ACE2046 Quantitative Techniques
Statistical Computing

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

5 Multiple regression

In many investigations, we would like to relate the response to several explanatory variables. In this lecture we extend simple linear regression to allow the use of several X–variables.

5.1 Example: Multiple linear regression

(I'm sorry that this is not a particularly relevant example for ACE2046).

The data in table 3 come from an experiment to investigate how the resistance of rubber to abrasion is affected by the hardness of the rubber and its tensile strength. Each of 30 samples of rubber was tested for hardness (in degrees Shore – the larger the number, the harder the rubber) and for tensile strength (in kg per square cm) and was then subjected to steady abrasion for a fixed time. The weight loss due to abrasion was measured in grams per hour.

We want to use multiple regression to predict abrasion loss based on hardness and tensile strength. Here, our multiple (linear) regression model will be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where Y: abrasion loss, X_1: hardness and X_2: tensile strength.

Figure 20 shows scatter plots of abrasion loss against hardness and tensile strength. There seems to be a strong negative relationship with hardness but the relationship with tensile strength is less clear.

5.2 Assumptions

As with simple linear regression, there are some assumptions underlying multiple linear regression which can be checked in `Minitab` . These are:

- Existence of a straight line relationship between Y and *each* X (explanatory) variable. This is not always easy to check because of the effects of other variables but scatter plots can often give an idea.
- Errors normally distributed.
- Errors independent.
- Errors have the same variance.

We check these assumptions using the residuals.

5.3 Multiple linear regression in `Minitab`

To perform multiple linear regression in `Minitab` , we use

```
Stat -> Regression -> Regression
```

Using `Minitab` for the example and entering the dependent variable as the `Response` and the explanatory variables as `Predictors`, gives the following output:

```
Regression Analysis: Abrasion loss versus Hardness, Tensile Strength

The regression equation is
Abrasion loss = 885 - 6.57 Hardness - 1.37 Tensile Strength


Predictor              Coef  SE Coef       T      P
Constant             885.16    61.75   14.33  0.000
Hardness            -6.5708   0.5832  -11.27  0.000
Tensile Strength    -1.3743   0.1943   -7.07  0.000
```

Abrasion loss	Hardness	Tensile strength	Abrasion loss	Hardness	Tensile strength
(g/h)	(degrees S)	(kg/cm^2)	(g/h)	(degrees S)	(kg/cm^2)
y	x_1	x_2	y	x_1	x_2
372	45	162	196	68	173
206	55	233	128	75	188
175	61	232	97	83	161
154	66	231	64	88	119
136	71	231	249	59	161
112	71	237	219	71	151
55	81	224	186	80	165
45	86	219	155	82	151
221	53	203	114	89	128
166	60	189	341	51	161
164	64	210	340	59	146
113	68	210	283	65	148
82	79	196	267	74	144
32	81	180	215	81	134
228	56	200	148	86	127

