Statistical Methods for Discrete Response, Time Series, and Panel Data: Live Session 4

Devesh Tiwari and Jeffrey Yau

1/30/2017

Agenda

1. Q&A (estimated time: 5-10 minutes)
2. Data Analysis and Modeling Exercises (estimated time: 80 minutes)

a. Instructor's introduction of the exercise and the dataset (estimated time: 5 minutes)
b. Discussion 1 (estimated time: 10 minutes)
c. Discussion 2 (estimated time: 10 minutes)
d. Discussion 3 (estimated time: 15 minutes)
e. Discussion 4 (estimated time: 20 minutes)
f. Discussion 5 (estimated time: 10 minutes)
g. Discussion 6 (estimated time: 10 minutes)

Introduction (estimated time: 5 minutes)

In this exercise, we will explore the relationship between voters' self identified party affiliation and their demographic characteristic. In particular, we seek to answer whether voters' age, race, and gender influence their party choice. For this exercise we will use the data from the **American National Election Survey**, which conducted a survey several months prior to the 2016 American Presidential elections. Note that the original survey data uses survey weights, which we will not be using.

The dataset "*w271_spring2017_anes.csv*" contains a handful of variables from the survey, and these variables have been cleaned and modified for this exercise. This dataset contains the following variables:

Variable Name	Explanations
ftwhite, ftblack, ftmuslim	Feeling thermometer variables where respondents are asked to rate their favorability of whites, blacks, and muslims, on a 0 – 100 scale.
Presjob	A seven point scale indicating respondents'
Srv_spend	Seven point scale representing the degree to
crimespend	A seven point scale representing degree to
ideo5	A five point scale of respondents' self
party	Categorical variable indicating respondents'
age	Respondents' age, as of 2016.
race_white	Dummy variable taking a value of one if the
female	Dummy variable taking a value of one if the

Discussion 1 (Estimated Time (10 minutes) - 5 mins in breakout session, 5 min class-wide discussion):

The US has two major political parties. The Democratic Party is considered to be the ideologically libearl party while the Republican Party is considered to be the ideologically conservative party. A non-trivial proportion of American voters either identify themselves as being Independnet or supporting other parties. In this dataset, voters are either Democratic, Republican, or Independent.

Question: What is the difference between modeling voters' party affiliation using a multinomial logistic regression model as opposed to using an ordinal logistic regression model? Under what circumstances would be OK to use an ordinal model?

Discussion 2: Assessing the independence of race, gender, and partisanship # (Estimated Time (10 minutes) - 5 mins in breakout session, 5 min class-wide discussion):

In a breakout session, discuss the following analysis of the dataset and EDA.

Take home exericse: Conduct a thorough EDA, including other variables in the dataset. For this live session, we instead focus on understanding a few bivariate relationships.

Insert a function to *tidy up* the code when they are printed out

```
rm(list = ls())
require(knitr)
```

```
## Loading required package: knitr
```

```
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

```
require(vcd)
```

```
## Loading required package: vcd
## Warning: package 'vcd' was built under R version 3.2.5
## Loading required package: grid
```

```
require(nnet)
```

```
## Loading required package: nnet
```

```
require(car)
```

```
## Loading required package: car
```

```
require(MASS)
```

```
## Loading required package: MASS
```

```
# path <- '/Users/DKT/Documents/Projects/anes2016'
path <- "~/Documents/JStuff/Teach/w271/LiveSessions/week04"
setwd(path)

df <- read.csv("w271_spring2017_anes.csv", stringsAsFactors = FALSE,
    header = TRUE, sep = ",")
```

Examine the data before conducting EDA

```
str(df)
```

```
## 'data.frame':    1200 obs. of  11 variables:
##  $ ftwhite   : int  100 74 50 64 58 51 70 70 50 90 ...
##  $ ftblack   : int  100 6 50 61 61 50 100 70 50 75 ...
##  $ ftmuslim  : int  20 22 5 61 22 11 100 40 12 72 ...
##  $ presjob   : int  1 3 7 2 7 7 2 7 7 2 ...
##  $ srv_spend : int  7 6 2 6 1 1 7 3 1 6 ...
##  $ crimespend: int  5 2 5 4 4 7 2 4 4 3 ...
##  $ party     : chr  "Democrat" "Independent" "Republican" "Democrat" ...
##  $ ideo5     : int  NA 2 4 2 4 4 1 5 4 2 ...
##  $ age       : int  56 59 53 36 42 58 38 65 43 80 ...
##  $ race_white: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ female    : int  0 1 0 0 0 0 0 0 0 0 ...
```

