with increasing skill level (x).

The predicted stress level for a student with skill level $x = 60$ is $\hat{\alpha} + \hat{\beta}x = 99.204 - 0.842754 \times 60 = 48.64$.

2. Summary statistics for the data set are:
$n = 5 \quad \sum x_i = 21 \quad \sum y_i = 12 \quad \sum x_i^2 = 111 \quad \sum y_i^2 = 46 \quad \sum y_i x_i = 69$
giving $\bar{x} = 4.2,\ \bar{y} = 2.4,\ ss_{xx} = 22.8,\ ss_{xy} = 18.6$ and $ss_{yy} = 17.2$.
Thus the least squares estimates are

$$\hat{\beta} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{ss_{xy}}{ss_{xx}} = 0.8518 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -1.0263$$

giving the fitted regression line $\quad y = \hat{\alpha} + \hat{\beta}x = -1.0263 + 0.8518x.$

Calculating the fitted values and residuals according to the formulae:

Predictor values (x_i)	1	3	4	6	7
Response values (y_i)	0	1	2	5	4
Fitted values ($\hat{y}_i$)	-0.2105	1.4211	2.2368	3.8684	4.6842
Residuals ($\hat{e}_i = y_i - \hat{y}_i$)	0.2105	-0.4211	-0.2368	1.1316	-0.6842

Finally, you were asked in this question to estimate σ^2 directly from the residuals, giving

$$\hat{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-2} = 0.6754.$$

The sum of the residuals is 0. This can easily be verified algebraically.

3. The summary statistics for the data set are:
$n = 7 \quad \sum x_i = 44 \quad \sum y_i = 9.6 \quad \sum x_i^2 = 344 \quad \sum y_i^2 = 13.36 \quad \sum y_i x_i = 57$, giving $\bar{x} = 6.285714,\ \bar{y} = 1.371429,\ ss_{xx} = 67.42857,\ ss_{xy} = -3.342857$ and $ss_{yy} = 0.1942857$.
Thus the least squares estimates are

$$\hat{\beta} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{ss_{xy}}{ss_{xx}} = -0.04957627 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1.68305085$$

giving the fitted regression line

$$y = \hat{\alpha} + \hat{\beta}x = 1.68 - 0.05x.$$

For a litter of size $x = 6$, this would predict an average piglet weight of

$$\mathrm{E}(Y|x = 6) = \hat{\alpha} + 6\hat{\beta} = 1.68 - 0.05 \times 6 = 1.38.$$

Once you have computed the linear regression analysis in **R** with the commands:

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

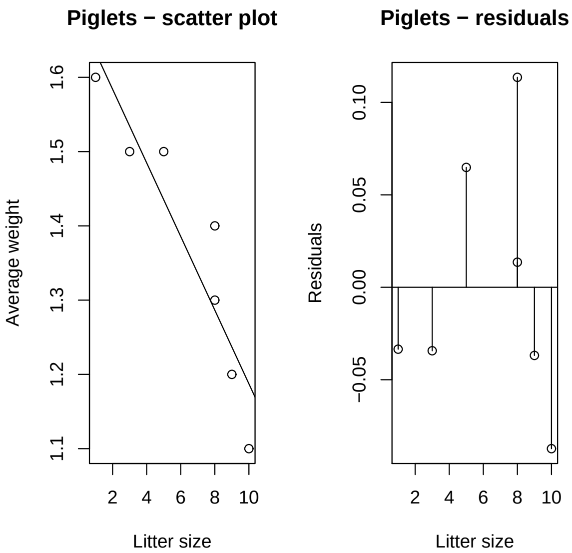
you can check your calculations using the command `coef` which gives the output:

```
> coef(piglets)
(Intercept) pig.littersize
 1.68305085 -0.04957627
```

The fitted values and the residuals can be calculated from the least squares estimates using the formulae $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ and $\hat{e}_i = y_i - \hat{y}_i$ or with the commands `fitted` and `residuals` in **R** which give the following output:

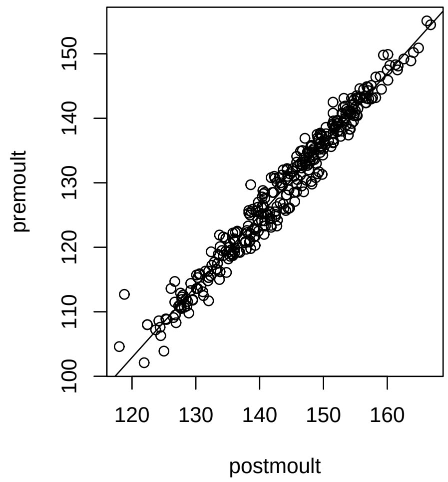
```
> fitted(piglets)
       1        2        3        4        5        6        7
1.633475 1.534322 1.435169 1.286441 1.286441 1.236864 1.187288
> residuals(piglets)
          1           2           3           4           5
-0.03347458 -0.03432203  0.06483051  0.01355932  0.11355932
          6           7
-0.03686441 -0.08728814
```

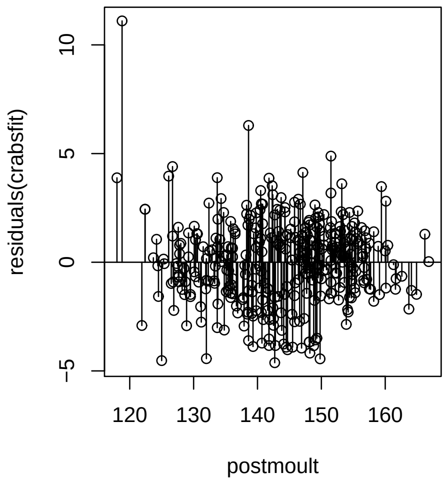
A scatter plot of the data is shown on the left below, together with the fitted regression line. There seems to be a reasonably good fit of the straight line to the data. A plot of the residuals against the corresponding spring rainfall is shown on the right. Note that there are two residual values at $x = 8$. There is no obvious sign of any systematic pattern. Overall, the fit is sufficiently good that we would have no reason to reject the linear regression model.



4. (a) Following the example in Section 4.7 of the notes to fit the regression line, and give the plot, below left:

```
> source("http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/crabs.R")
> attach(crabs)
> plot(postmoult,premoult)
> crabsfit<-lm(premoult~postmoult)
> abline(coef(crabsfit))
```





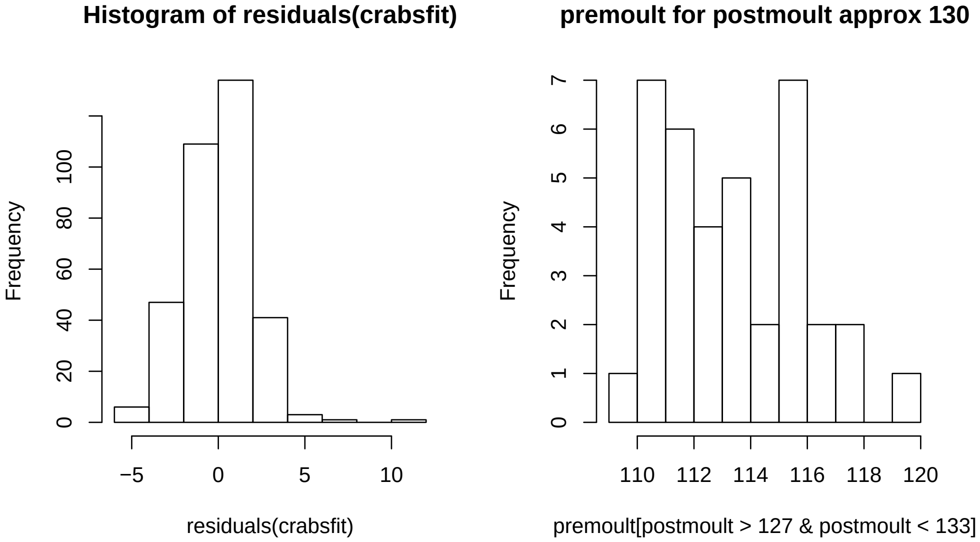
(b) We can print out the fitted values and residuals easily, but more usefully make a plot (above right) – the linear fit seems very good but there is a large outlier at the left hand end of the range:

```
> fitted(crabsfit)
       1        2        3        4        5        6
141.2521 127.7030 141.3623 121.0937 132.2194 137.6170
....
> residuals(crabsfit)
            1             2             3             4
 1.0478638826 -2.6030210395 -0.5622915245  1.3063033888
....

> plot(postmoult,residuals(crabsfit))
> segments(postmoult,0,,residuals(crabsfit))
> abline(h=0)
```