Table 3: Abrasion loss data

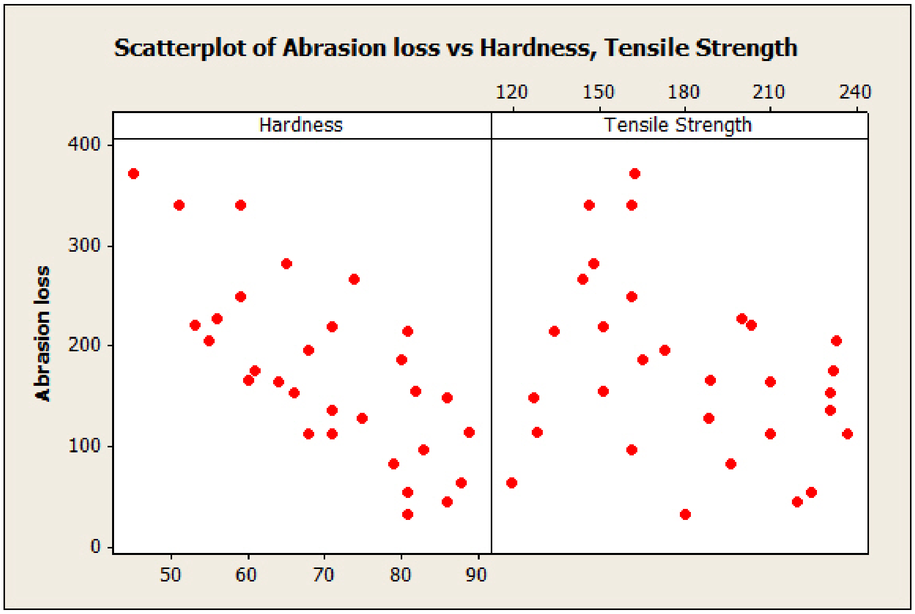

Figure 20: Plots of abrasion loss against potential explanatory variables

```
S = 36.4893   R-Sq = 84.0%   R-Sq(adj) = 82.8%


Analysis of Variance

Source          DF      SS     MS      F      P
Regression       2  189062  94531  71.00  0.000
Residual Error  27   35950   1331
Total           29  225011


Source            DF  Seq SS
Hardness           1  122455
Tensile Strength   1   66607


Unusual Observations

               Abrasion
Obs  Hardness      loss     Fit  SE Fit  Residual  St Resid
 14      81.0     32.00  105.55    9.12    -73.55     -2.08R
 19      88.0     64.00  143.38   14.83    -79.38     -2.38R

R denotes an observation with a large standardized residual.
```

So our multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

since we have two explanatory variables. Thus, referring to the `Minitab` output, we see our regression equation is

$$Y = 885 - 6.57X_1 - 1.37X_2 \ ,$$

where Y: abrasion loss and X_1 and X_2 are hardness and tensile strength respectively. So if we want to predict the abrasion loss of a synthetic rubber given that the hardness is 75 and the tensile strength is 200, then:

$$\hat{y} = 885 - 6.57 \times 75 - 1.37 \times 200 = 118.25 \ .$$

Notice that `Minitab` identifies two unusual observations. These should be checked.

5.4 Checking residuals

We should check the residuals in the usual way. Figure 21 shows the usual four-in-one plot for the example. There is some suggestion of asymmetry, and therefore non-normality, in the residuals but the main cause for concern here is in the plot against data order. There is a definite pattern here and some other variable seems to have changed during the course of the experiment and affected the results. We need to investigate this further.

It is also a good idea to plot the residuals against each of the explanatory variables. To do this we need to tell `Minitab` to store the residuals. These plots are shown in figure 22. The plot against hardness is reasonably acceptable, although there may be some suggestion of asymmetry. The plot against tensile strength suggests that the relationship might not be linear.

5.5 Fitting curves

We can introduce variables which are functions of other variables. For example, in the abrasion loss example, let $x_3 = x_2^2$. We can do such a calculation using the `Calculator` in `Minitab` .

Here is the result of this fit.

