SO5012 Analysising Data in the Real World
Multinomial 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_multinomial_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)

ukvote2010 <- read.csv("data/ukvote2010.csv")
```

To continue with the preparatory work, and **before** we dive into answering the questions, we first need to look at our data and see what it is, perhaps with some data cleaning if needed.

```
str(ukvote2010)
```

```
## 'data.frame':    16816 obs. of  8 variables:
##  $ VoteIntention : chr  "LiberalDemocrat" NA NA NA ...
##  $ Region        : chr  "west midlands" "greater london" "east midlands" "scotland" ...
##  $ Gender        : chr  "female" "female" "male" "female" ...
##  $ Ethnicity     : chr  "white" "white" "asian or asian british" "white" ...
##  $ MaritalStatus : chr  "Married/Widowed" "Seperated/Divorced" "Partner/Single" "Partner/Single" ...
```

```
##  $ Age           : int  62 70 30 42 61 79 78 22 26 62 ...
##  $ Housing       : chr  "Owns/Mortgage" "Owns/Outright" "Rent/Other" "Owns/Mortgage" ...
##  $ Qualifications: chr  "GCE" "Technical/None/Other" "Technical/None/Other" "Technical/None/Other" .
```

```
# Note - only do this if there are a limited number of variables, or you will end
# up with tonnes of output, often up to hundreds of pages!

names(ukvote2010)
```

```
## [1] "VoteIntention"  "Region"         "Gender"         "Ethnicity"
## [5] "MaritalStatus"  "Age"            "Housing"        "Qualifications"
```

```
# We see that "VoteIntention", "Region", "Gender", "Ethnicity", "MaritalStatus",
# "Housing", "Qualifications" are character variables - lets set them to factors

# and lets be smart and do it in a loop
factorvars <- c("VoteIntention", "Region", "Gender", "Ethnicity",
                "MaritalStatus", "Housing", "Qualifications")
for (v in factorvars) {
  ukvote2010[[v]] <- as.factor(ukvote2010[[v]])
}

summary(ukvote2010)
```

```
##          VoteIntention                     Region        Gender
##  Conservative   :3835   south east            :2915   female:8474
##  Labour         :2700   greater london        :2103   male  :8342
##  LiberalDemocrat:2349   north west            :1858
##  Other          :1119   scotland              :1609
##  NA's           :6813   south west            :1582
##                         yorkshire & humberside:1486
##                         (Other)               :5263
##                    Ethnicity               MaritalStatus       Age
##  asian or asian british :  280   Married/Widowed   :9994   Min.   : 18.00
##  black or black british :  161   Partner/Single    :5221   1st Qu.: 38.00
##  mixed background       :  174   Seperated/Divorced:1601   Median : 51.00
##  other ethnic background:  164                             Mean   : 49.89
##  white                  :16037                             3rd Qu.: 62.00
##                                                            Max.   :110.00
##
##          Housing                  Qualifications
##  Owns/Mortgage:7217   BA                  :5378
##  Owns/Outright:5670   GCE                 :4370
##  Rent/Other   :3929   GCSE                :1722
##                       Technical/None/Other:5346
##
##
##
```

```
# Now that we have set the factor variables correctly,
# this short cut will display most of the descriptives, but not the region
# as there are too many levels
table(ukvote2010$Region, useNA = "ifany")
```

```
##
##            east anglia          east midlands         greater london
```

```
##                 1192                 1186                 2103
##                north            north west             scotland
##                  732                 1858                 1609
##           south east            south west                wales
##                 2915                 1582                  831
##         west midlands yorkshire & humberside
##                 1322                 1486
```

```
# Lets pay special attention to the dependent, VotingIntention

table(ukvote2010$VoteIntention, useNA = "ifany")
```

```
##
##    Conservative          Labour LiberalDemocrat           Other            <NA>
##            3835            2700            2349            1119            6813
```

