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(MASS)) install.packages("MASS")
library(MASS)

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

if (!require(reshape2)) install.packages("reshape2")
library(reshape2)


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 ...
```

```r
# 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-"
```

```r
# 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

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?

```
# Each row is a series, and there is a variable related to year - so simply tabulate it!

table(cricket$year)
```

```
##
## 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975
##    3    3    4    2    1    6    2    3    4    3    2    4    2    4    5    5
## 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991
##    4    6    6    6    7    6    6    5    7    7    8    7    6    6    6    5
## 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007
##    3    7    8   10    9   14   16   10   14   17   15   12   17   15   15    7
## 2008 2009 2010 2011
##   14   13   11   11
```

b. List each country by their number of series wins

```
# Again, each row is a series and the winner is identified

table(cricket$winner)
```

```
##
##  Aus Draw  Eng  Ind   NZ  Pak   SA   SL   WI  Zim
##   67   81   55   40   19   34   30   22   40    1
```

c. How many series have been played in each country, and how many series in total has each country played?

```
# The first is easy

table(cricket$home)
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##  59  69  52  47  44  30  31  43  14
```

```
# The second requires us to sum the number of occurrences of each country across two columns

table(cricket$home) + table(cricket$visitor)
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
## 116 127 106  95  95  56  59  96  28
```

```
# This only works as the two tables have the same countries, in the same order.
# If they don't - its more complicated...
# see https://stackoverflow.com/questions/26986363/adding-two-vectors-by-names

# first save each table:
v1 <- table(cricket$home)
v2 <- table(cricket$visitor)
# make them into one long table, with repeated countries
```

```
v3 <- c(v1, v2)
v3
```

```
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##  59  69  52  47  44  30  31  43  14  57  58  54  48  51  26  28  53  14
```

```
# then a little trick to compress it
# this means 'create a crosstab of v3, by the names of v3'
total_series <- xtabs(v3 ~ names(v3))
total_series
```

```
## names(v3)
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
## 116 127 106  95  95  56  59  96  28
```

d. How many wins does each country have when they were a visitor? How many draws? And loses?

```
# wins
table(cricket$visitor[cricket$result == "Visitor"])
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##  26  19  11   6  14  11   5  17   1
```

```
# draw
table(cricket$visitor[cricket$result == "Draw"])
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##  11  11  12   7  16   6   6  11   1
```

```
# loses
table(cricket$visitor[cricket$result == "Home"])
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##  20  28  31  35  21   9  17  25  12
```

e. What is average difference between the home team's rating and the away team's rating? In which series was this largest?

```
summary(cricket$drating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -87.288 -16.770  -2.757  -2.629  13.110  62.667
```

```
cricket[cricket$drating == (summary(cricket$drating))[6],]
```

```
##       X series year home visitor matches winner  hrating   vrating  drating
## 288 288    494 2004  Aus     Zim       2    Aus 40.29146 -22.37519 62.66665
##     result  period per_60_69 per_70_79 per_80_89 per_90_02 per_02on
## 288   Home 4.2002-         0         0         0         0        1
```

```
# This looks confusing, but (summary(cricket$drating))[6] is simply returning
# the value of the summary that is the 6th entry in the summary table
# i.e. the maximum, and only returning the row of dataset where drating equals it.
```

Question 2

These question appear similar, but are harder and require some more detailed data manipulation.

a. Which country has the highest win ratio?

```
# This isn't easy...
# first we need the total games of each country, from question 1c
total_series
```

```
## names(v3)
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
## 116 127 106  95  95  56  59  96  28
```

```
# and then the total wins - i.e.
win_series <- table(cricket$winner)
win_series
```

```
##
##  Aus Draw  Eng  Ind   NZ  Pak   SA   SL   WI  Zim
##   67   81   55   40   19   34   30   22   40    1
```

```
# then we need to divide one by the other, but  the tables are different
# because of draws (look at both),

# so we need to create a little data frame of the two, using cbind (column bind)

wins_total <- data.frame(cbind(wins = win_series,
                               total = total_series[names(win_series)]))

# and now we can divide one column by the other
wins_total$win_ratio <- wins_total$wins/wins_total$total

# and then display it, ordering by win ratio descending
wins_total[order(-wins_total$win_ratio),]
```

```
##      wins total  win_ratio
## Aus    67   116 0.57758621
## SA     30    56 0.53571429
## Eng    55   127 0.43307087
## WI     40    96 0.41666667
## Ind    40   106 0.37735849
## SL     22    59 0.37288136
## Pak    34    95 0.35789474
## NZ     19    95 0.20000000
## Zim     1    28 0.03571429
## Draw   81    NA         NA
```