```
# Number of incomplete cases in the dataset There are a
# number of ways to accomplish this task The first one will
# list the entire dataframe (when printed out to a pdf or
```

```r
# html file) all of the observations with incomplete
# observations. The second one just count the number of
# missing data in each of the variables

# df[!complete.cases(df),]
sapply(df, function(x) sum(is.na(x)))
```

```
##    ftwhite    ftblack   ftmuslim    presjob  srv_spend crimespend
##          1          4          4          0          0          2
##      party     ideo5        age race_white     female
##         81         90          0          0          0
```

Let's select only the data that we need before conducting the analysis

```r
require(dplyr)
```

```
## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.2.5

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
df2 <- df %>% select(party, age, female, race_white)
str(df2)
```

```
## 'data.frame':    1200 obs. of  4 variables:
##  $ party     : chr  "Democrat" "Independent" "Republican" "Democrat" ...
##  $ age       : int  56 59 53 36 42 58 38 65 43 80 ...
##  $ female    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ race_white: int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Number of incomplete cases in the dataset
sapply(df2, function(x) sum(is.na(x)))
```

```
##      party        age     female race_white
##         81          0          0          0
```

There are still 81 observations with missing values.

Take-home Exercise: Examine if the missing value has any relationship with other variables? For instance, does all of the missing values in the party variable fall into certain age, gender, and/or race groups?

For now, we would simply exclude them in our analysis. Again, in practice, you do not just want to throw away observations without any investigation; we leave it as take-home exericse for this very simple case.

Include only complete cases

```r
df3 <- df2[complete.cases(df2), ]
str(df3)
```

```
## 'data.frame':    1119 obs. of  4 variables:
##  $ party     : chr  "Democrat" "Independent" "Republican" "Democrat" ...
##  $ age       : int  56 59 53 36 58 38 65 43 80 38 ...
##  $ female    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ race_white: int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
sapply(df3, function(x) sum(is.na(x)))
```

```
##      party        age     female race_white
##          0          0          0          0
```

Discussion 3: Assessing the independence of race, gender, and partisanship # (Estimated Time (15 minutes) - 7 mins in breakout session, 8 min class-wide discussion):

```r
# A few descriptive statistics
require(Hmisc)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

## Warning: replacing previous import by 'ggplot2::unit' when loading 'Hmisc'

## Warning: replacing previous import by 'ggplot2::arrow' when loading 'Hmisc'

## Warning: replacing previous import by 'scales::alpha' when loading 'Hmisc'

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
describe(df3)
```

```
## df3
##
##  4  Variables      1119  Observations
```

```
## --------------------------------------------------------------------------
## party
##       n missing  unique
##    1119       0       3
##
## Democrat (459, 41%), Independent (380, 34%)
## Republican (280, 25%)
## --------------------------------------------------------------------------
## age
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##    1119       0      72       1   48.25      22      25      34      49
##     .75     .90     .95
##      62      71      76
##
## lowest : 19 20 21 22 23, highest: 89 90 91 92 95
## --------------------------------------------------------------------------
## female
##       n missing  unique    Info     Sum    Mean
##    1119       0       2    0.75     593  0.5299
## --------------------------------------------------------------------------
## race_white
##       n missing  unique    Info     Sum    Mean
##    1119       0       2     0.6     813  0.7265
## --------------------------------------------------------------------------
```

```
party.gender.table <- xtabs(~party + female, data = df3)
prop.table(party.gender.table)
```

```
##              female
## party                 0         1
##   Democrat    0.1689008 0.2412869
##   Independent 0.1858803 0.1537087
##   Republican  0.1152815 0.1349419
```

```
chisq.test(party.gender.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  party.gender.table
## X-squared = 15.477, df = 2, p-value = 0.0004357
```

```
assocstats(party.gender.table)
```

```
##                     X^2 df   P(> X^2)
## Likelihood Ratio 15.501  2 0.00043048
## Pearson          15.477  2 0.00043571
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.117
## Cramer's V        : 0.118
```

