# Statistics 151 (Linear Modelling -Theory and Applications)
# Homework Two

**Due on 30 March, 2015**

16 March, 2015

1. Obtain an expression for the Cook's distance, $C_i$ in terms of the $i$th standardized residual $r_i$ and the leverage $h_{ii}$.

2. Show that the $i$th standardized predicted residual, $t_i$, satisfies $t_i = r_i\sqrt{(n-p-1)/(n-p-r_i^2)}$ where $r_i$ is the standardized residual.

3. In the Bodyfat dataset, consider the linear model

$$\text{BODYFAT} = \beta_0 + \beta_1\text{AGE} + \beta_2\text{WEIGHT} + \beta_3\text{HEIGHT} + \beta_4\text{THIGH} + e$$

   In R, plot the following graphs:

   (a) Residuals against fitted values.
   (b) Standardized Residuals against fitted values.
   (c) Residuals against Standardized Residuals.
   (d) Predicted residuals against fitted values.
   (e) Residuals against predicted residuals.
   (f) Residuals against leverage.
   (g) Predicted residuals against Standardized Predicted Residuals.
   (h) Standardized residuals against Standardized Predicted residuals.
   (i) Cooks Distance against the ID number of the subjects.

   Comment on these plots. Based on these plots, assess whether there are any outliers in the dataset; are there any influential observations. For each subject, calculate the p-value for testing whether the $i$th subject is an outlier based on the standardized predicted residual. Plot these p-values against the ID number of the subjects. How may of these p-values are less than 0.05? Does it make sense to rule all such subjects as outliers?

   Based on the analysis, does it make sense to fit the linear model with any of the subjects removed? If not, why not? If so, which ones; and in this case, report the summary for the linear model with the subjects removed.