ACE2046 Quantitative Techniques
Statistical Computing

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

4 Correlation and linear regression

In this section, we study relationships between random variables measured together. Many experiments focus on establishing links between variables, for example:

- Dosage of drug and recovery rate;
- Quality of fertiliser and growth of plant;
- Nutritional content of food and weight gain.

The data take the form of pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ collected together in an investigation.

We will look at two closely related techniques **Correlation** and **Regression**.

4.1 Correlation

The **sample correlation coefficient**, r, is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \times \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

When we calculate r using data, we typically regard it is an estimate of the *population correlation coefficient* ρ. It can be shown that $-1 \leq r \leq +1$, where $r = \pm 1$ corresponds to a perfect linear relationship and $r = 0$ is complete absence of such a relationship. The same is true for ρ.

We *can* calculate this by hand, but Minitab has a facility to do this for us:

```
Stat -> Basic Stats -> Correlation
```

A correlation coefficient of exactly -1 corresponds to a perfect **negative linear association**, or **indirect** association between variables (as one increases, the other decreases). A coefficient of exactly $+1$ corresponds to a perfect **positive linear association**, or **direct** association (as one increases, the other also increases). Figure 15 shows some examples.

Scatter plots between the two variables can be used to ascertain any such relationships, whilst the correlation coefficient quantifies this relationship.

Example

We are interested in the relationship between the number of calories consumed by parents and by their 5 year old children. Figure 16 shows a scatter plot between the number of calories consumed by parents and children.

We can then test to see whether there is **significant** evidence in the data of a linear association between the two variables. We set up the null and alternative hypotheses:

$$H_0 : \rho = 0 \quad H_1 : \rho \neq 0,$$

We test to see whether the sample correlation coefficient indicates that the population correlation coefficient is different from zero. For the data in figure 16, the `Minitab` output is shown below.

```
Correlations:
Adult, Child Pearson correlation of Adult and Child = 0.853

P-Value = 0.003
```

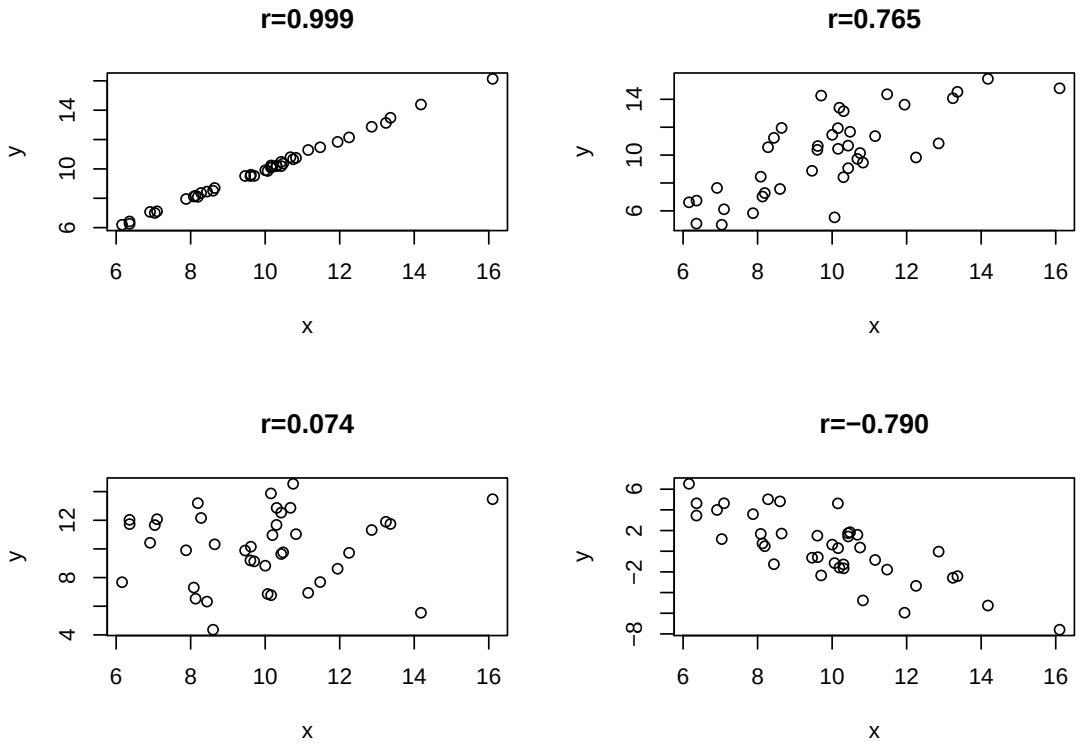

Figure 15: *Scatter plot to demonstrate correlation. The figures have sample correlation coefficients of 0.999, 0.765, 0.074,* -0.790 *respectively.*

So in this case we have $r = 0.853$. Since the p–value is less than 0.05, we reject H_0 in favour of H_1. There is evidence that there is a linear relationship between the amount of calories consumed by parent and child.