```
prop.table(table(ukvote2010$VoteIntention, useNA = "ifany"))
```

```
##
##    Conservative          Labour LiberalDemocrat           Other            <NA>
##      0.22805661      0.16056137      0.13968839      0.06654377      0.40514986
```

Here we can see, after setting the factor variables correctly, that none of the questions have any missing values, except for the voting intention variable, where 40.5% are missing values - presumably people who did not express an firm opinion (but to confirm this we'd need to look at the data documentation, which you do not have)

Questions and answers

Question 1

Cross-tabulate the variable VoteIntention with the variable Qualifications, setting the table to include the conditional probabilities of VoteIntention, given Qualifications. Which parties do better among those with higher qualification levels and which do worse?

```
table(ukvote2010$Qualifications,ukvote2010$VoteIntention)
```

```
##
##                         Conservative Labour LiberalDemocrat Other
##   BA                            1106    861             984   283
##   GCE                            991    645             573   308
##   GCSE                           425    302             205   112
##   Technical/None/Other          1313    892             587   416
```

```
# as this is two way table, we need to tell R which direction to calculate the
# percentages - this is done by the ,1 at the end of the prop.table command
prop.table(table(ukvote2010$Qualifications,ukvote2010$VoteIntention),1)
```

```
##
##                         Conservative     Labour LiberalDemocrat      Other
##   BA                     0.34199134 0.26623377      0.30426716 0.08750773
##   GCE                    0.39372269 0.25625745      0.22765197 0.12236790
##   GCSE                   0.40708812 0.28927203      0.19636015 0.10727969
##   Technical/None/Other   0.40928928 0.27805486      0.18298005 0.12967581
```

```
# and we can turn it into percentages and round...
round(100*prop.table(table(ukvote2010$Qualifications,ukvote2010$VoteIntention),1),1)
```

```
##
##                         Conservative Labour LiberalDemocrat Other
##   BA                            34.2   26.6            30.4   8.8
##   GCE                           39.4   25.6            22.8  12.2
##   GCSE                          40.7   28.9            19.6  10.7
##   Technical/None/Other          40.9   27.8            18.3  13.0
```

Tracing the vote shares (fractions) for each party, we see that Labour's share is steady at around 25-28%, regardless of education levels. The Conservative share declines with increasing education, from 41% at the lowest levels to 34% at the highest levels. The Liberal Democratic share increases with higher education, rising from 18 to 30%, while those supporting other parties declines from 13% to 9% as education levels increase.

It is worth noting two things: 1. that these statistics are based on 2010 data, before the coalition government that introduced austerity, and when the Liberal Democrats had their first experience of national government. In short, their vote share was *very* different to what is seen now. 2. a GCE - a General Certificate of Education Advanced Level - is commonly known as an A-level but this also includes equivalents (BTECs and the like)

Question 2

Fit a multinomial logistic regression model for VoteIntention, with Labour as the baseline outcome category, using only the variable Qualifications as an explanatory (factor) variable. How can we see from the coefficients which parties do better among those with higher education levels and which do worse? Check that you see the same general patterns as you saw when you cross-tabulated the same data.

```
levels(ukvote2010$VoteIntention)
```

```
## [1] "Conservative"    "Labour"          "LiberalDemocrat" "Other"
```

```
# so first we need to relevel Vote Intention
ukvote2010$VoteIntention <- relevel(ukvote2010$VoteIntention, ref = "Labour")

m1 <- multinom(VoteIntention~ Qualifications, data = ukvote2010)
```

```
## # weights:  20 (12 variable)
## initial  value 13867.102494
## iter  10 value 13013.059359
## final  value 12984.946892
## converged
```

