Unit 4 Live Session (Solutions)

Discrete Response Model Part 4

Figure 1: South Hall

Class Announcements

- No HW this week
- Lab-1 due in 2 weeks

Roadmap

Rearview Mirror

- Discusses how to estimate and make inferences about a Logistic Regression Model

Today

- Multinomial probability distribution
- IJ contingency tables and inference using contingency tables
- The notion of independence
- Multinomial logistic regression models
- Odds ratios in the context of multinomial regression models
- Ordinal logistical regression model
- Estimation and statistical inference of these models

Looking Ahead

- Poisson regression model: Parameter estimation and statistical inference
- Model Comparison Criteria, Model Assessment, Goodness of Fit

Start-up Code

```
# Start with a clean R environment
rm(list = ls())

# Load libraries
## Load a set of packages inclusing: broom, cli, crayon, dbplyr , dplyr, dtplyr, forcats,
## googledrive, googlesheets4, ggplot2, haven, hms, httr, jsonlite, lubridate , magrittr,
## modelr, pillar, purrr, readr, readxl, reprex, rlang, rstudioapi, rvest, stringr, tibble,
## tidyr, xml2

# Insert the function to *tidy up* the code when they are printed out
if(!"knitr"%in%rownames(installed.packages())) {install.packages("knitr")}
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

if(!"tidyverse"%in%rownames(installed.packages())) {install.packages("tidyverse")}
library(tidyverse)

## provide useful functions to facilitate the application and interpretation of regression analysis.
if(!"car"%in%rownames(installed.packages())) {install.packages("car")}
library(car)

## provides  many functions useful for data analysis, high-level graphics, utility operations like describe()
if(!"Hmisc"%in%rownames(installed.packages())) {install.packages("Hmisc")}
library(Hmisc)

## to work with "grid" graphics
if(!"gridExtra"%in%rownames(installed.packages())) {install.packages("gridExtra")}
library(gridExtra)

## provides function to for Visualization techniques, summary and inference procedures such as assocstats()
if(!"vcd"%in%rownames(installed.packages())) {install.packages("vcd")}
library(vcd)

## for multinomial log-linear models.
if(!"nnet"%in%rownames(installed.packages())) {install.packages("nnet")}
library(nnet)
```

```r
## To use plor()
if(!"MASS"%in%rownames(installed.packages())) {install.packages("MASS")}
library(MASS)

## To generate regression results tables and plots
if(!"finalfit"%in%rownames(installed.packages())) {install.packages("finalfit")}
library(finalfit)

## To produce LaTeX code, HTML/CSS code and ASCII text for well-formatted tables
if(!"stargazer"%in%rownames(installed.packages())) {install.packages("stargazer")}
library(stargazer)

## To do hypothesis testing in ordinal regression model
if(!"ordinal"%in%rownames(installed.packages())) {install.packages("ordinal")}
library(ordinal)
```

Case Study: National Election Survey

Introduction

ANES provides data that help explain election outcomes by supporting detailed hypothesis testing, measuring multiple variables, and promoting comparisons across people, contexts, and time. (ANES)

> In this exercise, we want to study how the evaluation of President Obama depends on voters' demographic characteristics such as gender, race, and age.

Data Description

The data was obtained from the **American National Election Survey**, conducted several months before the 2016 American Presidential elections.

The dataset "*voters.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:

- Presjob: Respondents' evaluation of President Obama(Approve, Neutral, Not Approve)
- age: (Respondents' age, as of 2016)
- race_white: Dummy variable taking a value of one if the respondent is white and is zero otherwise.
- female: Dummy variable taking a value of one if the respondent is female and is zero otherwise.

