Assignment 2

Kishan Sarpangala

**Question:**

Homework Sheet No. 2 Due Date: February 15, 2018 Maximum points: 30 Theme: Multinomial Logistic Regression vs Proportional Odds Model.

**Answer:**

One can download data directly from the internet into R.

The following is address for the dataset

<https://stats.idre.ucla.edu/stat/data/ologit.dta>

The data comes from the University of California Los Angeles.

It is in the format ‘.dta’ We need the package ‘foreign’ to get this data onto R.

Download and activate it. The following R code should facilitate downloading.

MB <- read.dta(“https://stats.idre.ucla.edu/stat/data/ologit.dta”)

**Defining our variables**:

apply = Likelihood of college juniors applying to grad school. (Self-reported)(very likely, somewhat likely, unlikely)

 pared = Does at least one parent have a graduate degree? (no=0, yes=1)

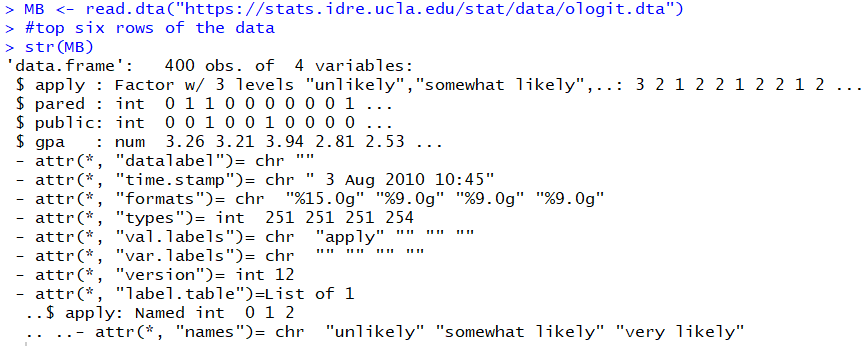
public = Undergrad was a private or public institution. (private = 0, public = 1)

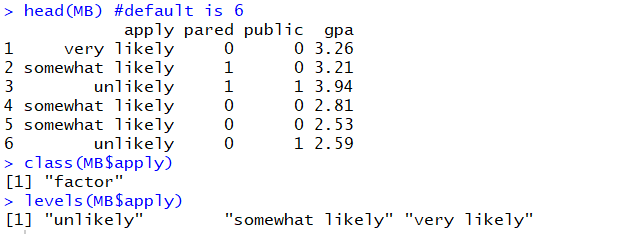
gpa = Undergrad grade point average

**Multinomial Logistic Regression**

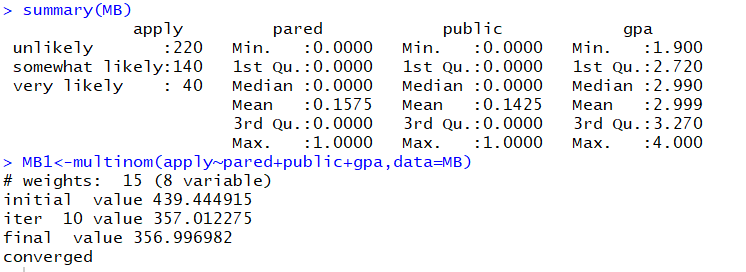
1. Write down the dimensions of the data.
2. Show the top six rows of the data.
3. Get the summary statistics of the data. The statistics should be meaningful.
4. Postulate the multinomial logistic regression model.
5. Fit the model.
6. Comment on the coefficients.
7. Check goodness-of-fit.
8. Present the model graphically.
9. Obtain the confusion matrix.
10. Calculate the misclassification rate.