```
summary(m1)
```

```
## Call:
## multinom(formula = VoteIntention ~ Qualifications, data = ukvote2010)
##
## Coefficients:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative      0.2504099         0.1790537         0.09125419
## LiberalDemocrat   0.1335318        -0.2518969        -0.52094697
## Other            -1.1126471         0.3734956         0.12071355
##                 QualificationsTechnical/None/Other
## Conservative                             0.1361937
## LiberalDemocrat                         -0.5519734
## Other                                    0.3498647
##
## Std. Errors:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative     0.04544887        0.06800786         0.08791927
## LiberalDemocrat  0.04666583        0.07398167         0.10181848
## Other            0.06852012        0.09742766         0.13013413
##                 QualificationsTechnical/None/Other
## Conservative                           0.06283542
## LiberalDemocrat                        0.07072732
## Other                                  0.09066384
##
## Residual Deviance: 25969.89
## AIC: 25993.89
```

This stores and displays the coefficients and the *standard errors* (a little like standard deviation, but how much variation there is in the estimate of a coefficient within a model). This needs to be used to calculate the p-values, as the `multinom` package does not do it automatically! The only way to work this out is to google it...

```
zvals <- coef(m1)/summary(m1)$standard.errors
pvals <- (1 - pnorm(abs(zvals), 0, 1)) * 2
```

```
# and then display the exponentiated coefficient along with p values afterwards
round(exp(coef(m1)),5)
```

```
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative        1.28455           1.19608            1.09555
## LiberalDemocrat     1.14286           0.77732            0.59396
## Other               0.32869           1.45280            1.12830
##                 QualificationsTechnical/None/Other
## Conservative                               1.14590
## LiberalDemocrat                            0.57581
## Other                                      1.41888
```

```
round(pvals,4)
```

```
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative         0.0000            0.0085             0.2993
## LiberalDemocrat      0.0042            0.0007             0.0000
## Other                0.0000            0.0001             0.3536
##                 QualificationsTechnical/None/Other
## Conservative                                0.0302
## LiberalDemocrat                             0.0000
## Other                                       0.0001
```

Since Labour is the baseline outcome level, all coefficients correspond to a comparison of support for some other party to support for Labour. The raw coefficients represent the log odds of a Labour voting intention vs Conservative. The exponentiated form is the ratio of the probabilities of choosing one outcome category over the baseline category (in this case, Labour).

Looking at the coefficients for Conservative, we see that lower education levels are associated with greater support for the Conservatives, relative to Labour (positive coefficients for education levels below BA which translates to values above 1 when they are exponentiated), although the GCSE coefficient is not significant. There is not much different between the various education levels that are less than a BA, which is consistent with the cross-tabulation results in question 1. What this means, is that the 'gap' (i.e. the ratio increases) between Labour and Conservative voting intention is larger at the lower education levels than it is at the BA level.

Looking at the coefficients for Liberal Democrat, we see that lower education levels are associated with lower levels of support for the Liberal Democrats, relative to Labour (increasingly negative raw coefficients/values below 1 when exponentiated as education levels go down). In this case, this means that at the lower education levels, the gap between Labour and Lib Dem votes increases, in favour of Labour.

Looking at the coefficients for Other, we see that education levels below BA are associated with higher levels of support for other parties, but there is not a clear trend among the lower education levels - i.e. that the gap between the parties is generally smaller at lower education levels, but there is not clear difference across the education levels.

All of these patterns are consistent with what we see in the cross-tabulation id you look at the 'gaps' between each parties vote intention share at each education level and compare them to the gap at the highest education level.

Question 3

As a sense check, run a logistic regression for a voting intention of Labour (as the baseline) vs conservatives. What do you notice about the results compared to those found in question 2?

