Regression Methods: Problems MATH-408
Anthony Davison

Problem 1 (Boston housing data)
In this practical, we will apply the concepts we have learned in class to the `Boston housing data`. The dataset could be loaded into `R` by first installing the `mlbench` package and then invoking `data(BostonHousing)`, or directly downloaded from here.

(a) First plot the response variable `medv`. Why do you think there are relatively many values of 50 and none above? Does this have implications for how the response should be modelled?

(b) Use the `lm` function in `R` to fit a linear model with `medv` as the response and all other variables as covariates. Use the `summary` function to get a summary of the fit. Do you think this fit is good? (Can you tell this from this output?) Which variables are significant? Give a careful interpretation of the regression coefficients, and give 95% confidence intervals for the effects of the most significant explanatory variables.

(c) Compute the residuals and plot them against the fitted values. Does the variance of the response depend on its mean? Do the standardized residuals seem normally distributed? Which points are the most influential? Are there any outliers?

It may be convenient to use the functions `glm.diag` and `glm.diag.plots` from the `SMPracticals` package, but if you do, you will need to fit the linear model using the `glm` function.

(d) Plot the standardized residuals against each explanatory variable. Do you see anything untoward? If so, suggest how to fix it, and discuss whether your fix works.

(e) Use the `boxcox` function from the `MASS` package to see if a response transformation might be helpful. What do you think?

(f) Using the `step` function, build models by selecting variables using forward selection, backward elimination, and stepwise regression. Do the models differ? Which do you think is best? Based on your discussion of significant variables in (b), are the variables selected by each method reasonable? What is the final AIC and BIC of the three models?

(g) Perform a 70/30 random split on the data and have the former as the training set and the latter as the testing. Fit a linear model to the training set, make predictions on the testing set and provide prediction confidence intervals. What is the average prediction error? Try this with different splits, and see how stable your conclusions are.