Descriptive Statistics

- First, load and check the data set and discuss its structure
- Discuss missing values and how you would typically handle them at work

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

head(df) %>%
  knitr::kable()
```

party	presjob	srv_spend	age	female	race_white
Democrat	Approve	High	56	Male	White
Independent	Neutral	High	59	Female	White
Republican	Not Approve	Low	53	Male	White
Democrat	Approve	High	36	Male	White
NA	Not Approve	Low	42	Male	White
Independent	Not Approve	Low	58	Male	White

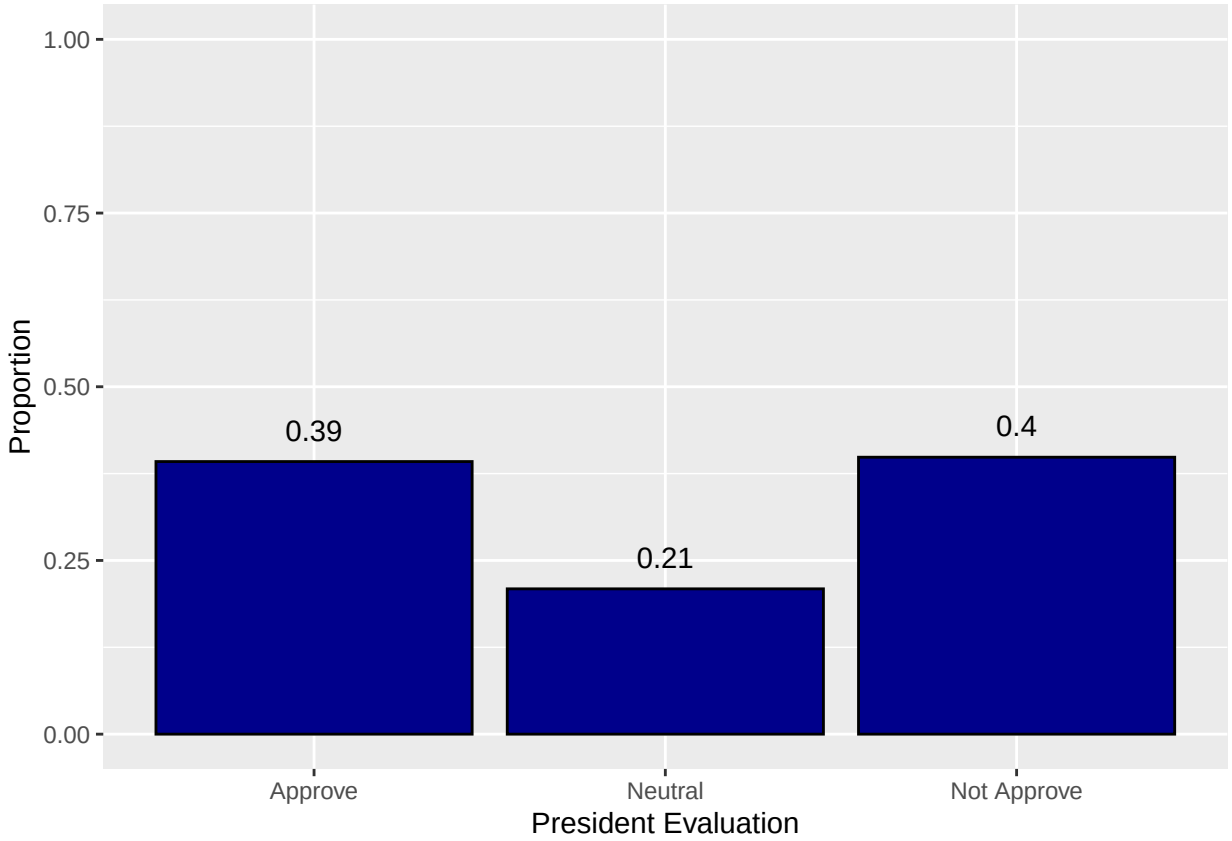
```
#str(df)
#describe(df)

## rows with NA
#df[!complete.cases(df),]
# counts NA in each column
#sapply(df, function(x) sum(is.na(x)))

# Keep only the complete cases in the dataset
df2 <- df[complete.cases(df),]
### Reorder the categories of presjob and convert female and race_white to factor

df2 <- df2 %>%
  mutate(
    race_white = factor(race_white),
    female = factor(female),
    presjob = factor(presjob)
      )
# Attach the dataset
attach(df2)
```

Univariate Analysis

- The response (or dependent) variable of interest, President evaluation, is a categorical variable with three levels.
- Use a barplot to examine the response variable. What do you learn about the President's evaluation?