```
# Here we need to limit the data to only include voting intentions that are
# Labour or Tory.

m_logit <- glm(VoteIntention ~ Qualifications,
          data = ukvote2010[ukvote2010$VoteIntention == "Labour"| ukvote2010$VoteIntention =="Conservati
          family = "binomial")
summary(m_logit)
```

```
##
## Call:
## glm(formula = VoteIntention ~ Qualifications, family = "binomial",
##     data = ukvote2010[ukvote2010$VoteIntention == "Labour" |
##         ukvote2010$VoteIntention == "Conservative", ])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.364  -1.326   1.001   1.018   1.073
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       0.25041    0.04545   5.510 3.59e-08 ***
## QualificationsGCE                 0.17905    0.06801   2.633  0.00847 **
## QualificationsGCSE                0.09125    0.08792   1.038  0.29932
## QualificationsTechnical/None/Other  0.13619    0.06284   2.167  0.03020 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8861.3  on 6534  degrees of freedom
## Residual deviance: 8853.3  on 6531  degrees of freedom
##   (6813 observations deleted due to missingness)
## AIC: 8861.3
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(m_logit))
```

```
##                       (Intercept)                 QualificationsGCE
##                          1.284553                          1.196085
##                QualificationsGCSE QualificationsTechnical/None/Other
##                          1.095544                          1.145903
```

We can see that the results here are the same the first line of the multinomial regression, and their interpretation should be familiar from previous seminars. This shows that multinomial regression is essentially just combining a series of logistic regressions which compare all the various levels in the dependent to your chosen reference into one handy method command.

Question 4

Add the variable Age to the model. For which outcome levels is there a significant association between age and vote intention, controlling for qualifications?

```
m2 <- multinom(VoteIntention~ Qualifications + Age, data = ukvote2010)
```

```
## # weights:  24 (15 variable)
## initial  value 13867.102494
## iter  10 value 13057.698986
## iter  20 value 12895.624225
## final  value 12895.623583
## converged
```

```
summary(m2)
```

```
## Call:
## multinom(formula = VoteIntention ~ Qualifications + Age, data = ukvote2010)
##
## Coefficients:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative     -0.5792916         0.1251710      -0.0065498275
## LiberalDemocrat   0.1948017        -0.2476404      -0.5124993357
## Other            -2.1749271         0.3060860       0.0008302082
##                 QualificationsTechnical/None/Other          Age
## Conservative                            -0.01414139  0.017526257
## LiberalDemocrat                         -0.53981612 -0.001352848
## Other                                    0.16225110  0.022198394
##
## Std. Errors:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative     0.09738985        0.06857176         0.08888796
## LiberalDemocrat  0.10247865        0.07425199         0.10259346
## Other            0.14594700        0.09809118         0.13119248
##                 QualificationsTechnical/None/Other         Age
## Conservative                            0.06506614 0.001826380
## LiberalDemocrat                         0.07299738 0.002014430
## Other                                   0.09345030 0.002626301
##
## Residual Deviance: 25791.25
## AIC: 25821.25
```

```
zvals <- coef(m2)/summary(m2)$standard.errors
pvals <- (1 - pnorm(abs(zvals), 0, 1)) * 2

exp(coef(m2))
```

```
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative      0.5602952         1.1333422          0.9934716
## LiberalDemocrat   1.2150700         0.7806406          0.5989966
## Other             0.1136164         1.3580992          1.0008306
##                 QualificationsTechnical/None/Other       Age
## Conservative                             0.9859581 1.0176807
## LiberalDemocrat                          0.5828554 0.9986481
## Other                                    1.1761555 1.0224466
```

```
pvals
```

```
##                  (Intercept) QualificationsGCE QualificationsGCSE
## Conservative    2.711534e-09      0.0679404379       9.412600e-01
## LiberalDemocrat 5.731508e-02      0.0008525802       5.870216e-07
## Other           0.000000e+00      0.0018059105       9.949509e-01
##                 QualificationsTechnical/None/Other       Age
## Conservative                         8.279444e-01 0.0000000
## LiberalDemocrat                      1.414424e-13 0.5018519
## Other                                8.252338e-02 0.0000000
```

The coefficients on Age are significant in the Conservative and Other equations, but not in the Liberal Democrat equation. That is, there is strong evidence of a positive association between age and support for Conservative relative to Labour, and strong evidence of a positive association between age and support for Other relative to Labour, holding qualifications constant. However, there does not appear to be a change associated with ageing for support of the Lib Dems over the Labour party.

Question 5

Perform a likelihood ratio test to see whether Age is a significant predictor across all outcome levels.