```
party.race.table <- xtabs(~party + race_white, data = df3)
prop.table(party.race.table)
```

```
##              race_white
## party                  0          1
```

```
##   Democrat    0.16711349 0.24307417
##   Independent 0.07327971 0.26630920
##   Republican  0.03306524 0.21715818
```

```r
chisq.test(party.race.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  party.race.table
## X-squared = 75.956, df = 2, p-value < 2.2e-16
```

```r
assocstats(party.race.table)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 77.480  2        0
## Pearson          75.956  2        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.252
## Cramer's V        : 0.261
```

Evidence suggests that party affiliation is not independent from respondents' gender or race. These contingnecy tables do not tell us if there is an ordered relationship between these demographic variables and party affiliation.

```r
prop.table(party.gender.table, 2)
```

```
##              female
## party                 0         1
##   Democrat    0.3593156 0.4553120
##   Independent 0.3954373 0.2900506
##   Republican  0.2452471 0.2546374
```

```r
prop.table(party.race.table, 2)
```

```
##              race_white
## party                 0         1
##   Democrat    0.6111111 0.3345633
##   Independent 0.2679739 0.3665437
##   Republican  0.1209150 0.2988930
```

Based on this output, do you think that there is an ordered relationship between the demographic variables and party affiliation? Why or why not?

Question: How could you explore the bivariate relationship between party and age? Can you think of a way where you can use a contingency table and chi-square test to test for independence?

Discussion 4: Assessing the independence of race, gender, and partisanship in the context of a Multinomial Logistic Regression Model (Estimated Time (20 minutes) - 10 mins in breakout session, 10 min class-wide discussion):

Multinomial Logistic Regression Model

We are going to use a multinomial logistic regression model to examine the relationship between repondents' party affiliation and their age, race, and gender. Remember that we usually do this only after we have conducted a thorough EDA and justified our modeling decision!

Question: Using the following results, interpret and discuss the model results. What do these coefficients mean? Why are there two sets of coefficients? What does it mean if we were to take the anti-log of the coefficients? If needed, write some R codes to transform the estimated parameters to interpret the results in terms of (a) odds ratios, and (b) probability in a particular party. Finally, discuss the results of the following hypothesis test.

Question: Suppose that you only know how to use logistic regression. How would you use binary logit to answer the questions that motivated this lab? Does it make sense to create three different dummy variables for Democrat, Independent, and Republican as the dependent variables?

```r
mod.nominal1 <- multinom(party ~ female + race_white + age, data = df3)
```

```
## # weights:  15 (8 variable)
## initial  value 1229.347151
## iter  10 value 1158.681844
## final  value 1158.296836
## converged
```

```r
summary(mod.nominal1)
```

```
## Call:
## multinom(formula = party ~ female + race_white + age, data = df3)
##
## Coefficients:
##             (Intercept)     female race_white          age
## Independent  -0.3398992 -0.5347606   0.938838 -0.004691314
## Republican   -1.8071816 -0.1967837   1.463081  0.006711978
##
## Std. Errors:
##             (Intercept)    female race_white         age
## Independent   0.2363950 0.1426915   0.1601700 0.004265171
## Republican    0.2874293 0.1579126   0.2025863 0.004646449
##
## Residual Deviance: 2316.594
## AIC: 2332.594
```

```r
exp(coefficients(mod.nominal1))
```

```
##             (Intercept)    female race_white       age
## Independent   0.7118421 0.5858095   2.557008 0.9953197
## Republican    0.1641160 0.8213683   4.319245 1.0067346
```

```r
# Examine statistical signficance of model and coef.
Anova(mod.nominal1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: party
##            LR Chisq Df Pr(>Chisq)
## female       14.367  2  0.0007592 ***
## race_white   72.677  2  < 2.2e-16 ***
## age           5.904  2  0.0522392 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
test.stats <- summary(mod.nominal1)$coefficients/summary(mod.nominal1)$standard.errors
test.stats
```

```
##             (Intercept)    female race_white       age
## Independent   -1.437844 -3.747670    5.86151 -1.099912
## Republican    -6.287395 -1.246156    7.22201  1.444539
```