dimensions of the data and top six rows of the data presented below





summary statistics of the data is given below



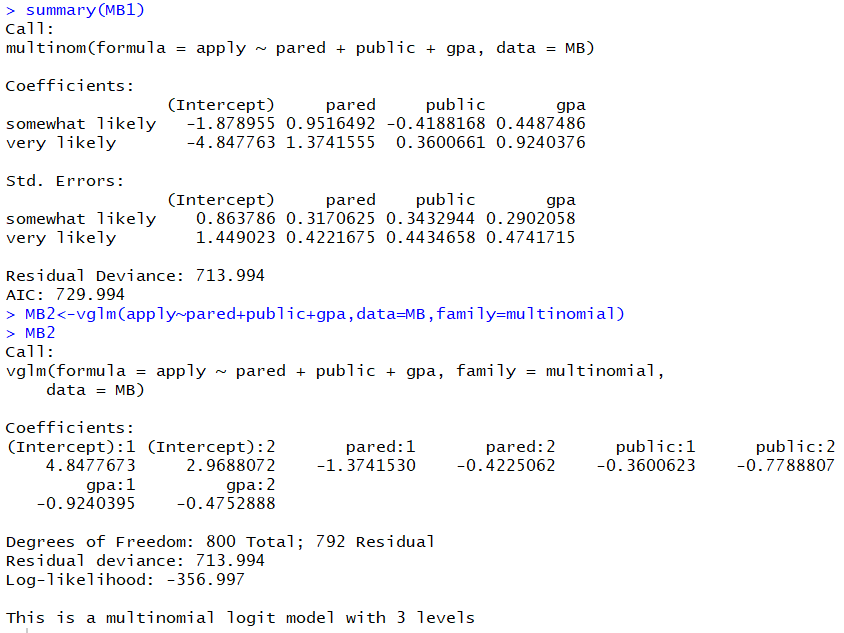
We fit the model using the multinom function

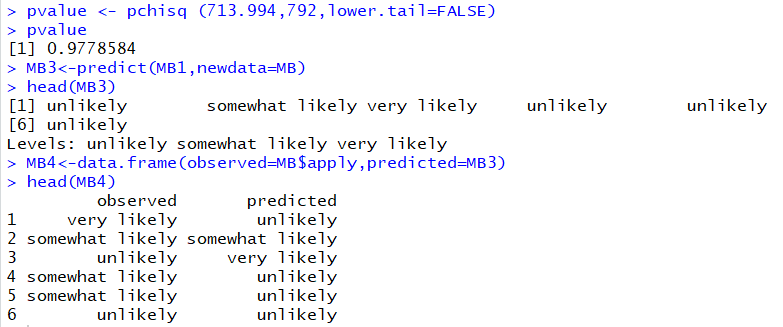
Comment on the coefficients:

Pr of somewhat likely = exp(-1.88 +0.95\*pared-0.41\*public+0.44\*gpa)/D

Pr of very likely = exp(-4.85+1.37\*pared+0.36\*public+0.92\*gpa)/D

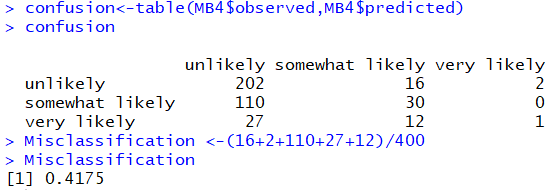
Pr of unlikely = 1/D

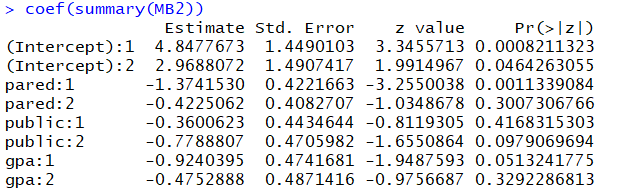




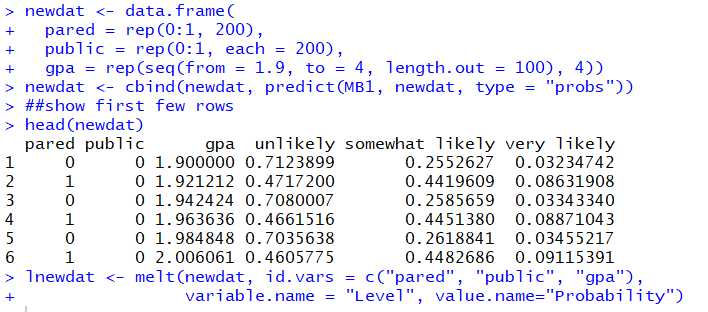
Obtained the confusion matrix.

Calculated the misclassification rate.

