In the first part of the practical you will fit a multiple regression models to examine the factors that influence vitamin C levels. The aim of this exercise is to familiarise you with fitting models with mutiple predictors in Stata and strengthen understanding of why results change when different predictors are added.

In the second optional part you will explore the matrix algebra involved in multivariable linear regression.

**Data**  
The data are in the file vitC.dta. This can be found in folder "Linear regression datasets" on the Moodle page.

The variables in this dataset are:

serial Patient identifier

age Age of subject in years

height Height in metres

cigs Smoking status (0= non-smoker; 1=smoker)

weight Weight in kg

sex Sex (0=male; 1=female)

seruvitc Serum vitamin C level (µmol/l)

ctakers Vitamin C supplements taken (1=yes, 0=no)

# Initial analysis

As usual open Stata and get set up to start the practical. First we'll carry out some preliminary analyses.

1) Summarise the data to get a feel for the distribution of each variable (think about the best way to describe each variable). Check whether there are any missing values.

2) Examine the correlation between vitamin C levels and the other continuous variables, using plots or summary statistics

3) Tabulate vitamin C by the categorical variables

4) In your group discuss whether any of the variables look like they might be important predictors of vitamin C levels.

# Univariable regression

1) Fit separate regression models relating vitamin C levels separately to each potential predictor variable.

2) How should we interpret the slope coefficient in each of these models?

3) What do you conclude about the association between each of the predictors and vitamin C level? Do these conclusions match what you expected from the preliminary analysis?

# Multivariable regression

1) Fit a regression model relating vitamin C levels jointly to sex, smoking status and use of vitamin supplements.

2) Write out an equation to describe this model.

3) Carefully interpret the parameter estimate for smoking and write a short paragraph summarising your findings on how smoking influences serum vitamin C.

4) Share your paragraphs with each other and discuss any differences in what you included and how you interpreted the results.

5) Compare the coefficient for smoking from the multiple regression model to that in the simple linear regression models relating smoking to vitamin C levels.

6) Discuss with your colleagues the possible reasons for any material changes in results.

# Comparing nested models

1) Add age, weight and height as additional predictors to the regression model relating vitamin C levels jointly to sex, smoking status and use of vitamin supplements.

2) Can you spot something that has changed about the data used by Stata to fit this model, compared to the previous models?

3) Use the ANOVA tables from the two models to conduct a partial F test for the three additional variables.

4) Use Stata's test command to confirm your result. What do you conclude from the results of this test?

# Summarising results

1) On the basis of the analysis you have done, which model would you select to present findings on whether serum vitamin C differs between smokers and non-smokers?

2) What are your conclusions about the factors that influence vitamin C levels?

# Matrix version

For this part we use a file called haem.dta which contains observations of haemoglobin, PCV measurements and ages of 12 adult women. The variables in this file are:

hb Haemoglobin (gm/dl)

pcv Pack cell volume %

age Age (years)

As there are just 12 rows and three variables, you should browse (or list) the data to inspect it before proceeding to further analysis.

1) Fit it a linear regression model for haemoglobin on PCV and age. Save the fitted values.

We will now use matrix algebra to (hopefully) calculate the same numbers Stata has provided us.

The first step is to convert the dataset into two matrices, one for the outcome (Y) and one for the predictors (X). We will use the Stata command "mkmat".

mkmat hb, matrix(Y)

To make the matrix for the predictors, we will first need to make a dummy variable containing the value one for every observation. We will use the variable to represent the constant when we create our predictor matrix.

gen one = 1

mkmat one pcv age, matrix(X)

Note that we follow the convention of using captial letters for matrix names. We can check this has done what we expected by listing the contents of each matrix:

matrix list Y

matrix list X

We can now use these matrices to generate the β matrix and the predicted values.First, the βmatrix:

matrix B = inv(X' \* X) \* X' \* Y

matrix list B

As an intermeidiate step to get the fitted values, we create the 'hat' matrix:

matrix P = X \* inv(X' \* X) \* X'

matrix list P

And finally, the matrix of fitted values:

matrix Q = P \* Y

matrix list Q

Compare these values to the variable you created earlier to check that they are the same.

To further understand the matrix algebra we will look in more depth at the estimation of the βmatrix.

matrix B = inv(X' \* X) \* X' \* Y

Lets examine contents of the matrix X'Y.

matrix XY = X'\*Y

mat list XY

The contents of this matrix are the sum of the sum of cross products between each of the predictors and Y. You can this by calculating these statistics. An easy way to do this is to use the Stata command "egen" (this is a very useful command in many different situations).

egen sum1y = sum(one\*hb)

egen sumx1y = sum(pcv\*hb)

egen sumx2y = sum(age\*hb)

l sum\* in 1

Similarly the X'Y matrix contains the sum of the squares of each predictor variable (on the diagonal) and the sum of the cross products between the predictors as the off diagonal terms. Go ahead and double check this yourself.