QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Exercise Sheet 6

1. Based on the liver.csv dataset we have seen in the Practical session,

   (a) Fit Model 3: $\log_{10} Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, where $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and obtain the fitted values

   (b) Assess the assumptions of normality and constant variance of the random errors.

   (c) Compare the model 3 with the two models find in the Practical session. Which one is best and why?

   (d) For the best model find the leverage values and Cook's distances and comment.

2. Consider a set of equity returns from four different markets across 12 different periods. The data are available in the marketdata.txt. Define the forth variable as your response variable (Y). Define the following three models:

   - Model 1 $\Longrightarrow Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$;
   - Model 2 $\Longrightarrow Y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$
   - Model 3 $\Longrightarrow Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

   (a) What could you say about the significance of the parameters in the three models?

   (b) Are the data satisfying the normality assumption in each model?

   (c) The adjusted R^2 tells us which is the best model, when we have different number of exogenous variables across the models. As for the R^2, the model with the highest adjusted R^2 is the best. By using the adjusted R^2, state which is the best model across Model 1, Model 2 and Model 3.

   (d) For the best model, compute the predicted values when the new data are the X_1 vector of values.

   (e) Show the predicted values and the prediction interval (at 95% and at 90%) in a plot.

   (f) Are the predicted values far away from the real values of Y_i for the 90% and 95% case?