```
lrtest(m1, m2)
```

```
## Likelihood ratio test
##
## Model 1: VoteIntention ~ Qualifications
## Model 2: VoteIntention ~ Qualifications + Age
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  12 -12985
## 2  15 -12896  3 178.65  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio statistic is 2((-12895.624)-(-12984.947))=178.65. Since we added three coefficients (age in each of three equations), the likelihood ratio test has three degrees of freedom. The likelihood ratio test is strongly significant ($p < 0.0001$), indicating that holding qualifications constant, there is evidence that voter support for the parties has some association with age.

Question 6

Now, add the variable Gender to the model. Relative to men of the same age and qualifications, which parties are women more/less likely to vote for? Does gender improve the model?

```
m3 <- multinom(VoteIntention~ Qualifications + Age + Gender, data = ukvote2010)
```

```
## # weights:  28 (18 variable)
## initial  value 13867.102494
## iter  10 value 13048.625619
## iter  20 value 12872.188089
## final  value 12866.003464
## converged
```

```
summary(m3)
```

```
## Call:
## multinom(formula = VoteIntention ~ Qualifications + Age + Gender,
##     data = ukvote2010)
##
## Coefficients:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative     -0.5177754         0.1225635        -0.02031279
## LiberalDemocrat   0.1972151        -0.2476788        -0.51285167
## Other            -2.3567793         0.3138330         0.03672553
##                 QualificationsTechnical/None/Other          Age   Gendermale
## Conservative                           -0.01336818  0.017871020 -0.151621809
## LiberalDemocrat                        -0.53969678 -0.001347221 -0.005114873
## Other                                   0.16199214  0.021388164  0.378437170
##
## Std. Errors:
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative      0.0995174        0.06860414         0.08904507
## LiberalDemocrat   0.1051109        0.07425206         0.10269243
## Other             0.1509826        0.09820945         0.13148227
##                 QualificationsTechnical/None/Other         Age Gendermale
## Conservative                           0.06510873 0.001831183 0.05074703
## LiberalDemocrat                        0.07299600 0.002019117 0.05692825
## Other                                  0.09346332 0.002630122 0.07339890
##
## Residual Deviance: 25732.01
## AIC: 25768.01
```

```
zvals <- coef(m3)/summary(m3)$standard.errors
pvals <- (1 - pnorm(abs(zvals), 0, 1)) * 2

exp(coef(m3))
```

```
##                 (Intercept) QualificationsGCE QualificationsGCSE
## Conservative     0.59584461         1.1303909          0.9798921
## LiberalDemocrat  1.21800599         0.7806106          0.5987856
## Other            0.09472481         1.3686611          1.0374082
##                 QualificationsTechnical/None/Other       Age Gendermale
## Conservative                             0.9867208 1.0180317  0.8593132
## LiberalDemocrat                          0.5829250 0.9986537  0.9948982
## Other                                    1.1758510 1.0216185  1.4600011
```

```
pvals
```

```
##                  (Intercept) QualificationsGCE QualificationsGCSE
## Conservative    1.962418e-07      0.0740130722       8.195544e-01
## LiberalDemocrat 6.062007e-02      0.0008510047       5.912444e-07
## Other           0.000000e+00      0.0013956590       7.799998e-01
##                 QualificationsTechnical/None/Other          Age   Gendermale
## Conservative                          8.373215e-01 0.000000e+00 2.809963e-03
## LiberalDemocrat                       1.429967e-13 5.046234e-01 9.284082e-01
## Other                                 8.305724e-02 4.440892e-16 2.524193e-07
```

Relative to women of the same age and qualification (i.e. the effect of these two variables is controlled for in the model), men are less likely to vote Conservative than Labour, and are more likely than women to vote for other parties. There is little difference between women and men's propensity to vote for the Liberal Democrats versus Labour, given the same age and qualifications.