b. Which country has the largest difference between the percentage of wins at home compared to away?

```
# No of wins at home per country
win_home <- table(cricket$winner[cricket$result == "Home"])

# Number of home series
home_series <- table(cricket$home)

# and using the method from 2a
```

```
home_total <- data.frame(cbind(win_home = win_home,
                               no_home = home_series[names(win_home)]))

home_total$home_ratio <- home_total$win_home / home_total$no_home

# and then the same for away series wins
win_away <- table(cricket$winner[cricket$result == "Visitor"])
visitor_series <- table(cricket$visitor)
visitor_total <- data.frame(cbind(win_away = win_away,
                                  no_visitor = visitor_series[names(win_away)]))
visitor_total$visit_ratio <- visitor_total$win_away/visitor_total$no_visitor

ratios_home_visit <- cbind(home_total, visitor_total)

ratios_home_visit$ratio_diff <- ratios_home_visit$home_ratio - ratios_home_visit$visit_ratio

ratios_home_visit[order(-ratios_home_visit$ratio_diff),]
```

```
##      win_home no_home home_ratio win_away no_visitor visit_ratio  ratio_diff
## SL         17      31  0.5483871        5         28  0.17857143  0.36981567
## Ind        29      52  0.5576923       11         54  0.20370370  0.35398860
## Aus        41      59  0.6949153       26         57  0.45614035  0.23877490
## WI         23      43  0.5348837       17         53  0.32075472  0.21412900
## SA         19      30  0.6333333       11         26  0.42307692  0.21025641
## Eng        36      69  0.5217391       19         58  0.32758621  0.19415292
## Pak        20      44  0.4545455       14         51  0.27450980  0.18003565
## NZ         13      47  0.2765957        6         48  0.12500000  0.15159574
## Zim         0      14  0.0000000        1         14  0.07142857 -0.07142857
## Draw        0      NA         NA        0         NA          NA          NA
```

c. 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?

```
loss_high_rating <- cricket[cricket$drating > 0 & cricket$result != "Home",]
loss_high_rating[,3:4]
```

```
##     year home
## 11  1963  Aus
## 13  1964  Eng
## 15  1965   SA
## 19  1965  Eng
## 26  1968   WI
## 27  1968   NZ
## 28  1968  Eng
## 35  1971   WI
## 37  1971  Eng
## 39  1972  Eng
## 43  1973  Eng
## 48  1974  Eng
## 49  1975  Ind
## 82  1980  Eng
## 107 1985  Pak
## 110 1985  Ind
## 116 1986  Aus
## 120 1986  Eng
```

```
## 124 1987  Aus
## 132 1988   NZ
## 133 1988   WI
## 141 1990  Pak
## 156 1993  Aus
## 178 1995   WI
## 180 1996  Pak
## 185 1996   NZ
## 192 1997  Pak
## 208 1998  Ind
## 212 1998   SA
## 222 1999  Pak
## 228 1999  Eng
## 243 2001   SL
## 255 2001  Eng
## 262 2002  Aus
## 270 2002   NZ
## 274 2002  Eng
## 292 2004  Aus
## 294 2004   NZ
## 298 2004  Pak
## 305 2005  Pak
## 311 2005  Ind
## 330 2006  Eng
## 340 2007  Eng
## 341 2008  Pak
## 359 2009  Aus
## 371 2010   SA
## 382 2011   SL
```

```
# and without the intermediate step for visitor wins
cricket[cricket$drating > 0 & cricket$result == "Visitor",3:4]
```

```
##     year home
## 13  1964  Eng
## 15  1965   SA
## 19  1965  Eng
## 26  1968   WI
## 27  1968   NZ
## 35  1971   WI
## 37  1971  Eng
## 43  1973  Eng
## 49  1975  Ind
## 82  1980  Eng
## 110 1985  Ind
## 120 1986  Eng
## 124 1987  Aus
## 156 1993  Aus
## 178 1995   WI
## 180 1996  Pak
## 222 1999  Pak
## 228 1999  Eng
## 243 2001   SL
## 294 2004   NZ
## 298 2004  Pak
```

```
## 340 2007  Eng
## 341 2008  Pak
## 359 2009  Aus
```

```
# and just to see who did most often:

table(cricket[cricket$drating > 0 & cricket$result == "Visitor",]$home)
```

```
##
## Aus Eng Ind  NZ Pak  SA  SL  WI Zim
##   3   8   2   2   4   1   1   3   0
```

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.

Question 3.

