Department of Statistics

COURSE STATS 330

Model answers for Assignment 2, 2008

1. *Load the data into R, and make a data frame **earthquakes.df** to contain the data. Check for any typographical errors. Create a new variable SeisMmt from the variables RootSeisMmt and ExpSeisMmt as described above and add it to the data frame. [5 marks]*

   The following code will load the data in and create a basic data frame:

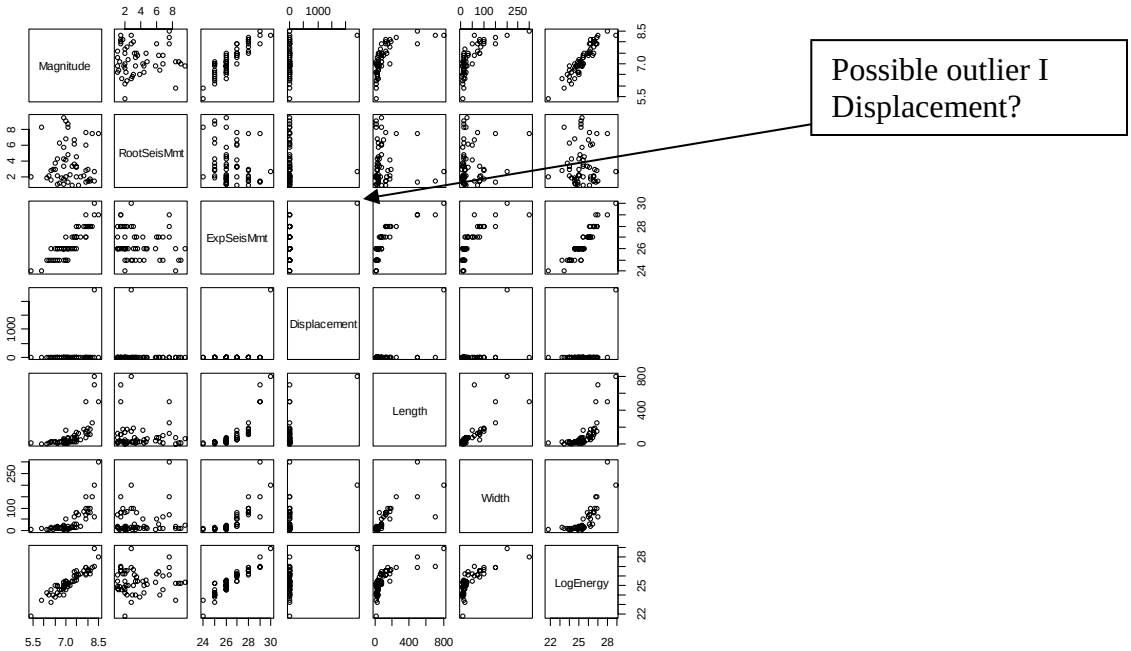
   ```
   earthquakes.df = read.table(file.choose(), header=T)
   ```

   Note that you can read the data direct from the web by using the URL of the data file

   ```
   mydata =
   "http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/earthquakes.txt"
   earthquakes.df = read.table(mydata, header=T)
   ```

   A pairs plot is a good screening device:

   ```
   pairs(arthquakes.df)
   ```



print out the variable Displacement, sorted into order:

```
> sort(earthquakes.df$Displacement)
 [1]    0.05    0.20    0.30    0.48    0.58    0.60    0.65    0.65    0.66
[10]    0.72    0.87    0.90    0.90    0.92    1.00    1.00    1.05    1.10
[19]    1.20    1.20    1.25    1.40    1.50    1.50    1.60    1.60    1.63
[28]    1.70    1.90    2.00    2.00    2.10    2.10    2.20    2.50    2.50
[37]    2.50    2.50    2.50    2.60    2.90    3.00    3.00    3.00    3.10
[46]    3.10    3.30    3.30    3.50    4.10    4.60    5.00    7.00 2400.00

> order(earthquakes.df$Displacement)
 [1] 37  6 38 24 41 20 34 50 21 48 25 35 44 43  4 54 22  8 28 33 27 51 31 52 49
[26] 53 36 40 14  7 17  1 45 11  9 13 19 32 46 39 47  2  3 23 10 12  5 30 29 42
[51] 15 16 26 18
```

Seems that earthquake 18 is the culprit – a look at the data sheet indicates that the correct value is 24, not 2400: the decimal point was omitted.