```r
# It appears as if age might not be statistically
# significant!  Let's examine statistical sig using LRT

mod.nominal.noage <- multinom(party ~ female + race_white, data = df3)
```

```
## # weights:  12 (6 variable)
## initial  value 1229.347151
## iter  10 value 1161.249009
## final  value 1161.248758
## converged
```

```r
summary(mod.nominal.noage)
```

```
## Call:
## multinom(formula = party ~ female + race_white, data = df3)
##
## Coefficients:
##             (Intercept)     female race_white
## Independent  -0.5400052 -0.5394630  0.9106016
## Republican   -1.5124478 -0.1887458  1.5056499
##
## Std. Errors:
##             (Intercept)    female race_white
## Independent   0.1515344 0.1425658  0.1578609
## Republican    0.2003729 0.1576405  0.2005126
##
## Residual Deviance: 2322.498
## AIC: 2334.498
```

```r
anova(mod.nominal.noage, mod.nominal1)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: party
##                        Model Resid. df Resid. Dev   Test    Df LR stat.
## 1       female + race_white      2232   2322.498
## 2 female + race_white + age      2230   2316.594 1 vs 2     2 5.903844
```

```
##      Pr(Chi)
## 1
## 2 0.05223921
```

According to the LRT, age is not a statistically signficant variable. However, it might be worth visualizing its impact on predicted probabilities. Let's examine the impact of age on respondents' party affiliation. We will generate predicted probability plots for white men between the ages of 20 and 80.

Discussion 5: Discuss the visuals of the model results. # (Estimated Time (10 minutes) - 5 minutes breakout room discussion; 5 min class-wide discussion):

Question: Why do you think that the dashed lines are parallel to the solid lines? What does this chart tell you about the impact of race and gender on party affiliation?

Suppose you were interested in whether the relationship between party affiliation and age is different for white respondents than persons of color? How would you test this? Estimate this model, comment on the statistical signficance of this model and the interaction term, and graph the predicted probability of the party affiliation of white men by age.

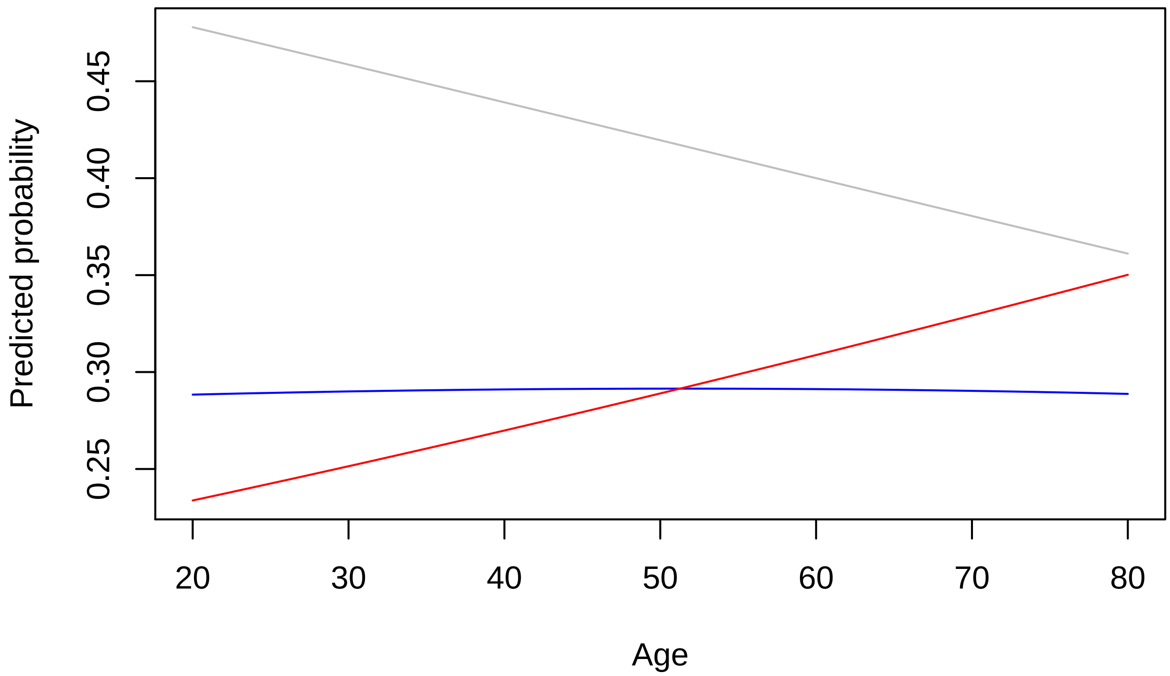
```
simulated.data <- data.frame(female = 0, race_white = 1, age = 20:80)

pi.hat.nom1 <- predict(mod.nominal1, newdata = simulated.data,
    type = "probs")
head(pi.hat.nom1)
```