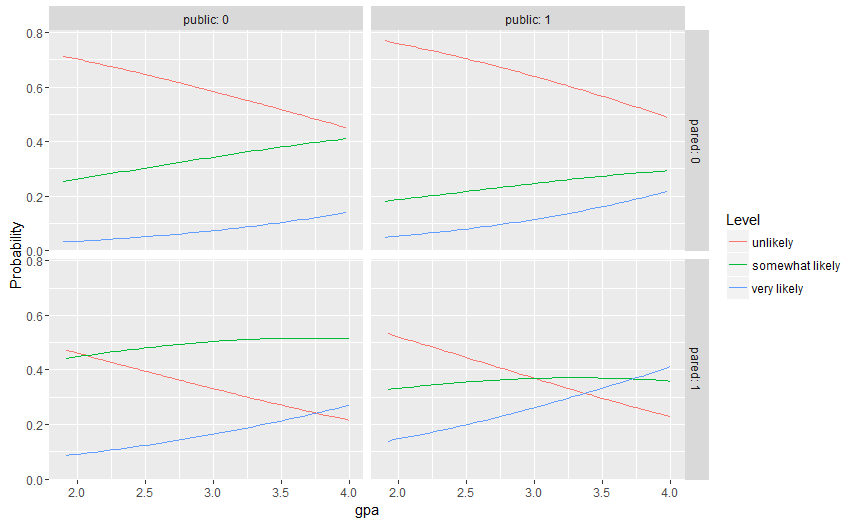




Error row in the Deviance Table which depicts goodness of fit.



Presenting the model graphically (below)

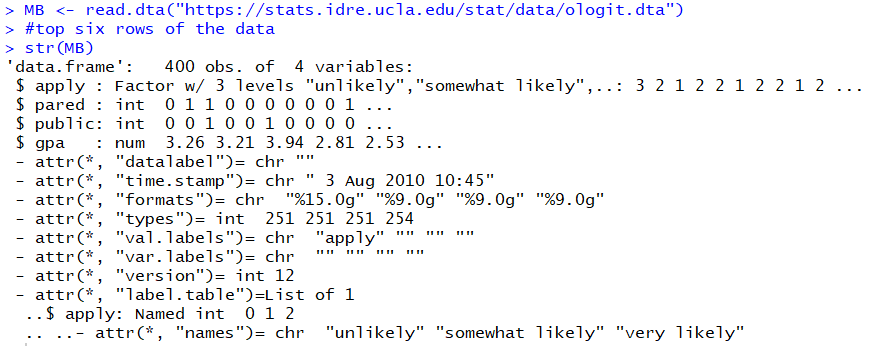


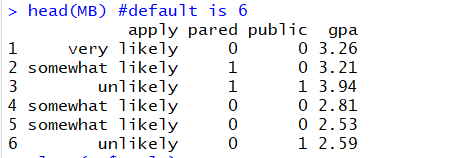
**Proportional Odds Model**:

1. Postulate the proportional odds model.
2. Fit the model.
3. Comment on the coefficients.
4. Check goodness-of-fit.
5. Present the model graphically.
6. Obtain the confusion matrix.
7. Calculate the misclassification rate.
8. Compare the models.

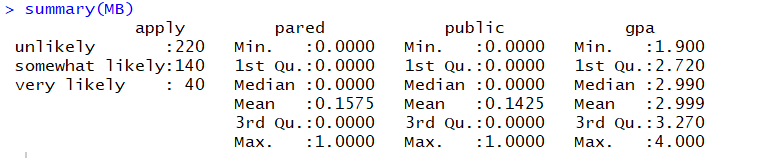
**Terminal Output**:

dimensions of the data and top six rows of the data presented below

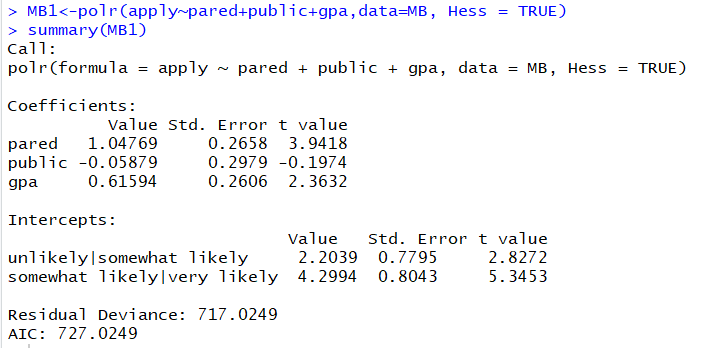




summary statistics of the data



We fit the model using the polr function from the MASS package. “polr” stands for Proportional Odds Linear Regression. The MASS package comes with R. (Incidentally, MASS stands for Modern Applied Statistics)

