# Multinomial Regression for Student Program Choice Prediction

*Rahul Raghatate [rraghata@iu.edu]*

*October 26, 2017*

Students entering high school can make program choices among general program, vocational program and academic program.

So, to predict the program type ,

Lets try to fit multinomial logistic regression for their choice of program with respective to other categorical and ordinal variables

Data: https://stats.idre.ucla.edu/stat/data/hsbdemo.dta

## Data Import

Lets import the data and have a look at summary of it.

```
# library(rio)
hsb_data = read.dta("hsbdemo.dta")
# Required predictors and dependent variable
head(hsb_data[, c(3, 7, 5)])
```

```
##      ses write     prog
## 1    low    35 vocation
## 2 middle    33  general
## 3   high    39 vocation
## 4    low    37 vocation
## 5 middle    31 vocation
## 6   high    36  general
```

```
summary(hsb_data[, c(3, 7, 5)])
```

```
##      ses         write             prog
##  low   :47   Min.   :31.00   general : 45
##  middle:95   1st Qu.:45.75   academic:105
##  high  :58   Median :54.00   vocation: 50
##              Mean   :52.77
##              3rd Qu.:60.00
##              Max.   :67.00
```

```
# baseline level of 'prog' dependent variable
hsb_data$prog_l <- relevel(hsb_data$prog, ref = "academic")
```

## Exploring bivariate relation counts

```
# ses ~ prog_l
with(hsb_data, table(ses, prog_l))
```

```
##         prog_l
## ses      academic general vocation
##   low          19      16       12
##   middle       44      20       31
##   high         42       9        7
```

```
# write ~ prog_l
with(hsb_data, do.call(rbind, tapply(write, prog_l, function(x) c(M = mean(x),
    SD = sd(x)))))
```

```
##                 M       SD
## academic 56.25714 7.943343
## general  51.33333 9.397775
## vocation 46.76000 9.318754
```

## Applying Multinomial Logistic Regression Model

```
# Multinomial Regression
model <- multinom(prog_l ~ ses + write, data = hsb_data)
```

```
## # weights:  15 (8 variable)
## initial  value 219.722458
## iter  10 value 179.982880
## final  value 179.981726
## converged
```

```
## Call:
## multinom(formula = prog_l ~ ses + write, data = hsb_data)
##
## Coefficients:
##          (Intercept)  sesmiddle    seshigh      write
## general     2.852198 -0.5332810 -1.1628226 -0.0579287
## vocation    5.218260  0.2913859 -0.9826649 -0.1136037
##
## Std. Errors:
##          (Intercept) sesmiddle   seshigh      write
## general     1.166441 0.4437323 0.5142196 0.02141097
## vocation    1.163552 0.4763739 0.5955665 0.02221996
##
## Residual Deviance: 359.9635
## AIC: 375.9635
```

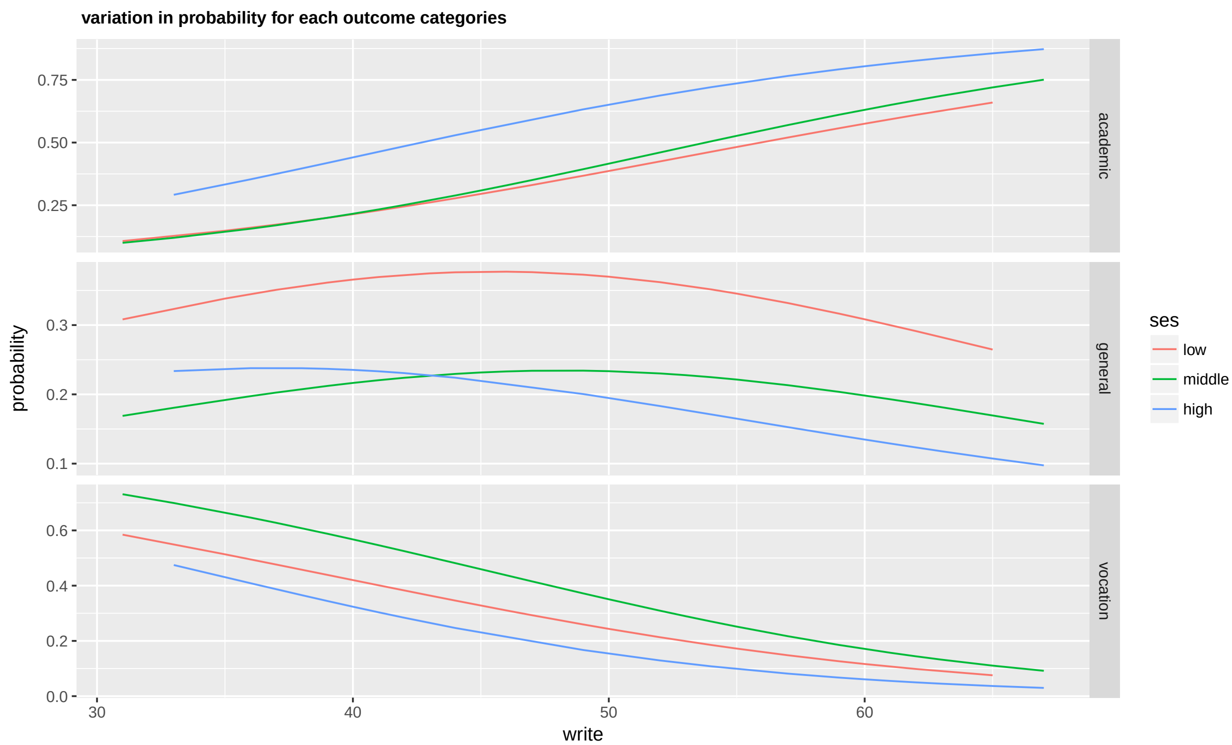
```
##          (Intercept) sesmiddle   seshigh      write
## general     1632.582 -41.33231 -68.73974  -5.628277
## vocation   18361.262  33.82809 -62.56877 -10.738840
```