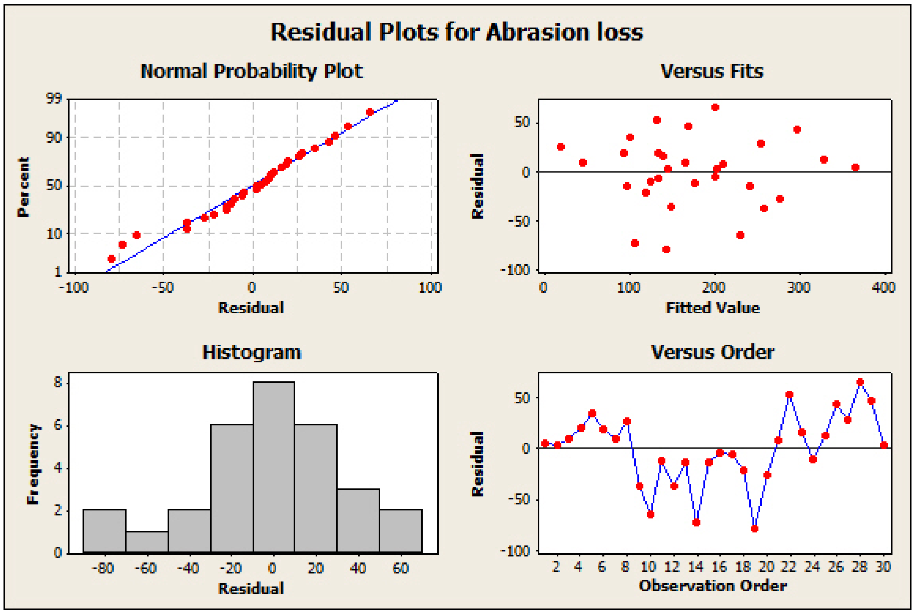

Figure 21: Residual plots

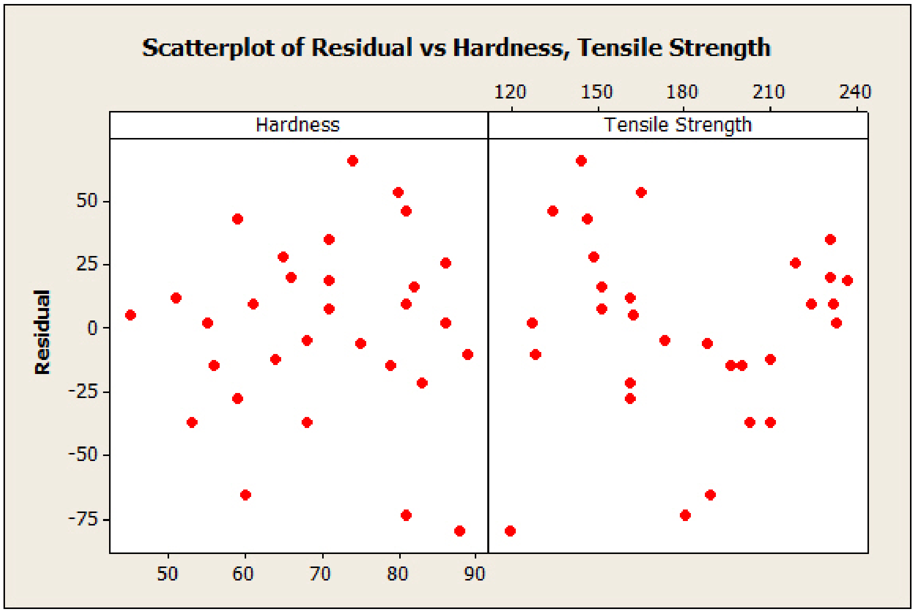

Figure 22: Residuals against explanatory variables

```
Regression Analysis: Abrasion loss versus Hardness, Tensile Stre, ...

The regression equation is
Abrasion loss = 1082 - 6.76 Hardness - 3.46 Tensile Strength
                + 0.00569 TS squared


Predictor             Coef   SE Coef       T      P
Constant            1082.5     232.3    4.66  0.000
Hardness           -6.7605    0.6239  -10.84  0.000
Tensile Strength    -3.461     2.377   -1.46  0.157
TS squared        0.005690  0.006457    0.88  0.386


S = 36.6414   R-Sq = 84.5%   R-Sq(adj) = 82.7%


Analysis of Variance

Source          DF      SS     MS      F      P
Regression       3  190104  63368  47.20  0.000
Residual Error  26   34907   1343
Total           29  225011


Source            DF  Seq SS
Hardness           1  122455
Tensile Strength   1   66607
TS squared         1    1042


Unusual Observations

               Abrasion
Obs  Hardness      loss     Fit  SE Fit  Residual  St Resid
 19      88.0     64.00  156.22   20.84    -92.22     -3.06R

R denotes an observation with a large standardized residual.
```

The residual plots are slightly better now but still not very good. There are other problems with these data!

