# DEPARTMENT OF STATISTICS
# Course STATS 330: Advanced Statistical Modelling
# Tutorial Sheet 2: August 5, 2010

This tutorial is designed to give you practice in the following:

- Fitting a regression model
- Making predictions
- Handling missing data
- Testing a linear combination

In this tutorial we will be using the **moisture evaporation data** , and the **cherry data** on the website.

After completing the tutorial, you should have the skills you need to attempt Assignment 2.

## Task 1: Read in the data

Download the data set **`evap.txt`** from the web. Make a data frame **`evap.df`**. The data consists of daily readings of moisture evaporation from soil, along with certain environmental variables. The variables are

**evap:** the amount of moisture evaporating from the soil in the 24 hour period (response)

**maxst:** maximum soil temperature over the 24 hour period
**minst:** minimum soil temperature over the 24 hour period
**avst:** average soil temperature over the 24 hour period

**maxat:**maximum air temperature over the 24 hour period
**minat**: minimum air temperature over the 24 hour period
**avat:** average air temperature over the 24 hour period

**maxh:** maximum humidity over the 24 hour period
**minh:** minimum humidity over the 24 hour period
**avh:** average humidity over the 24 hour period
**wind:** average wind speed over the 24 hour period.

Some of these data have been recoded. The response variable of interest is **`evap`**, and we want to model this in terms of the environmental variables.
You can use the form

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

to get data directly from the web.

## Task 2: make a pairs plot

Make a pairs plot of the data. Are any of the explanatory variables closely related to each other? Calculate the correlations between the variables to confirm what you see in the plot.

```
> pairs(evap.df)
> cor(evap.df)
```

## Task 3: Fit the regression

Fit a regression model using **evap** as the response and all the other variables as explanatories.

```
evap.lm <- lm(evap~.,data=evap.df)
summary(evap.lm)
```

Which of the explanatory variables are insignificant? Is this because (a) they are unrelated to the response, or (b) they are related to the response but are also strongly related to other explanatory variables? Hint: Look at the correlations.

## Task 4: Predict the soil evaporation

In view of the high correlations, added variable plots and VIF's, lets retain only one temperature variable, say **maxst**, since this has the highest correlation with the response. (See Lectures 14 and 15 for more on how to select variables). Using the variables **humid**, **wind** and **maxst**, predict the evaporation when avh = 400, wind = 300, maxst = 200.

```
newevap.lm <-lm(evap~maxst + avh + wind,data=evap.df)
new.df<-data.frame(avh=400, wind=300, maxst=200)
predict(newevap.lm, new.df, interval="p")
```

## Task 5: Dealing with missing values

Suppose there were some missing values in the data. Introduce a missing value for the variable **avst** in the first observation:

```
evap.df[1, 3]=NA
```

Then, fit the regression with all the variables as you did before:

```
evap.lm <- lm(evap~.,data=evap.df)
summary(evap.lm)
```

The degrees of freedom are different because R automatically dropped the first (incomplete) observation.

Now fit a model with **maxst, avh** and **wind** as before:

```
newevap.lm <- lm(evap~ maxst + avh + wind, data=evap.df)
summary(newevap.lm)
```

This has used all the data since the model doesn't require **avst.**

Suppose we compare the two models using **anova.** We get an error because different sets of data have been used. One solution to this is to make a data frame containing only the cases with complete data.

You can use the following code to delete the incomplete observations from the data set:

```
temp.mat = !is.na(evap.df)
# look at temp.mat to see what this does
use = apply(temp.mat, 1 ,all)
# check out apply in the documentation
no.missing.evap.df = evap.df[use,]
# now all is ok
```

## Task 6: Download the "330 functions" from the web site

Download the file of 330 functions (file **R330.txt**) from the 330 web site by clicking on the link "Useful R functions" and saving the file. Then, in R, to load the functions, type

**source(*filename*)**

where *filename* is the name of your file, enclosed in "". Or, if you are connected to the web, type

**source("http://www.stat.auckland.ac.nz/~lee/330/R330.txt")**

to load the file directly from the website.

## Task 7: Test a linear combination

Create the file **cherry.df** from the cherry data. Then test the hypothesis that $\beta_1+\beta_2 = 3$ which was discussed in Lecture 6. Use the function **test.lc** in the "330 functions".