First row compares prog = "general" to baseline prog = "academic" . Second row compares prog = "vocation" to baseline prog = "academic".

Interpretations for the General vs. Academic model:
$\ln(\frac{P(prog="general")}{P(prog="academic")}) = b_{10} + b_1 1 * (ses = 2) + b_{12} * (ses = 3) + b_{13} * write$
Interpretations for the Vocational vs. Academic model:
$\ln(\frac{P(prog="vocation")}{P(prog="academic")}) = b_{20} + b211 * (ses = 2) + b_{12} * (ses = 3) + b_{23} * write$

### Plot the fitted values against writing score and social economic status

```
model.df <- data.frame(ses = hsb_data$ses, write = hsb_data$write)
model.df.prob <- cbind(model.df, fitted.values(model))
# dataframe for ggplot
pred_prob.ggplot <- melt(model.df.prob, id.vars = c("ses", "write"), value.name = "probability")
## plot of fitted probabilities facet over program type
ggplot(pred_prob.ggplot, aes(x = write, y = probability, colour = ses)) + geom_line() +
    facet_grid(variable ~ ., scales = "free") + ggtitle(" variation in probability for each outcome cate
    theme(plot.title = element_text(size = 10, face = "bold"))
```




### Based on the models defined, for "ses=middle and write=54", making a prediction for prog

```
data <- data.frame(ses = "middle", write = 54)
cat("\nFor ses=middle and write=54, the predicted probabilities for prog_type are:\n",
    predict(model, newdata = data, "probs"))
```

```
##
## For ses=middle and write=54, the predicted probabilities for prog_type are:
##  0.504927 0.2247932 0.2702798
```

```
probs <- predict(model, newdata = data, "probs")
# Calculating required ratios

vocation_over_academic = probs[3]/probs[1]
```

```r
general_over_academic = probs[2]/probs[1]

cat("\n\nProbability of choosing each outcome category over the baseline category are \n")
```

```
##
##
## Probability of choosing each outcome category over the baseline category are
```

```r
vocation_over_academic
```

```
## vocation
## 0.535285
```

```r
general_over_academic
```

```
##   general
## 0.4451995
```

## Proportional odds logistic regression

Lets treat prog as an ordered categorical variable and fit a proportional odds logistic regression with the same predictors ses and write.

```r
# Model
hsb_data.polr = polr(factor(prog) ~ ses + write, data = hsb_data)
display(hsb_data.polr)
```

```
## polr(formula = factor(prog) ~ ses + write, data = hsb_data)
##                   coef.est coef.se
## sesmiddle          0.68     0.36
## seshigh            0.40     0.38
## write             -0.04     0.02
## general|academic  -3.15     0.87
## academic|vocation -0.71     0.83
## ---
## n = 200, k = 5 (including 2 intercepts)
## residual deviance = 397.0, null deviance is not computed by polr
```

```r
probs <- predict(hsb_data.polr, data.frame(ses = "middle", write = 54), type = "probs")
cat("\nProbability for each outcome category.\n")
```

```
##
## Probability for each outcome category.
```

```r
probs
```

```
##   general  academic  vocation
## 0.1780447 0.5357214 0.2862339
```

**Therefore, prediction for prog given ses=middle and write=54 is "academic".**

We also have their reading, math, science and standardized test score for social studies in the data set ("read", "math", "science" and "socst").

## PCA for above variables and "write" to explore first two principal components

```
pca_cols <- c("read", "math", "science", "socst", "write")
pca_data <- hsb_data[pca_cols]
data.pca = prcomp(pca_data, scale. = TRUE)
data.pca
```

```
## Standard deviations (1, .., p=5):
## [1] 1.8387006 0.7465777 0.6378031 0.5967980 0.5466638
##
## Rotation (n x k) = (5 x 5):
##               PC1         PC2           PC3         PC4        PC5
## read    0.4664184 -0.02727868  0.5312736731 -0.02057541 -0.7064239
## math    0.4587755 -0.26090184  0.0005952692 -0.78003732  0.3361498
## science 0.4355824 -0.61089329  0.0069539237  0.58947561  0.2992449
## socst   0.4256688  0.71757896  0.2595770518  0.20131689  0.4426938
## write   0.4483893  0.20754742 -0.8064237887  0.05575345 -0.3200677
```

```
ggbiplot(data.pca, obs.scale = 1, groups = hsb_data$prog) + ggtitle("Biplot for first two principal comp
    theme(plot.title = element_text(size = 10, face = "bold"))
```