Note that the test used by Minitab is based on the assumption that the two variables have normal distributions (technically, that they have a bivariate normal distribution).

Spurious correlations

Correlation is a useful tool, but it can easily mislead. It is important to realise that a high correlation does not necessarily mean that there is a *causal* link (for example storks and birthrate in Denmark). Equally, a low correlation can hide a strong, but non–linear relationship. Two variables might also be linearly dependent given a third, hidden variable (for example number of pubs and churches in English towns).

Also remember to plot a scatter plot of the data. For example, in Figure 17, all plots have a correlation of 0.816.

4.2 Simple linear regression

In this model, we regard one variable, Y, as **dependent** and the other, X, as **explanatory**. The aim is to formulate a model for predicting Y from X. We write

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are unknown parameters (intercept and slope respectively), and ϵ is a random variable' known as an "error", representing the scatter about the line. We usually estimate β_0 and β_1 by the method of "least squares". We can do this by hand or, again, we can use Minitab.

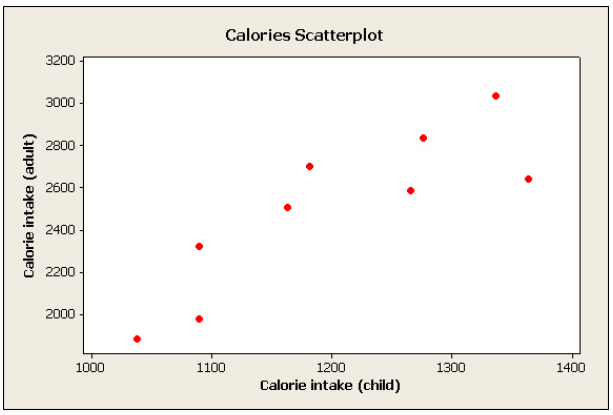

Figure 16: *The number of calories consumed by parents and their 5 year old children.*

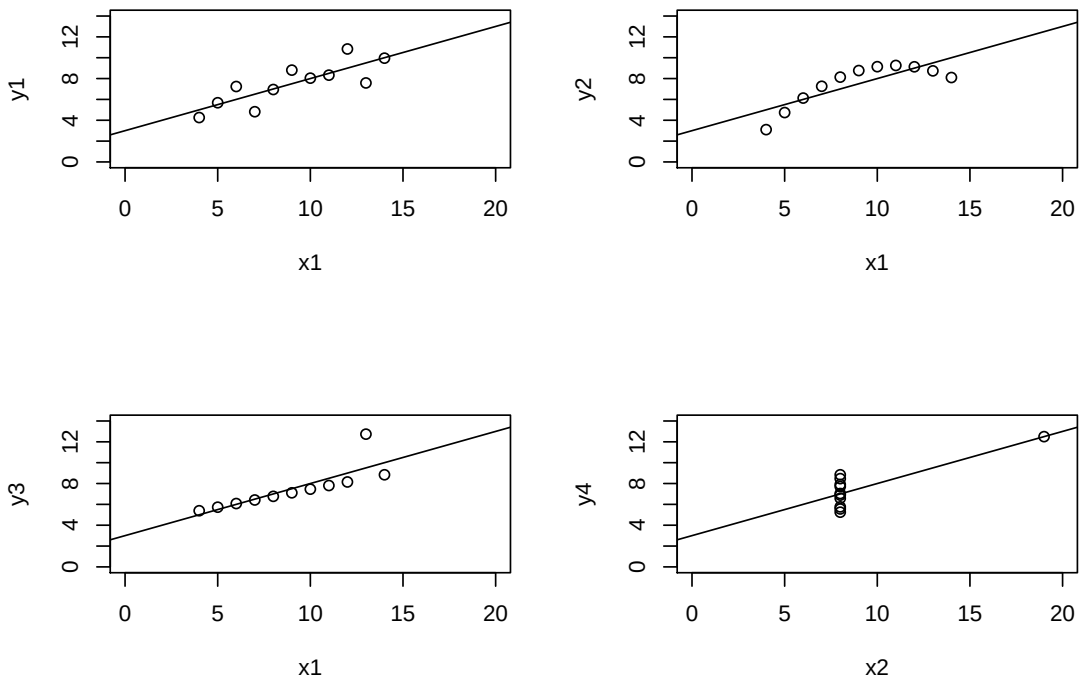

Figure 17: *Each of these scatter plots has a correlation coefficient of r=0.816, but some are clearly not linear.*

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

The key assumptions in this regression analysis are:

- Existence of a straight line relationship (look at scatter-plot);
- Errors normally distributed;
- Errors are independent;
- Errors have the same variance.

The following example demonstrates a regression analysis in `Minitab` .