The rest of the data seem OK.

Fix up the error:

```
earthquakes.df$Displacement[18]=24
```

Now calculate the seismic moment and add it to the data set

```
M0 = earthquakes.df$RootSeisMm *
10^earthquakes.df$ExpSeisMmt

earthquakes.df = data.frame(earthquakes.df, SesMmt=M0)
```

Print out the first 3 rows as a check

```
> earthquakes.df[1:3,]
  Magnitude RootSeisMmt ExpSeisMmt Displacement Length Width LogEnergy  SesMmt
1       7.9         7.6         27          2.1    130    70     26.09 7.6e+27
2       7.5         4.6         26          3.0     35    13     25.60 4.6e+26
3       7.0         2.0         26          3.0     20    11     25.48 2.0e+26
```

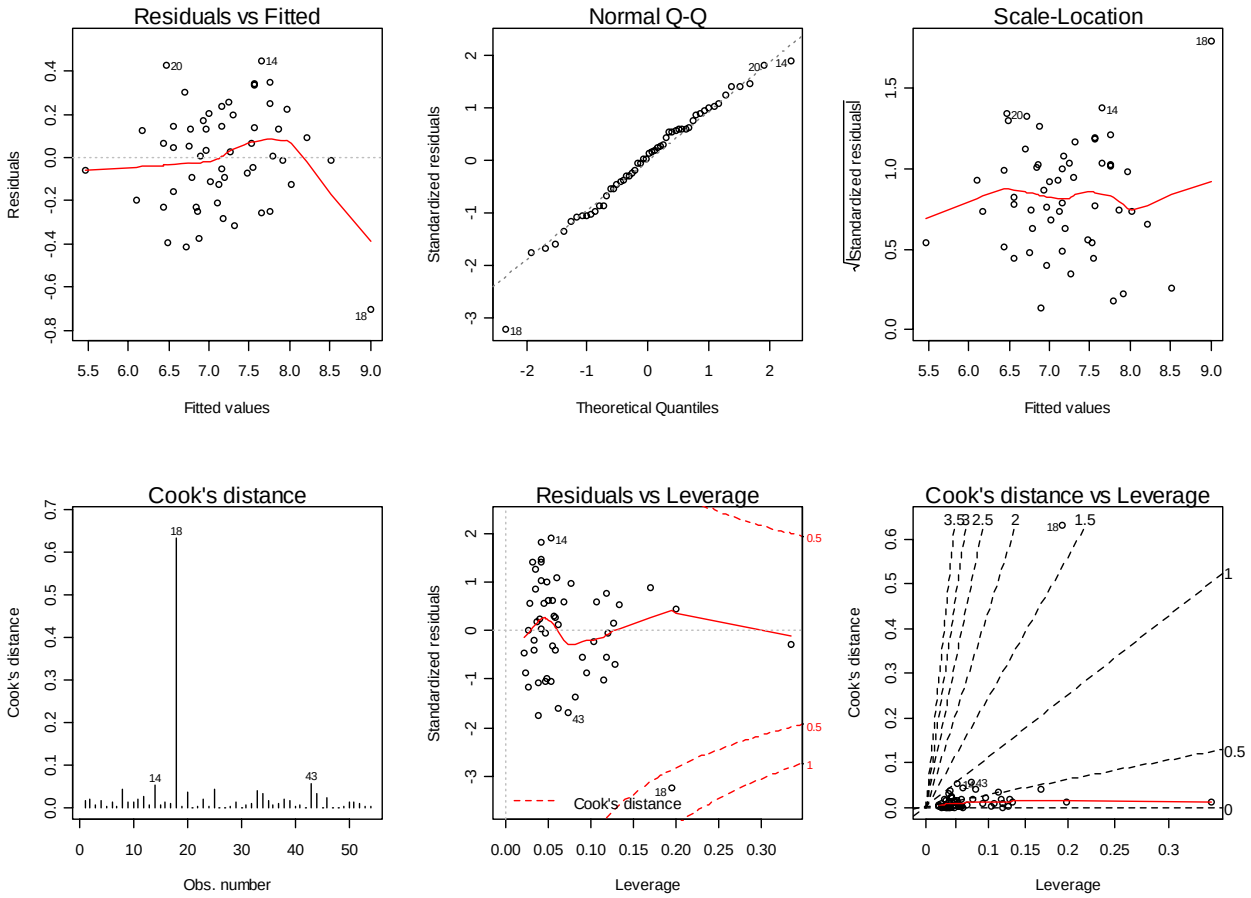
2. *Fit a regression to the earthquake data, using Magnitude as the response and the variables log(SeisMmt), log(Displacement) and log(Length) as the explanatory variables. Comment on the fit. [20 marks, 10 for the fitting and diagnostics and 10 for the discussion]*

