STAT 847
Reading Assignment 2

Chris Binoi Verghese
21092999

1. What is George Box's maxim about models?

According to George Box's maxim, "All models are wrong, but some are useful."

2. Give an example of the principle of parsimony in action in statistical modelling.

An example of the principle of parsimony in statistical modelling involves being willing to extend a linear regression model to include an additional quadratic term if this significantly improves prediction quality. However, if the quality does not improve, the principle of parsimony requires you remove it, since only the parameter of importance need to be included in the model.

3. In one sentence, why would you employ model averaging?

One would employ model averaging when there are several candidate models that have model scores similar to the winner model, in which case there may be better to combine inference output across these best models.

4. (1 mark) What are the three data sets used in this model selection problem?

The datasets were extracted from three 50,000-word corpora (in the original Russian editions):

(i) Sh, which is a published work guaranteed to be by Sholokhov,
(ii) Kr, which with equal trustworthiness is by Kriukov, and
(iii) QD, which is the Nobel winning text 'The Quiet Don'.

5. (2 marks) What are the three competing models that were suggested?

The three competing model frameworks were as follows:

i) M1: This proves Sholokhov is the author of 'The Quiet Don'. This can be done by showing text corpora Sh and QD come from the same statistical distribution, while Kr represents another.
ii) M2: D and Solzhenitsyn were correct in denouncing Sholokhov as the author. As such, the text corpus Sh is not statistically compatible with QD, while QD and Kr have the same distribution
iii) Sh, Kr, and QD possess three statistically disparate corpus distributions.

6. (3 marks) How were the models compared to each other? (A vague description in your own words is good. Details about the methodology aren't necessary, but do mention what data specifically was important)

The data involving the corpus and their sentence length distributions are first fitted into a four parameter mixture of a Poisson (a degenerate negative binomial) and another negative binomial. This model gives a prediction of number of sentences in various length groups and the three models with these distributions are compared and selected via statistical methodology involving closeness of distributions.
During the model choosing stages, the Poisson model was considered too simple while the mixed Poisson model had patterns that were more variegated than those dictated by a negative binomial resulting in a final four parameter model mixture Poisson and negative Binomial. The data used involve the corpus information from QD, Sh and Kr, as well as these corpus sentence lengths.

7. Describe the dataset being used in this problem.

The dataset utilized comprises football (soccer) match results from five prominent events:

i) the 1998 World Cup held in France,
ii) the 2000 European Cup in Belgium and the Netherlands,
iii) the 2002 World Cup in Korea and Japan,
iv) the 2004 European Cup in Portugal, and
v) the 2006 World Cup held in Germany.

The World Cups consist of 64 matches involving 32 national teams, while the European Cups consist of 31 matches among 16 teams. The numeric data values include the match number, the number of goals scored by each team (excluding extra time or penalty shoot-outs), along with the teams' FIFA rankings and their ratio fifa1/fifa2, followed by team names.

8. Figure 1.5 shows the distribution of football scores, which are whole numbers. Why are there clouds of data points instead of values only at the whole numbers?

The data values were jittered to enhance visibility of individual match results, preventing overlapping values and providing a clearer representation of scores. This approach facilitates the easy portrayal of the number of scores per value.