4.3 Example: Regression analysis

An American pizza chain carried out a survey which included measurements of sales and of the local student population. Ten restaurants were included, and the results are shown below.

Restaurant	1	2	3	4	5	6	7	8	9	10
Sales Y ($1000)	58	105	88	118	117	137	157	169	149	202
Students X ('000)	2	6	8	8	12	16	20	20	22	26

Before we proceed with any regression analysis, we need to check that there exists a linear relationship between the two variables. Figure 18 shows a scatter plot of pizza sales against number of students. From this, there certainly appears to be a positive linear association.

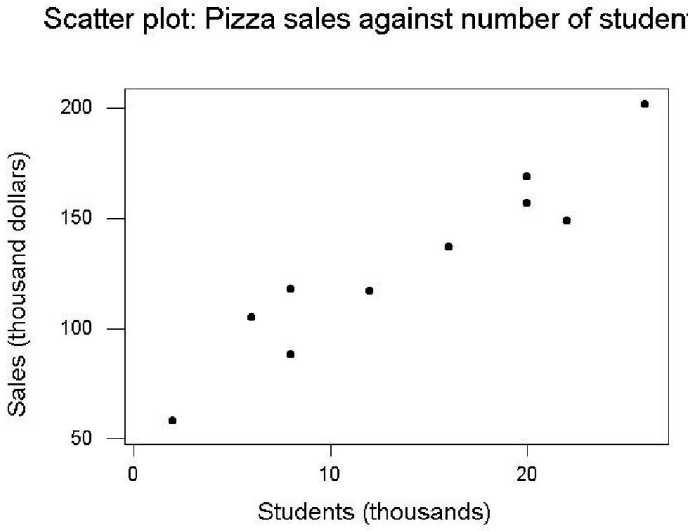

Figure 18: Scatter plot of pizza sales against number of students

Regression Output

The table below shows the `Minitab` output from a regression analysis of the pizza sales data.

```
Regression Analysis: Sales versus Students

The regression equation is
Sales = 60.0 + 5.00 Students

Predictor  Coef   SE Coef  T     P
Constant  60.000  9.226    6.50  0.000
Students  5.0000  0.5803   8.62  0.000

S=13.83  R-Sq = 90.3%  R-Sq(adj) = 89.1%
```

From this output, you can see that the regression equation is given as

$$\text{Sales} = 60.0 + 5.00\text{Students}.$$

Thus our estimates of intercept and slope are $\beta_0 = 60$ and $\beta_1 = 5$ respectively. Both the intercept ("`Constant`" in the table) and slope ("`Students`") have p–values of less than 0.05, so we can safely say that both estimates are significantly different from zero. The R^2 statistic measures the proportion of variation explained by the analysis. Here, we have an R^2 value of 90.3%, which is reassuring ($R^2 = 100\%$ indicates that *all* points lie *exactly* on the regression line). Note, when we have a single explanatory variable, as in this case, $R^2 = r^2$, which is just the correlation coefficient squared.

Residual diagnostics

To check whether our regression model assumptions were correct, we must check our **residuals**. The residuals are estimates of the 'errors' about the fitted regression line (the ϵ in the model). In fact, if our estimates of β_0 and β_1 are $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively, then the residual from observation i is

$$\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

If our model is appropriate, then the residuals should be (approximately) independent, normally distributed and have the same variance. Figure 19 shows some residual plots constructed in `Minitab` . Both the normal plot and histogram of residuals verify the assumption that the residuals are normally distributed. The scatter plot of residuals versus fitted values also suggests that the residuals are of equal varince, and that the relationship is straight, since there seems to be no structure in the plot (random scatter). A lack of independence in the errors (or other problems) might show as patterns in the plot of residuals versus observation order.

Making predictions

Suppose we wish to predict the sales at another restaurant,based on the size of the local student population. For example, if we'd like an estimate of pizza sales in a town with 14,000 students, then our estimate

$$\text{Sales} = 60.0 + 5.00 \times 14 = 130,$$

i.e. about $130,000. We can also put limits on it using Minitab, to get an estimated value of between (120,140).

To use Minitab for prediction, click on the `Options` box under the `Regression` dialog. Then enter your new x value (14 in this case) in the box `Prediction interval for new observations`. Minitab will then calculate a *confidence interval* (`CI`) for $\beta_0 + \beta_1 x$ and a *prediction interval* (`PI`) for the new y value. The prediction interval is wider than the confidence interval because it allows for the new error ϵ..

