SO5012 Analysising Data in the Real World
Ordinal regression
Solutions and commentary

Contents

Introduction

This seminar sheet is intended as a introduction to multinomial regression and is a combination of code and **interpretation** for the worksheet *SO5012_semX_ordinal_regression.docx*.

As with all the previous weeks, first we need to: 1. Set the working directory 2. Load the packages we'll be using 3. Load the data

Here we'll use the `results='hide', message=FALSE` command on the r chunk so that our output is not filled up by this set up code, although the code will be visible.

```
# setwd(whereeveryousavestuff)
# Note to RP - this isnt needed for a project, but you'll need to change if just posting rmd

if (!require(nnet)) install.packages("nnet")
library(nnet)

if (!require(lmtest)) install.packages("lmtest")
library(lmtest)

cricket <- read.csv("data/cricket.csv")
```

As with each work sheet, and all analysis, we need to do basic checks on the data before starting any analysis proper.

```
str(cricket)
```

```
## 'data.frame':    389 obs. of  17 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ series   : int  123 124 125 127 128 129 130 131 133 134 ...
##  $ year     : int  1960 1960 1960 1961 1961 1961 1962 1962 1962 1962 ...
##  $ home     : chr  "Pak" "Ind" "WI" "Ind" ...
##  $ visitor  : chr  "Aus" "Aus" "Eng" "Pak" ...
##  $ matches  : int  3 5 5 5 5 5 3 5 5 5 ...
##  $ winner   : chr  "Aus" "Aus" "Eng" "Draw" ...
##  $ hrating  : num  -2.72 -33.33 4.07 -26.33 47.54 ...
##  $ vrating  : num  50.09 53.95 17.55 5.83 6.71 ...
##  $ drating  : num  -52.8 -87.3 -13.5 -32.2 40.8 ...
```

```
##  $ result   : chr  "Visitor" "Visitor" "Visitor" "Draw" ...
##  $ period   : chr  "1960-69" "1960-69" "1960-69" "1960-69" ...
##  $ per_60_69: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ per_70_79: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ per_80_89: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ per_90_02: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ per_02on : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# we see that home, visitor, winner, result and period are character vectors.
# Lets convert them into factors.

factorvars <- c("home", "visitor", "winner", "result", "period")
for (v in factorvars) {
  cricket[[v]] <- as.factor(cricket[[v]])
  print(levels(cricket[[v]])) # this simply report the resulting levels
}
```

```
## [1] "Aus" "Eng" "Ind" "NZ"  "Pak" "SA"  "SL"  "WI"  "Zim"
## [1] "Aus" "Eng" "Ind" "NZ"  "Pak" "SA"  "SL"  "WI"  "Zim"
##  [1] "Aus"  "Draw" "Eng"  "Ind"  "NZ"   "Pak"  "SA"   "SL"   "WI"   "Zim"
## [1] "Draw"    "Home"    "Visitor"
## [1] "1960-69"     "1970-79"     "1980-89"     "1990-3.2002" "4.2002-"
```

```
# the result variable is currently in the wrong order, running "Draw", "Home", "Visitor"
# this needs re-levelling.

cricket$result <- factor(cricket$result, levels = c("Visitor", "Draw", "Home"))
levels(cricket$result)
```

```
## [1] "Visitor" "Draw"    "Home"
```

Only now are we ready to start the questions!

Questions and answers

A friend once said:

> It's always better to give than to receive.

Question 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): a. How many series were played in each year, in total? b. List each country by their number of series wins c. How many wins does each country have when they were a visitor? How many when draws? And loses? d. Which country has the highest win ratio? e. Which country has the largest difference between the percentage of wins at home compared to away? f. What is average difference between the home team's rating and the away team's rating? In which series was this largest? g. Are there any occurrences when the home team had a higher rating but failed to win the series? List them by year.