(c) Histogram of the residuals (below left): the impression is of a roughly symmetric shape centred at 0, but with several large positive outliers:

```
> hist(residuals(crabsfit))
```



(d) We can make the prediction for $x = 130$ by calculating $\hat{\alpha} + \hat{\beta} \times 130$ by hand:

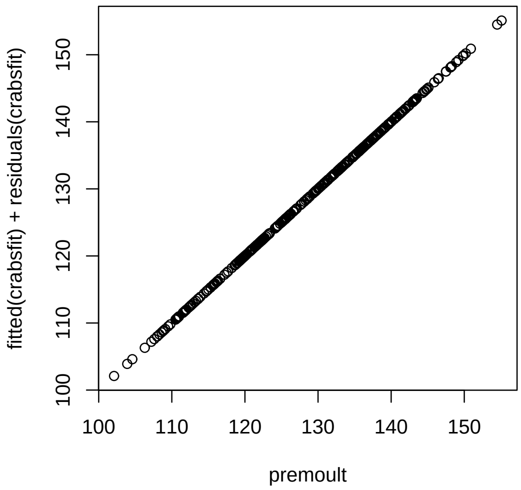
```
> coef(crabsfit)
(Intercept)   postmoult
 -29.268434    1.101554
> -29.268434+1.101554*130
[1] 113.9336
```

and compare to the histogram of data values for y from x near to 130 (above right); the prediction seems perfectly consistent with these data:

```
> hist(premoult[postmoult>127&postmoult<133])
```

(e) Finally we check the 3 assertions numerically – for the first we chose to make a plot (below), but there are other ways we could have done it, e.g. by typing `range(fitted(crabsfit)+residuals(crabsfit))`:

```
> plot(premoult,fitted(crabsfit)+residuals(crabsfit))
> sum(residuals(crabsfit))
[1] 1.170244e-14
> fit2<-lm(fitted(crabsfit)~postmoult)
> residuals(fit2)
            1             2             3             4
 1.603193e-13 -3.088760e-15  4.482824e-15 -1.041277e-15
....
```



5. The least squares estimates of the regression parameter(s) are defined to be the values that minimise the sum of squares of the differences between the observed y_i and the fitted values, i.e. the values that minimise $\sum_{i=1}^n (y_i - \mathrm{E}(Y_i|x_i))^2$. For this model, $\mathrm{E}(Y_i \mid x_i) = \gamma x_i$, so the least squares estimate of γ here is the value that minimises $\sum_{i=1}^n (y_i - \gamma x_i)^2$.

   From standard calculus, the minimising value satisfies the equation

$$\frac{\partial}{\partial \gamma} \sum_{i=1}^n (y_i - \gamma x_i)^2 = 0$$

   giving

$$\begin{aligned} 0 &= \textstyle\sum_{i=1}^n (y_i - \gamma x_i)(-2x_i) \\ \text{i.e.} \qquad 0 &= -2[\textstyle\sum_{i=1}^n y_i x_i - \gamma \sum_{i=1}^n x_i^2] \end{aligned}$$

   so the least square estimate under the new model is

$$\hat{\gamma} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

   Finally, the residual sum of squares is given by

$$\begin{aligned} RSS &= \textstyle\sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \textstyle\sum_{i=1}^n (y_i - \hat{\gamma} x_i)^2 \\ &= \textstyle\sum_{i=1}^n y_i^2 - 2\hat{\gamma} \sum_{i=1}^n y_i x_i + \hat{\gamma}^2 \sum_{i=1}^n x_i^2 \\ &= \textstyle\sum_{i=1}^n y_i^2 - 2\left(\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}\right) \sum_{i=1}^n y_i x_i + \left(\frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}\right)^2 \sum_{i=1}^n x_i^2 \\ &= \textstyle\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i x_i)^2}{\sum_{i=1}^n x_i^2} \end{aligned}$$

   Since the values of $y_i - \hat{y}_i$ $(i = 1, \ldots, n)$ satisfy the equation above determining $\hat{\gamma}$, there are effectively only $(n-1)$ independent values of $y_i - \hat{y}_i$ in the sum, so the appropriate estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{(n-1)} = \frac{\sum y_i^2 - (\sum y_i x_i)^2 / \sum x_i^2}{(n-1)}$$