5.6 Variable selection

When fitting a multiple regression model we might have a large number of potential variables which might be included in the model. We might wish to include only those where there is evidence that they have a real effect. We therefore need a method to select these. There are various methods available to do this. In this module we will adopt the following simple procedure.

1. Fit the model including all of the explanatory variables.

2. Check the P-values for the explanatory variables. Find the largest. If this is greater than 0.05 remove this variable. If it is less than 0.05 stop here.

3. Fit the model again with the variable removed and go back to step 2.

5.7 Air pollution example

These data are a subsample from a study of the effect of air pollution on lung function. The variables measured were age, sex, height, weight and forced vital capacity (FVC). FVC is the total volume of air in liters which an individual can expel regardless of how long it takes. We wish to predict FVC using age, sex, height, and weight. (Sex is coded 0/1). Using `Minitab` we get the following:

```
The regression equation is
FVC = - 3.34 - 0.609 Sex - 0.0437 Age + 0.158 Height - 0.00265 Weight

Predictor        Coef   SE Coef      T      P
Constant       -3.339     2.131  -1.57  0.126
Sex           -0.6085    0.2485  -2.45  0.020
Age          -0.04366   0.01250  -3.49  0.001
Height        0.15791   0.03602   4.38  0.000
Weight      -0.002655  0.005314  -0.50  0.621

S = 0.539470   R-Sq = 72.6%   R-Sq(adj) = 69.4%
```

So our regression equation is:

$$\text{FVC} \; = -3.34 - 0.609 \text{ Sex} - 0.0437 \text{ Age} + 0.158 \text{ Height} - 0.00265 \text{ Weight}$$

We should now check the residuals (not shown). Notice this time, that the p-value for weight is 0.621. This suggests that we should remove weight from our model and refit the data.

So on removing Weight, and fitting the remaining variables using `Minitab` gives:

```
FVC = - 3.06 - 0.547 Sex - 0.0430 Age + 0.145 Height

Predictor       Coef  SE Coef      T      P
Constant      -3.057    2.033  -1.50  0.142
Sex          -0.5473   0.2138  -2.56  0.015
Age         -0.04296  0.01228  -3.50  0.001
Height       0.14536  0.02552   5.69  0.000

S = 0.533655   R-Sq = 72.4%   R-Sq(adj) = 70.1%
```

Notice that:

- We have a new model, i.e.

$$\text{FVC} \; = -3.06 - 0.547 \text{ Sex} - 0.0430 \text{ Age} + 0.145 \text{ Height}$$

- The R^2 has reduced from 72.6% to 72.4%. The more variables in the model, the larger the R^2 value.
- The p-value for the constant is greater than 0.05 (it is actually 0.142). However, we usually keep the constant in the model, because the value zero is not particularly special.

5.8 Practical 5

Instructions

1. You have an individual reference number. You should use the data assigned to your reference number. Please write your reference number on your report as well as your name.
2. Please also mark on your report my name (Dr Malcolm Farrow) and "School of Mathematics & Statistics".
3. Answer both questions.
4. This assignment is to be submitted *via* NESS, by no later than 12.00 noon on Wednesday 14th November.
5. Write each solution in the form of a (brief) report. This should have an introduction to the problem, a description of the analysis and a clear statement of conclusions, illustrated, where appropriate, with graphs and tables.
6. Graphs should be properly labelled with appropriate axis labels etc.
7. In your report explain (briefly) how you obtained your results using Minitab so that, if anything has gone wrong, I might be able to see where it is.

Questions

1. Data are provided which refer to ice cream consumption over 30 four-week periods in the 1950s. The data are:

   - Y: ice cream consumption (pints per capita).
   - X_1: the price of ice cream per pint ($).
   - X_2: Average weekly family income ($).
   - X_3: Mean temperature (°F).

   The aim is to use a regression of Y on X_1, X_2 and X_3 to investigate how Y depends on these variables.