```
##    Democrat Independent Republican
## 1 0.2883616   0.4778647  0.2337737
## 2 0.2885527   0.4759433  0.2355040
## 3 0.2887379   0.4740199  0.2372422
## 4 0.2889173   0.4720944  0.2389883
## 5 0.2890907   0.4701670  0.2407423
## 6 0.2892583   0.4682377  0.2425040
```

```
tail(pi.hat.nom1)
```

```
##     Democrat Independent Republican
## 56 0.2896048   0.3707795  0.3396157
## 57 0.2894451   0.3688406  0.3417143
## 58 0.2892788   0.3669034  0.3438179
## 59 0.2891058   0.3649678  0.3459264
## 60 0.2889263   0.3630340  0.3480397
## 61 0.2887401   0.3611021  0.3501579
```

```
plot.new()
x <- 20:80
plot(x, pi.hat.nom1[, 1], type = "l", col = "blue", ylim = range(min(pi.hat.nom1),
    max(pi.hat.nom1)), xlab = "Age", ylab = "Predicted probability")
lines(x, pi.hat.nom1[, 2], col = "gray")
lines(x, pi.hat.nom1[, 3], col = "red")
```



Age seems to have little relationship with the probability a white male is a Democrat, but it makes a large substantative impact on the probability a white male is Independent or Republican.

Now, let's see if the same holds for women of color!

```
simulated.data <- data.frame(female = 1, race_white = 0, age = 20:80)

pi.hat.nom1.womenofcolor <- predict(mod.nominal1, newdata = simulated.data,
    type = "probs")
head(pi.hat.nom1.womenofcolor)
```

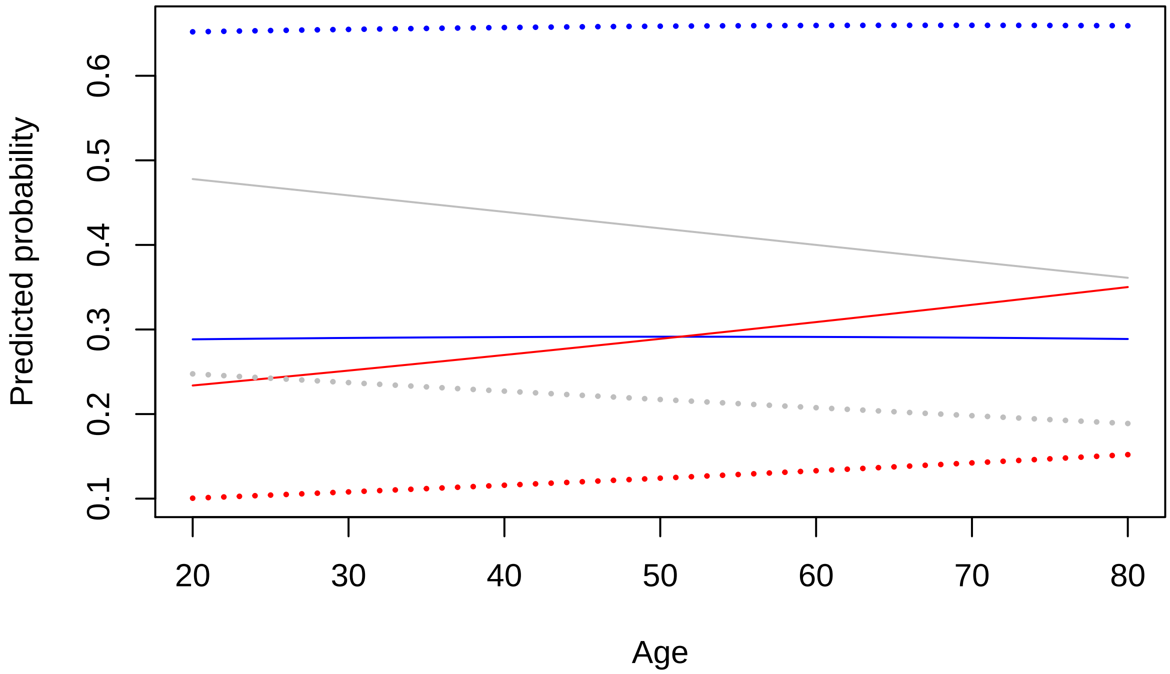
```
##    Democrat Independent Republican
## 1 0.6519656   0.2475235  0.1005109
## 2 0.6522797   0.2464838  0.1012365
## 3 0.6525876   0.2454459  0.1019664
## 4 0.6528893   0.2444101  0.1027006
## 5 0.6531847   0.2433763  0.1034390
## 6 0.6534739   0.2423444  0.1041817
```