```
p1<- df2 %>%
  ggplot(aes(x= presjob, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=round(..prop..,2)), vjust=-1) +
  xlab("President Evaluation") +
  ylab("Proportion") +
  ylim(0,1)

p1
```

```
## Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(prop)` instead.
```

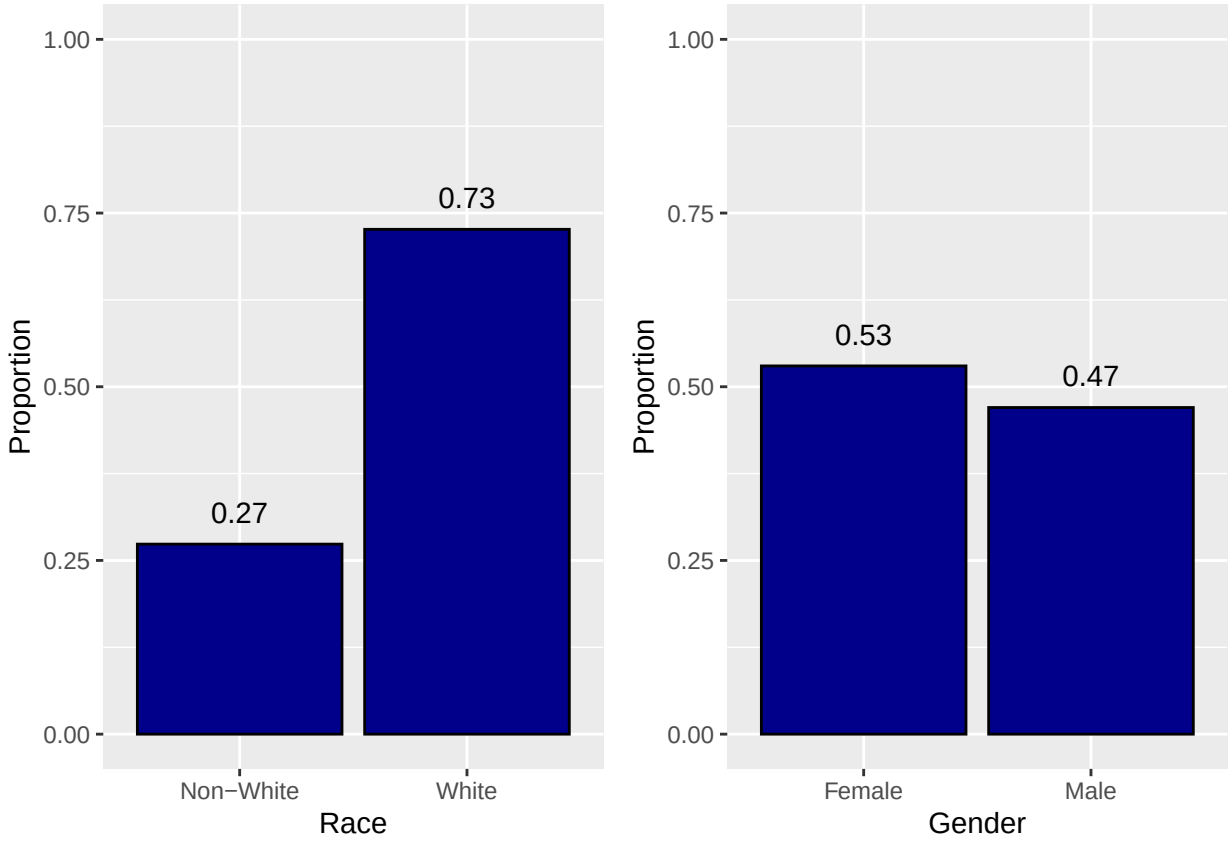


From the bar plot, 39% of subjects in our sample approved of the president's job, 40% did not approve of the president's job, and 21% were neutral.

- Use the following command to produce a barplot for race and gender. What do you discover?