Fit the regression:

```
earthquakes.lm = lm (Magnitude~log(SeisMmt) +
 log(Displacement) + log(Length), data=earthquakes.df)
```

draw the diagnostic plots

```
par(mfrow=c(2,3))
plot(earthquakes.lm, which=1:6) # Draw all six plots
```



The plots indicate that 18, even when corrected, is still an outlier. The magnitude is much less that is suggested by the model. If we delete point 18 and refit, we get an excellent fit (R^2 almost 90%).

Note that we can refit using the code

```
Earthquakes18.lm = lm (Magnitude~log(SeisMmt) +
log(Displacement) + log(Length), data=earthquakes.df,
subset=-18)
```

The residual versus fitted values plot shows no gross non-planarity. There is a hint of non-planarity in the coplots (not shown) using Length and Distance as conditioning variables, but putting the data into Ggobi and rotating again shows no great degree of non-linearity.

3. If the fit is unsatisfactory, modify the model to improve the fit. [10 marks]

Based on the discussion above, we will accept the model, with the outlier point 18 deleted.

4. Do you think the data are well described by the model (1)? That is, is it plausible that in model (2), we have $\beta_1 = \beta_2$ and $\beta_1 = -\beta_3$? [5 marks]

The summary of the final regression is

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.17739    2.48627  -0.474  0.63792
log(SeisMmt)       0.12995    0.04652   2.793  0.00742 **
log(Displacement)  0.30005    0.06554   4.578 3.23e-05 ***
log(Length)        0.04380    0.10223   0.428  0.67018
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.218 on 49 degrees of freedom
Multiple R-squared: 0.8978,     Adjusted R-squared: 0.8916
F-statistic: 143.5 on 3 and 49 DF,  p-value: < 2.2e-16
```

The estimated coefficients don't seem to support the idea that $\beta_1 = \beta_2$ and $\beta_1 = -\beta_3$. The coefficients of log(SeisMmt) and log(Displacement) are not very similar, and the The coefficients of log(SeisMmt) and log(Length) are not opposite in sign and of similar magnitudes. Still, the standard errors are quite big – so are these values consistent with the hypothesis? We can use the "330 function" `test.lc.` To test $\beta_1 = \beta_2$ we type

```
> cc = c(0, 1,-1,0)
> test.lc(cc, 0, earthquakes18.lm)
$est
[1] -0.1701021
$std.err
[1] 0.1050176
$df
[1] 49
$t.stat
[1] -1.619748
$p.val
[1] 0.1117055
```

So that there is no evidence against $\beta_1 = \beta_2$ (p-value is 0.1117). On the other hand, when testing $\beta_1 = -\beta_3$

```
> cc = c(0, 1,0,1)
> test.lc(cc, 0, earthquakes18.lm)
$est
[1] 0.1737500
$std.err
[1] 0.06110683
$df
[1] 49
$t.stat
[1] 2.843381
$p.val
[1] 0.006492787
```

So that there is strong evidence against $\beta_1 = -\beta_3$ (p-value is very small). Seems like the theory doesn't explain these data very well.

Extra Question for STATS 762 only

Test the hypothesis that $\beta_1 = \beta_2 = -\beta_3$ by performing a simultaneous test of the hypotheses that $\beta_1 = \beta_2$ and $\beta_1 = -\beta_3$, using the theory below.

The following code does the test:

```
> A = matrix(c(0, 1,-1,0,0, 1,0,1),2,4, byrow=T)
> est = A%*%coef(earthquakes18.lm)
> cov.beta = summary(earthquakes18.lm)$cov.unscaled*
                      (summary(earthquakes18.lm)$sigma^2)
> cov.mat = A%*%cov.beta%*%t(A)
> F.stat = sum(est* (solve(cov.mat)%*%est))/2
> df = earthquakes18.lm$df.residual
> p.val = 1-pf(F.stat, 2,df)
> F.stat
[1] 4.043426
> p.val
[1] 0.02369572
> A
     [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    1    0    1
```

Evidence against the hypothesis is reasonably strong – some doubt if the earthquake theory explains these data.