Fit an ordinal regression model for result, with drating as the only explanatory variable. Confirm that the effect of drating is statically significant.*

```
m1 <- polr(result ~ drating , data = cricket)
summary(m1)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = result ~ drating, data = cricket)
##
## Coefficients:
##           Value Std. Error t value
## drating 0.04416   0.005327    8.29
##
## Intercepts:
##              Value   Std. Error t value
## Visitor|Draw -1.2369  0.1295    -9.5480
## Draw|Home    -0.1659  0.1127    -1.4721
##
## Residual Deviance: 715.1328
## AIC: 721.1328
```

```
# as with multinomial regression this table does not include p values which need
# to be calculated separately - see https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/

ctable1 <- coef(summary(m1))
```

```
##
## Re-fitting to get Hessian
```

```
p <- pnorm(abs(ctable1[, "t value"]), lower.tail = FALSE) * 2
ctable <- cbind(ctable1, "p value" = p)
round(ctable1,4)
```

```
##                Value Std. Error t value
## drating       0.0442     0.0053  8.2896
```

```
## Visitor|Draw -1.2369     0.1295 -9.5480
## Draw|Home    -0.1659     0.1127 -1.4721
```

```
# and of course we need to exponential of the coeff to interpret
exp(coef(m1))
```

```
##  drating
## 1.045145
```

The P-value of the coefficient of drating is $P < 0.001$, so the coefficient is statistically significant. The estimated coefficient is 0.044, and its exponential is $\exp(0.044) = 1.045$. The interpretation of this is that a 1-point increase in the difference between home and visiting teams' ratings is associated with 4.5% increase in the odds of a more favourable outcome for the home team. This same increase applies to both the odds of home win, against draw or visitor winning, and the odds of home win or draw, against visitor winning.

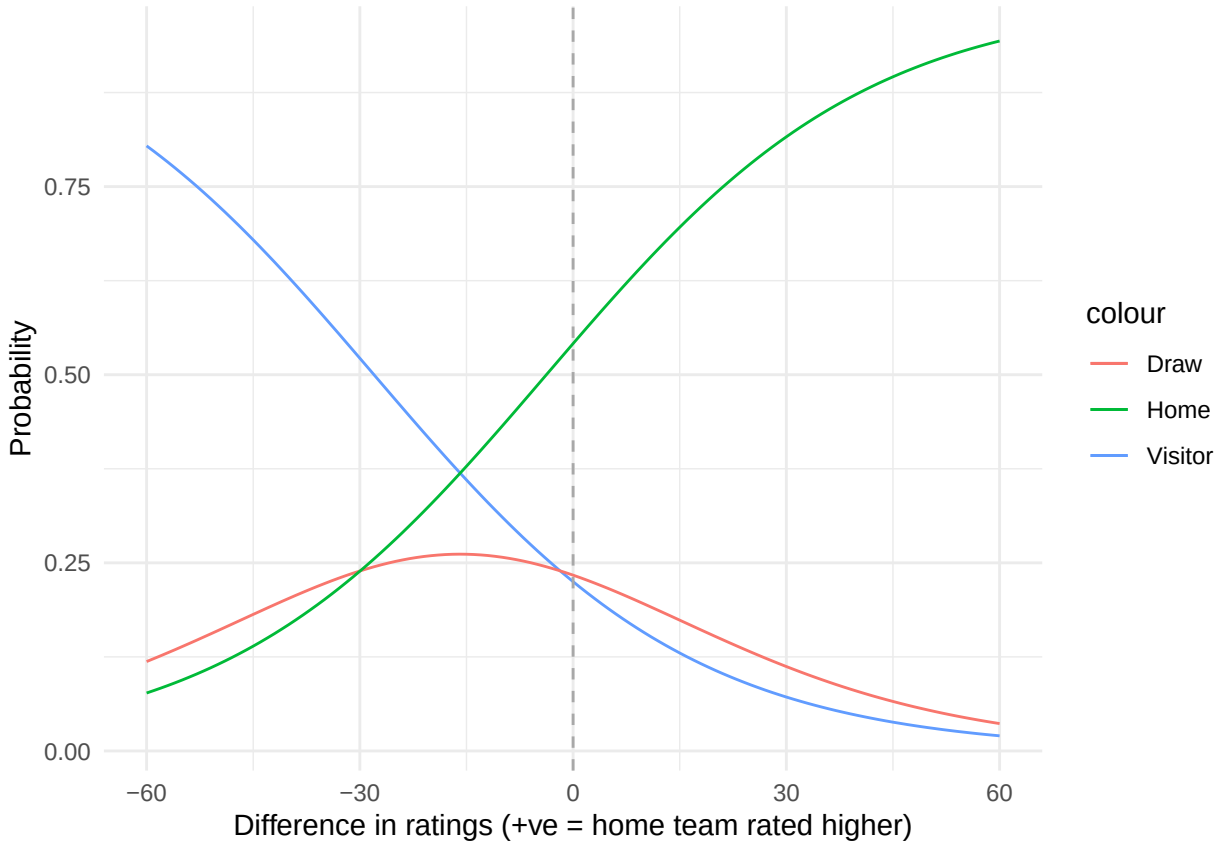
Question 4.