```
tail(pi.hat.nom1.womenofcolor)
```

```
##     Democrat Independent Republican
## 56 0.6594920   0.1934389  0.1470691
## 57 0.6594359   0.1925171  0.1480470
## 58 0.6593726   0.1915977  0.1490297
## 59 0.6593021   0.1906806  0.1500173
## 60 0.6592244   0.1897658  0.1510098
## 61 0.6591395   0.1888533  0.1520072
```

```
x <- 20:80
plot.new()
plot(x, pi.hat.nom1[, 1], type = "l", col = "blue", ylim = range(min(cbind(pi.hat.nom1,
    pi.hat.nom1.womenofcolor)), max(cbind(pi.hat.nom1, pi.hat.nom1.womenofcolor))),
    xlab = "Age", ylab = "Predicted probability")
lines(x, pi.hat.nom1[, 2], col = "gray")
```

```r
lines(x, pi.hat.nom1[, 3], col = "red")
points(x, pi.hat.nom1.womenofcolor[, 1], pch = 19, col = "blue",
    cex = 0.25)
points(x, pi.hat.nom1.womenofcolor[, 2], pch = 19, col = "gray",
    cex = 0.25)
points(x, pi.hat.nom1.womenofcolor[, 3], pch = 19, col = "red",
    cex = 0.25)
```



Discussion 6: Discuss the results of the followig ordinal logistic regression model. # (Estimated Time (10 minutes) - 5 minutes breakout room discussion; 5 min class-wide discussion):

Proportional Odds Logistic Regression (or Ordinal Logistic Regression) Model

For illustration purposes, let's model the relationship between respondents' party affiliation and their demographic characteristics.

Question: Why do you think these two charts are different? If you were to conduct a thorough EDA, how could you determine which model is the correct one?

```r
mod.ordered1 <- polr(as.factor(party) ~ female + race_white +
    age, data = df3, method = "logistic", Hess = TRUE)
summary(mod.ordered1)
```

```
## Call:
## polr(formula = as.factor(party) ~ female + race_white + age,
##     data = df3, Hess = TRUE, method = "logistic")
##
```

```
## Coefficients:
##                 Value Std. Error t value
## female     -0.204806    0.11263  -1.818
## race_white  1.109528    0.13465   8.240
## age         0.003567    0.00335   1.065
##
## Intercepts:
##                         Value   Std. Error t value
## Democrat|Independent    0.4949  0.1936     2.5560
## Independent|Republican  2.0382  0.2025    10.0637
##
## Residual Deviance: 2332.911
## AIC: 2342.911
```

```r
summary(mod.nominal1)
```

```
## Call:
## multinom(formula = party ~ female + race_white + age, data = df3)
##
## Coefficients:
##             (Intercept)     female race_white          age
## Independent  -0.3398992 -0.5347606   0.938838 -0.004691314
## Republican   -1.8071816 -0.1967837   1.463081  0.006711978
##
## Std. Errors:
##             (Intercept)    female race_white         age
## Independent   0.2363950 0.1426915  0.1601700 0.004265171
## Republican    0.2874293 0.1579126  0.2025863 0.004646449
##
## Residual Deviance: 2316.594
## AIC: 2332.594
```

