THE UNIVERSITY OF SYDNEY

MATH1905 STATISTICS ADVANCED

Semester 2 **Tutorial Week 3** 2012

1. In `R` type `data(airquality)` to obtain the 'airquality' data set and `help(airquality)` to read infos about this data set.

(a) Type `cor(airquality,use='complete')` to obtain the matrix of pairwise correlations. What are the 2 most highly correlated variable pairs?

(b) Type `pairs(airquality)` to obtain the paired scatter plots. Comment on the plots as well as on the pairwise correlations.

(c) Type `attach(airquality)` to obtain the 6 variables from the data frame. Fit a linear model $y = a + bx$ using $y =$ Ozone and $x =$ Temp. Comment on the residual plots.

Hints: There are some missing values in the y = Ozone variable which create problems when you'll try to `plot(x,residuals)`. The following R-code fixes the problem:

```
> attach(airquality)
> x = Temp; y = Ozone;
> lm.out = lm(y~x)
> plot(x,lm.out$resid) # check the error message that occurs
> x1 = x[!is.na(y)]
> plot(x1,lm.out$resid) # note that no error message occurs
```

(**Explanation**: Square brackets [..] select components of vectors in R. E.g. x[1] returns 1st component of x. The command `!is.na(y)` asks whether or not a component of y is missing (=NA). So `[!is.na(y)]` selects the non missing y-components. Hence, `x1 = x[!is.na(y)]` shortens the x vector accordingly and is stored in `x1 = ....`)

(d) Fit a linear model $y = a + bx$ using $y =$ Ozone and $x =$ Wind. Comment on the residual plots.

(e) Assess the quality of the 2 previous linear models. Which of the 2 models better predicts Ozone concentration?

(f) Based on the Ozone $= a + b *$ Temp model, what is the expected Ozone concentration (in ppb) on a day where the temperature hits 75 degrees F.

2. The relation between carbon monoxide concentration and traffic density is of environmental interest. In the following table density (vehicles per hour to the nearest 500 vehicles) and carbon monoxide concentration (CO) in ppm are given for a particular street corner in Newtown.

x: Traffic density(in thousands)	y: CO (in ppm)
1.0	9.0 6.8 7.7
1.5	9.6 6.8 10.3
2.0	12.3 11.8
3.0	20.7 20.2 21.6 20.6

To enter the data in R you have to create two vectors as follows:

```
> x = c(1,1,1,1.5,1.5,1.5,2,2,3,3,3,3)
> y = c(9,6.8,7.7,9.6,6.8,10.3,12.3,11.8,20.7,20.2,21.6,20.6)
```

(a) Draw a scatter plot of the data.

(b) Fit a linear model $y = a + bx$ relating CO (y) to traffic density (x).

(c) What is the expected CO concentration on a day with traffic density 2,500?

(d) Writing $\hat{y}_i = a + bx_i$, the estimate of y at x_i, and $\hat{e}_i = y_i - \hat{y}_i$, the residual at x_i, draw a residual plot and comment on it.

(e) Check your results with R.

3. Recall $\hat{e}_i = y_i - \hat{y}_i = (y_i - \bar{y}) - b(x_i - \bar{x})$. Show that $\sum_{i=1}^{n} \hat{e}_i = 0$.

Assignment 1 for MATH1905 STATISTICS (due on Tuesday, 23 August, in week 5) will consist of selected questions from the Problem Sheets for weeks 1, 2, 3, 4.

1. J.B. Haldane is responsible for showing how carbon dioxide levels in blood influence breathing rates by affecting the acidity of the blood. In one experiment he administered varying doses of sodium bicarbonate with the following results:

Dose (in grams):	x	30	40	50	60	70	80	90	100
Breathing rate:	y	16	14	13	13	11	12	9	9

Use R to help answer the following.

(a) Calculate the coefficient of correlation.

(b) Find the least squares regression line of breathing rate on dose.

(c) What breathing rate would you predict for a dose of 85g?

(d) What proportion of variance of the response variable can be explained by the least squares regression line?

(e) Use R to construct a scatterplot of the data and mark the regression line on the diagram.

2. Consider the paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(a) Show that $\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$.

(b) For the linear regression line $y_i = a + bx_i$ the ith residual is $\hat{e}_i$. Show that $\sum_{i=1}^{n} \hat{e}_i x_i = 0$.

(c) Deduce from the previous question that $\{x_i\}$ and $\{\hat{e}_i\}$ are uncorrelated.

Extra questions to try: *A Primer of Statistics:* Ch I page 37 Q19-28.