   The data file `icecreamdat.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

   ```
   File -> Other files -> Import special text
   ```

   entering `c8-c40` in the box marked `Store data in column(s)` and entering

   ```
   http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/icecreamdat.txt
   ```

   for the file name.

   - The values of X_1 are in column `c8`. Name this column `Price` or some other suitable name.
   - The values of X_2 are in column `c9`. Name this column `Income` or some other suitable name.
   - The values of X_3 are in column `c10`. Name this column `Temperature` or some other suitable name.
   - Your consumption figures are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column `c20`. Name your column `Consumption` (or some other suitable name).

   (a) Use scatterplots to see whether a linear model is reasonable. In `Minitab` select

   ```
   Graph -> Scatterplot -> With regression -> OK
   ```

- Under `Y variables` enter your `Consumption` column in each of the first three rows.
- Under `X variables` enter your `Price` column in the first row, your `Income` column in the second row and your `Temperature` column in the third row. Select `Multiple graphs -> In separate panels of the same graph` .
- Click `OK` twice.

(b) Fit a regression of Y on $X_1,\ X_2,\ X_3$.
Select

```
Stat -> Regression -> Regression
```

- For the `Response` enter your `Consumption` column.
- For the `Predictors` enter your `Price` , `Income` and `Temperature` columns.

(c) As described in the lecture, use a variable selection procedure to find a model containing only variables with significant tests. State clearly your final model.

(d) For your final model produce residual plots to examine the model assumptions. Look at your residual plots and comment.

(e) Report your conclusions clearly.

2. Data are provided which refer to an experiment to calibrate a near infrared reflectance instrument for the measurement of protein content of ground wheat samples. The data are:

- Y: the protein content (%) determined by a standard method.
- $X_1, \ldots, X_6$: reflectance measurements at six wavelengths.

The aim is to use a regression of Y on $X_1, \ldots, X_6$, or some subset of them to predict Y.

The data file `groundwheat.txt` can be downloaded from the module Web page (via Blackboard or otherwise) or it can be read directly into `Minitab` using

```
File -> Other files -> Import special text
```

entering `c5-c40` in the box marked `Store data in column(s)` and entering

```
http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/groundwheat.txt
```

for the file name.

The values of $X_1, \ldots, X_6$ are in columns `c5-c10`. Name these columns `X1, X2, X3, X4, X5, X6` or some other suitable names.

Your protein values are in the column corresponding to your reference number. For example, if your reference number is 20 then your data are in column `c20`. Name your column `Protein %` (or some other suitable name).

(a) Carry out a regression analysis, going through steps similar to those for question 1. Select a suitable model by eliminating variables which do not give significant tests. Use residual plots to look for any signs of problems with the model assumptions.

(b) Report your conclusions clearly.

Student Reference Numbers

In the practicals for this week and later weeks, in at least some of the exercises, each student will have different data. In order to identify which data each student should have, each student is given a reference number according to the table below.

Agapiou	Kristy Marie	11
Au	Hoi Ching	12
Begin	Abigail Ruth	13
Breininger	Stella Panagio	14
Cepelis	Aivaras	15
Choo	Siew Li	16
Clarke	Timothy Oliver	17
Crawley	Michael	18
Darwin	Darwin	19
Drayton	Chloe Roseanne	20
Faleti	Hannah Ifedolapo	21
Golding	Carolina De Arouca	22
Grant	Georgia Kate	23
Hui	Angeline Kosim	24
Jenkins	Grace Evangeline	25
Linn	Sophia Campbell	26
Lo	Kin Wa	27
MacKinlay	Brittney Lauren	28
Monroe	Sophie Isabella	38
Obertelli	Georgia May	29
Rathod	Neel Harish	30
Rowland	Maisie Katharine	31
Saberioon	Ghazaleh	32
Song	Anqi	33
Tam	Wai Tsun Rachel	34
Trikerioti Chatziioannou	Danai	35
Wright	Jessica Laura Kate	36
Zawarska	Santana Anna	37