STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 4

1. The file `house.csv` contains data on 27 houses sold in Pennsylvania in 1977. We have the option of treating the number of bedrooms as numerical or categorical. For this analysis, we will fit an ANCOVA model to predict house prices, treating the number of bedrooms as categorical (which may be more appropriate - why?)

   (a) Fit a parallel regression ANCOVA model using price as the response variable, the number of bedrooms as the categorical predictor and the continuous variables Taxes, LotSize, LivingSpace and Age. Test whether the number of bedrooms is related to the house price. Create the appropriate plots to check diagnostics.

   (b) What is the expected difference in price between a three and a four-bedroom house? Find a 95% confidence interval for this value. Repeat this exercise for the difference in price between a four and a five bedroom house. Comment on the results.

   (c) Calculate the correlation matrix of the predictors (including the indicator variables associated with the number of bedrooms) and comment on its structure, particularly in relation to your results in part (b).

   (d) A stepwise variable selection procedure identifies the model including the predictors Taxes, LivingSpace and Bedrooms as the best model. Perform a cross-validation to assess the predictive capability of the model by using the last 7 data points as the validation set and the first 20 data points as the modelling set. Compare these predictions to the actual values using the formula:

$$\sum_{i=21}^{27} (Y_i - \hat{Y}_i)^2,$$

   where $\hat{Y}_i$ is the predicted value from the regression fit on the modelling set. Do the same for the model fit in part (a). Which model seems best from this perspective?

2. As an example of an Analysis of Covariance (ANCOVA) model, the lecture notes presents an analysis of some data on Teacher Effectiveness (Example 2 on pages 13 to 18). The data for this example are available on Wattle. Use R to repeat the analyses described in the lecture notes (the original analysis used S-Plus rather than R). Can you still fit the models described in the lecture notes using R and still get the same output shown in the lecture notes?