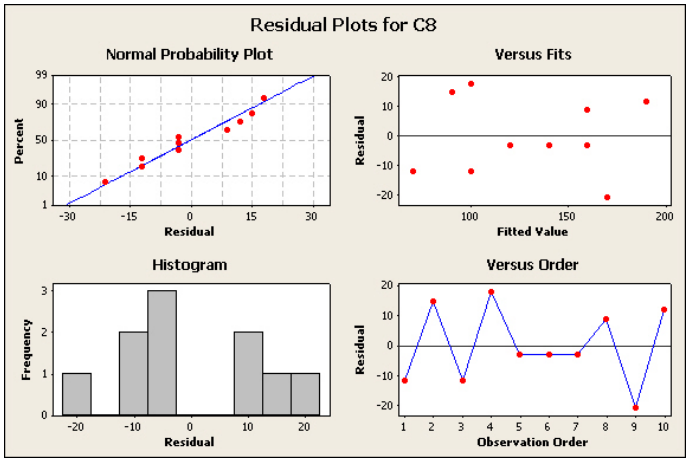

Figure 19: Residual diagnostics of regression analysis

4.4 Practical 4

Instructions

1. You have an individual reference number. A list of refernce numbers is given on the last page of this handout. 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 31st October.

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.

8. Sixty percent of the marks will be awarded for successfully doing the calculations, obtaining the correct results and showing the required graphs. The remaining forty percent will be awarded for good presentation, clear explanations and conclusions, good comments *etc.*

Questions

1. The data come from an experiment used in the calibration of an instrument to measure blood lactic acid concentration.Samples were prepared with known true lactic acid concentrations and several measurements with the instrument were made at each true concentration. The data are:

   - X: the true concentration of lactic acid.
   - Y: the concentration as given by the instrument.

   The units are mmol/l.

   The aim is to use a regression of Y on X to see how the instrument reading depends on the true value.

The data file `lactic.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 `c10-c40` in the box marked `Store data in column(s)` and entering

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

for the file name.

The values of X are in column `c10`. Name this column `True concentration` or some other suitable name.

Your instrument readings 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 `Instrument reading` (or some other suitable name).

(a) Use a scatterplot to see whether a straight-line model is reasonable. In `Minitab` select

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

- Under `Y variables` enter your `Instrument reading` column.
- Under `X variables` enter your `True concentration` column.

(b) Calculate the sample correlation coefficient. In `Minitab` select

`Stat -> Basic Stats -> Correlation`

In the `Variables` box enter your `True concentration` column and your `Instrument reading` column.

(c) Fit a regression of Y on X. Before going any further enter the numbers `2 4 6 8 10` in the first five cells of a new column, for example `c41`. Name the column `New concentrations`.

Then select

`Stat -> Regression -> Regression`

- For the `Response` enter your `Instrument reading` column.
- For the `Predictors` enter your `True concentration` column.
- Select

  `Options -> Prediction intervals for new observations`

  and enter your `New concentrations` column then select `OK` .
- Select

  `Graphs -> Four in one`

(d) Look at your residual plots and comment. Do you think that there might be a problem with the model at low concentrations?

(e) Report your conclusions clearly.

(f) Of course, we really want to be able to predict the *true* concentration from and instrument reading. Can you suggest how we might do this?

2. Data are provided which refer to 30 eleven-year-old girls at a middle school. The data are:

- X: the height in cm.
- Y: the weight in kg.

The aim is to use a regression of Y on X to find prediction intervals for Y given X.

The data file `height.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 `c10-c40` in the box marked `Store data in column(s)` and entering

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

for the file name.

The values of X are in column `c10`. Name this column `Height` or some other suitable name.

Your weight 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 `Weight` (or some other suitable name).

(a) Carry out a regression analysis, going through steps similar to those for question 1. Find prediction intervals for the weights of girls with heights 130, 135, 140, 145, 150, 155, 160. Looking at the residual plots, do you see any sign of problems with the model assumptions?

(b) Repeat the regression analysis but this time using the *natural logarithms* of the heights and weights.

i. Select

`Calc -> Calculator`

- Enter a suitable column, eg. `c41` , for `Store result in variable` .
- Find `Natural log` from the list of functions and double click on this. You should get `LN()` in the `Expression` box.
- You now need to enter your Weight variable inside the brackets. You can do this by double clicking on it in the list of variables on the left.
- Click `OK` .
- Name your new variable `Log weight` (or some other suitable name).

ii. By a similar process, calculate the logarithms of the heights and store them in a column (eg. `c42`) named `Log height` (or some other suitable name).

iii. To find the prediction intervals you will need the logarithms of the heights which you used before. Put these in, eg., `c43` .

iv. Carry out the regression analysis using log weight as the response (Y) variable and log height as the predictor (X) variable. Store the prediction intervals by ticking the appropriate bix.

v. You will need to use the calculator to convert your stored prediction interval limits back from the log scale. Use the function `Exponential` to do this.

vi. Do you think the residual plots look better this time?

3. 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. Please make a note of your reference number because we will use the same numbers in later weeks. **Please note that some of the numbers have changed since last week.**

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