```r
# Generate predicted probability chart for white men, between
# ages 20 and 80
simulated.data <- data.frame(female = 0, race_white = 1, age = 20:80)

pi.hat.ord1 <- predict(mod.ordered1, newdata = simulated.data,
    type = "probs")
head(pi.hat.ord1)
```

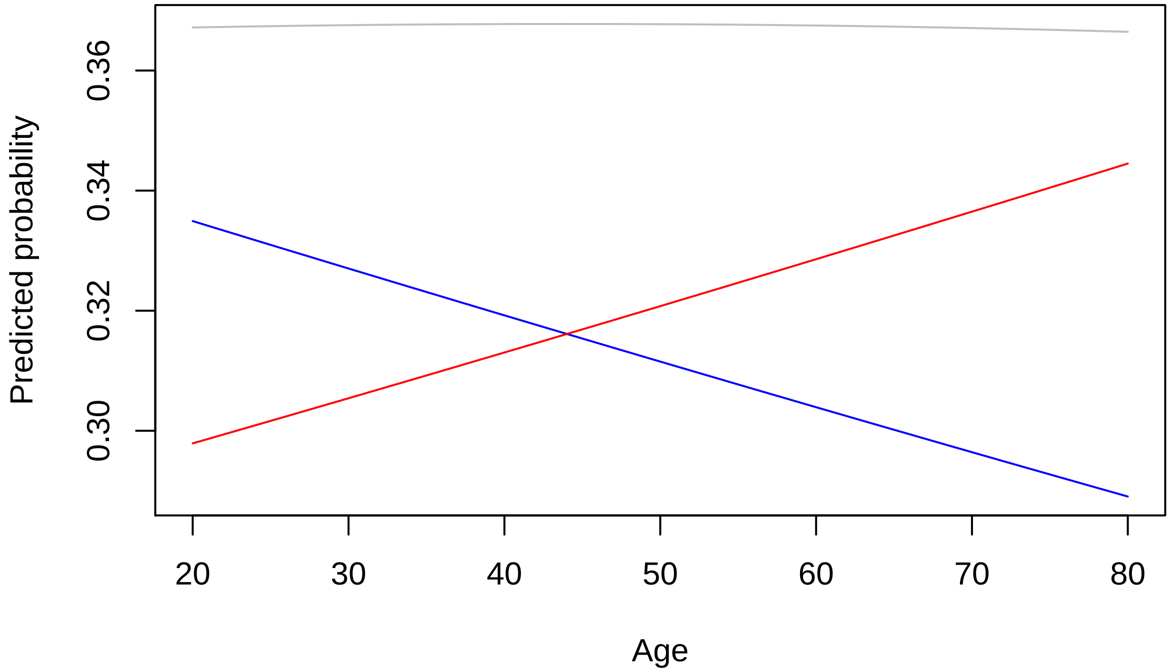
```
##    Democrat Independent Republican
## 1 0.3349249   0.3671756  0.2978996
## 2 0.3341307   0.3672231  0.2986462
## 3 0.3333375   0.3672685  0.2993939
## 4 0.3325453   0.3673120  0.3001427
## 5 0.3317540   0.3673535  0.3008925
## 6 0.3309636   0.3673929  0.3016435
```

```r
tail(pi.hat.ord1)
```

```
##     Democrat Independent Republican
## 56 0.2927240   0.3667886  0.3404874
## 57 0.2919860   0.3667251  0.3412889
## 58 0.2912491   0.3666596  0.3420913
## 59 0.2905133   0.3665921  0.3428946
## 60 0.2897786   0.3665226  0.3436988
```

```
## 61 0.2890450   0.3664511  0.3445039
```

```
plot.new()
x <- 20:80
plot(x, pi.hat.ord1[, 1], type = "l", col = "blue", ylim = range(min(pi.hat.ord1),
    max(pi.hat.ord1)), xlab = "Age", ylab = "Predicted probability")
lines(x, pi.hat.ord1[, 2], col = "gray")
lines(x, pi.hat.ord1[, 3], col = "red")
```



This predicted probability chart looks very different from the one generated from the multinomial model! In the chart generated by the multinomial model, age has no impact on whether a white male is a Democrat, whereas in this chart, age has no impact on whether a white male is Independent. The models are generating very different results!