**SO5012 Analysing Data in the Real World**

**Seminar X – Ordinal Regression**

**Introduction and Data**

﻿This week’s seminar focuses on cricket, more specifically, the greatest form of cricket – test match cricket, ﻿a form of the game which involves matches lasting up to five days, played between national teams. A test series is a set of 1 to 6 consecutive test matches played in one country between the same two teams. Each match, and thus also each series, ends in a win for one of the two teams, or a draw when neither team wins. Each series is played in the country of one of the two teams, so we can talk about the home team and the visiting team of the series.

﻿The data set shows information on every test series of at least 2 matches played between 9 of the test-playing nations (Australia, England, South Africa, the West Indies, New Zealand, India, Pakistan, Sri Lanka, and Zimbabwe) in the period between 13 November 1959 and 22 August 2011. The variables are

* series: ID number of the series
* year: the year when the series was played (with the English and Southern seasons combined: e.g. 1970 refers to matches played approximately between September 1969 and September 1970)
* home and visitor: the home team and the visiting team
* matches: number of matches in the series
* winner of the series ("Draw" if the series was drawn)
* hrating and vrating: the ratings of the home and visiting teams before the series began, based on the results of each team in the preceding 3-4 years. Higher ratings correspond to more successful teams. These ratings will be treated as continuous, interval-level variables.
* drating: the difference hrating-vrating
* result of the series (1=Win for the visiting team; 2=Draw; 3=Win for the home team). This will be treated as the response variable Y in the analyses.
* period in which the series was played, approximately a decade. Dummy variables for different periods are also included

**Seminar tasks**

1. ﻿As always, spend some time playing with the data to understand how it works. In particular, answer these following questions (HINT: you may need to do some data manipulation):
   1. How many series were played in each year, in total?
   2. List each country by their number of series wins
   3. How many series have been played in each country, and how many series in total has each country played?
   4. How many wins does each country have when they were a visitor? How many draws? And loses?
   5. What is average difference between the home team’s rating and the away team’s rating? In which series was this largest?
2. These question appear similar, but are harder and require some more detailed data manipulation.
   1. Which country has the highest win to loss ratio?
   2. Which country has the largest difference between the percentage of wins at home compared to away?
   3. Are there any occurrences when the home team had a higher rating but failed to win the series? List the teams involved by year. What about when the home team with a higher rating lost?

Overall, these questions should demonstrate that there are multiple ways to interrogate the dataset and, for anyone with an interest in cricket at least, there are numerous interesting findings within it. The issue is that one can quickly become lost just pulling our various statistics, which don’t answer a fundamental and general questions such as ﻿(i) how well do the team ratings predict results of test series, and (ii) what is the extent of home advantage? Ordinal regression can be used here.

1. ﻿Fit an ordinal regression model for result, with drating as the only explanatory variable. Confirm that the effect of drating is statically significant.
2. ﻿Plot the fitted values of the probabilities of the individual categories of Y (i.e. the three values of result options) against drating
3. ﻿What does this show you in terms of who is likely to win? What can you say about the probability of a draw? And is there a home advantage?
4. It has been argued that the rating system has become less relevant since the mid 90s, and particularly after 2002. Similarly, the effect of rating is thought to be less pronounced when there are few matches in series. Can you test these hypotheses?