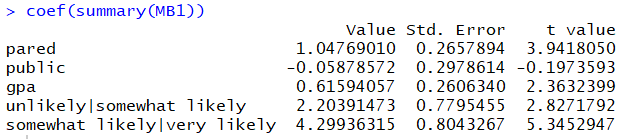


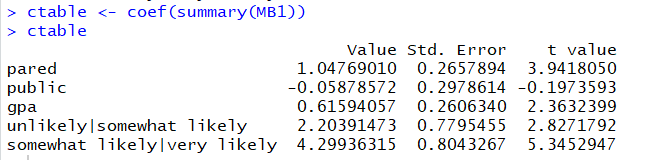
Coefficients generated

Pared and gpa have a direct relationship with the probability of applying, while public has a negative relationship to the outcome.

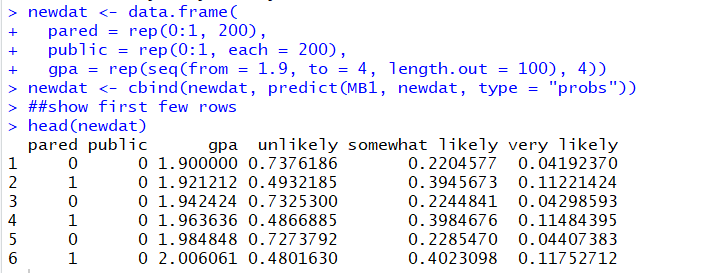
Pr of unlikely/somewhat likely = exp(2.2 +1.04\*pared-0.05\*public+0.61\*gpa)/ (1+ exp(2.2 +1.04\*pared-0.05\*public+0.61\*gpa)

Pr of somewhat likely/very likely = exp(4.2+1.04\*pared-0.05\*public+0.61\*gpa) /(1+ exp(4.2+1.04\*pared-0.05\*public+0.61\*gpa)

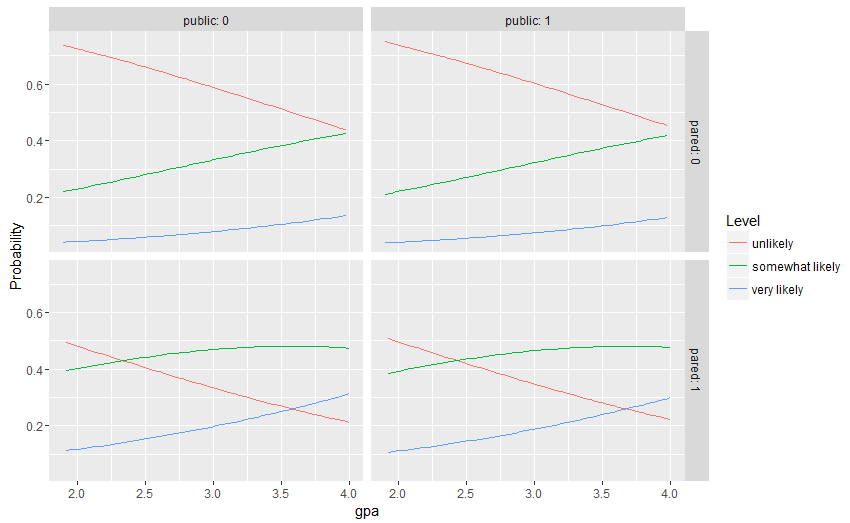


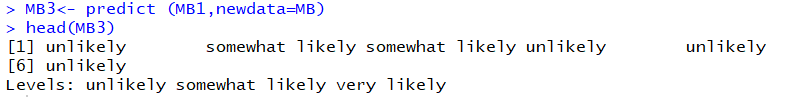


Error row in the Deviance Table which depicts goodness of fit.