Plot the fitted values of the probabilities of the individual categories of Y (i.e. the three values of result options) against drating

```
# create a little fake dataset to run the model against - lets go with drating
# from - 60 to 60, which is a good part of the rnage seen in question 1 e

test_data <- data.frame(drating = seq(-60,60,1))
prediction <- data.frame( cbind(drating = test_data$drating,
                          predict(m1, test_data, type = "probs") ))


ggplot(data = prediction,
       aes(drating)) +
        geom_line(aes(y = Visitor, colour = "Visitor")) +
        geom_line(aes(y = Draw, colour = "Draw")) +
        geom_line(aes(y = Home, colour = "Home")) +
        geom_vline(xintercept = 0, linetype = "dashed", color = "grey66") +
  theme_minimal() +
  ylab("Probability") +
  xlab("Difference in ratings (+ve = home team rated higher)")
```



Question 5

What does this show you in terms of who is likely to win? What can tou say about the probability of a draw? And is there a home advantage?

The plot of the probabilities of the individual outcomes shows that the probability of the home team winning is very high when drating is large and positive, i.e. when the home team is much stronger than the visiting team. Similarly, the probability of the visitors winning is high when drating is large and negative. The ratings of the teams before a series are thus strongly predictive of the result of the series. A draw is never the likeliest outcome, and has probabilities of roughly 0.2 for most of the values of drating.

As to whether there is a home advantage: the probability curves for wins by home and visiting teams are quite clearly not mirror images of each other - at each value of the ratings difference between the same two teams, a team has a much higher predicted probability of winning if it plays at home than if it plays away. In particular, in a series of two equally strong teams the probability that the home team wins is around 0.55, and the probability that the visiting team wins is about 0.22. This shows evidence of a substantial home advantage in test cricket.

Question 6.

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? *

```
m2 <- polr(result ~ drating + matches + period, data = cricket)
summary(m2)
```

```
##
```

```
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = result ~ drating + matches + period, data = cricket)
##
## Coefficients:
##                       Value Std. Error t value
## drating             0.04479   0.005397  8.3001
## matches            -0.04791   0.097320 -0.4922
## period1970-79       0.22963   0.474503  0.4839
## period1980-89       0.74614   0.448677  1.6630
## period1990-3.2002   0.44616   0.422565  1.0558
## period4.2002-       0.43070   0.433761  0.9930
##
## Intercepts:
##               Value   Std. Error t value
## Visitor|Draw -0.9710  0.5551    -1.7493
## Draw|Home     0.1098  0.5524     0.1987
##
## Residual Deviance: 711.2552
## AIC: 727.2552
```

```
# as with multinomial regression this table does not include p values which need
# to be calculated separately - see https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/

ctable2 <- coef(summary(m2))
```

```
##
## Re-fitting to get Hessian
```

```
p2 <- pnorm(abs(ctable2[, "t value"]), lower.tail = FALSE) * 2
ctable2 <- cbind(ctable2, "p value" = p2)
round(ctable2,4)
```

```
##                     Value Std. Error t value p value
## drating            0.0448     0.0054  8.3001  0.0000
## matches           -0.0479     0.0973 -0.4922  0.6225
## period1970-79      0.2296     0.4745  0.4839  0.6284
## period1980-89      0.7461     0.4487  1.6630  0.0963
## period1990-3.2002  0.4462     0.4226  1.0558  0.2910
## period4.2002-      0.4307     0.4338  0.9930  0.3207
## Visitor|Draw      -0.9710     0.5551 -1.7493  0.0802
## Draw|Home          0.1098     0.5524  0.1987  0.8425
```

```
# and of course we need to exponential of the coeff to interpret
exp(coef(m2))
```

```
##           drating           matches     period1970-79     period1980-89
##         1.0458115         0.9532239         1.2581299         2.1088453
## period1990-3.2002     period4.2002-
##         1.5623026         1.5383415
```

```
p2
```

```
##           drating           matches     period1970-79     period1980-89
##      1.040588e-16      6.225450e-01      6.284350e-01      9.631671e-02
## period1990-3.2002     period4.2002-      Visitor|Draw         Draw|Home
##      2.910413e-01      3.207325e-01      8.024443e-02      8.424790e-01
```

We can see from this output that neither the number of matches nor the time period in which the game was played are signficant. Similarly the value of drating coefficeient does not change much so we can say that in all likelihood, that the rating of teams is important regardless of the number of matches and is as relevent now as it was in the 1960s.

Note: we have used here period as a categorical variable here to see if there are non-liner changes over the time period. We could have added the raw year value if our hypothesis was that the outcomes had changed the probability of occuring over over time