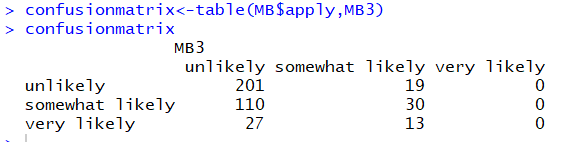


Model presented graphically.





Obtained the confusion matrix.



Calculated the misclassification rate



The **proportional odds model** is a class of generalized linear models used for modelling the dependence of an ordinal response on discrete or continuous covariates

**Multinomial logistic regression** (often just called 'multinomial regression') is used to predict a nominal dependent variable given one or more independent variables. It is sometimes considered an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

**Ordinal Logistic Regression: The Proportional Odds Model**

When the response categories are ordered, you could run a multinomial regression model.  The disadvantage is that you are throwing away information about the ordering. An ordinal logistic regression model preserves that information, but it is slightly more involved.

In the Proportional Odds Model, the event being modeled is not having an outcome in a single category, as is done in the binary and multinomial models.

**Difference/ Compare the models**:

In the case of the multinomial one has no intrinsic ordering; in contrast in the case of ordinal regression there is an association between the levels.

Code: Multinomial Logistic Regression

require(foreign)

require(ggplot2)

require(MASS)

library(reshape2)

require(reshape2)

require(nnet)

library (VGAM)

MB <- read.dta("https://stats.idre.ucla.edu/stat/data/ologit.dta")

#top six rows of the data

str(MB)

head(MB) #default is 6

class(MB$apply)

levels(MB$apply)

MB$pared <- as.factor(MB$pared)

MB$public <- as.factor(MB$public)

summary(MB)

MB1<-multinom(apply~pared+public+gpa,data=MB)

summary(MB1)

MB2<-vglm(apply~pared+public+gpa,data=MB,family=multinomial)

pvalue <- pchisq (713.994,792,lower.tail=FALSE)

pvalue

MB3<-predict(MB1,newdata=MB)

head(MB3)

MB4<-data.frame(observed=MB$apply,predicted=MB3)

head(MB4)

confusion<-table(MB4$observed,MB4$predicted)

confusion

Misclassification <-(16+2+110+27+12)/400

Misclassification

newdat <- data.frame(

pared = rep(0:1, 200),

public = rep(0:1, each = 200),

gpa = rep(seq(from = 1.9, to = 4, length.out = 100), 4))

newdat <- cbind(newdat, predict(MB1, newdat, type = "probs"))

##show first few rows

head(newdat)

lnewdat <- melt(newdat, id.vars = c("pared", "public", "gpa"),

variable.name = "Level", value.name="Probability")

ggplot(lnewdat, aes(x = gpa, y = Probability, colour = Level)) +

geom\_line() + facet\_grid(pared ~ public, labeller="label\_both")

Code: Proportional odds model

require(foreign)

require(ggplot2)

require(MASS)

library(reshape2)

require(reshape2)

require(nnet)

library (VGAM)

MB <- read.dta("https://stats.idre.ucla.edu/stat/data/ologit.dta")

#top six rows of the data

str(MB)

head(MB) #default is 6

class(MB$apply)

levels(MB$apply)

summary(MB)

MB1<-polr(apply~pared+public+gpa,data=MB, Hess = TRUE)

summary(MB1)

coef(summary(MB1))

ctable <- coef(summary(MB1))

ctable

newdat <- data.frame(

pared = rep(0:1, 200),

public = rep(0:1, each = 200),

gpa = rep(seq(from = 1.9, to = 4, length.out = 100), 4))

newdat <- cbind(newdat, predict(MB1, newdat, type = "probs"))

##show first few rows

head(newdat)

lnewdat <- melt(newdat, id.vars = c("pared", "public", "gpa"),

variable.name = "Level", value.name="Probability")

## view first few rows

head(lnewdat)

ggplot(lnewdat, aes(x = gpa, y = Probability, colour = Level)) +

geom\_line() + facet\_grid(pared ~ public, labeller="label\_both")

MB3<- predict (MB1,newdata=MB)

head(MB3)

confusionmatrix<-table(MB$apply,MB3)

confusionmatrix

misclassificationrate<-(19+110+40)/(201+110+27+19+30+13)

misclassificationrate