```
p2<- df2 %>%
  ggplot(aes(x= race_white, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=round(..prop..,2)), vjust=-1) +
  xlab("Race") +
  ylab("Proportion") +
  ylim(0,1)
```

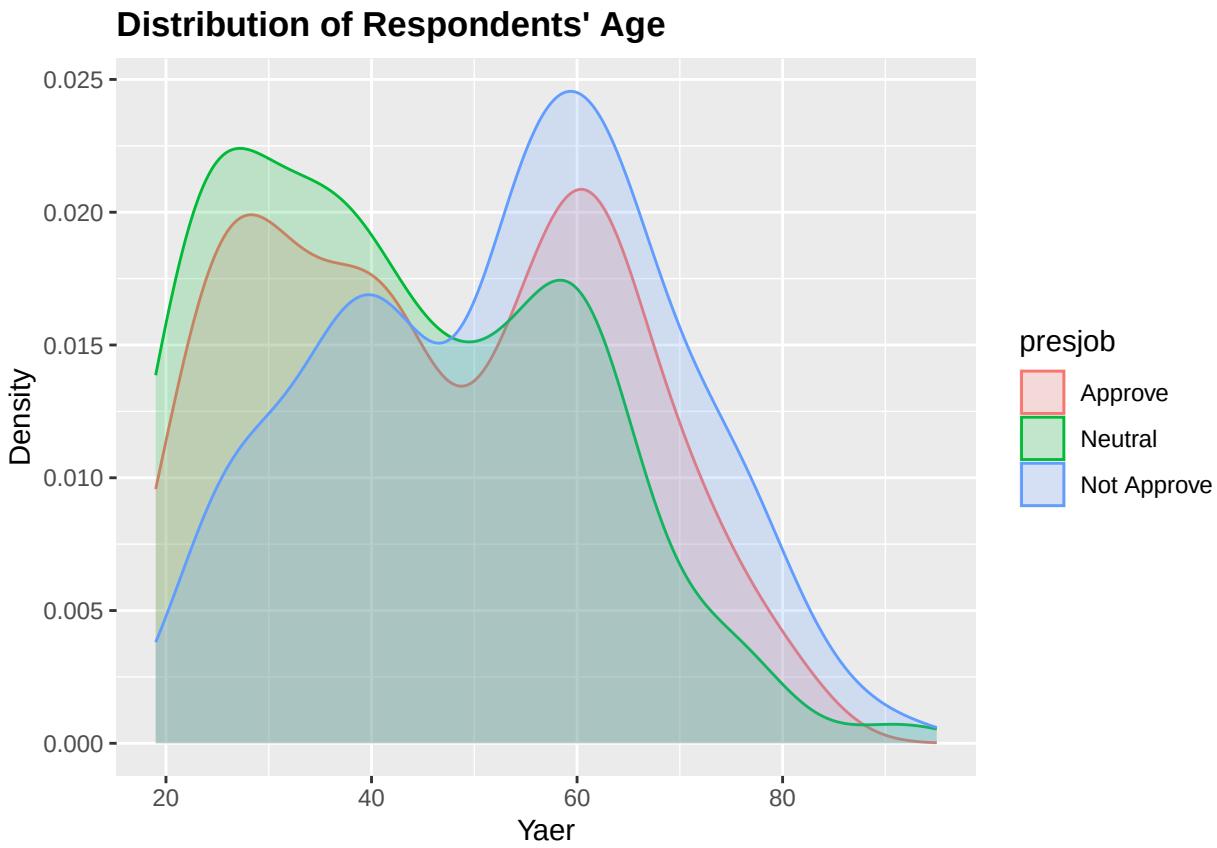
```
p3<- df2 %>%
  ggplot(aes(x= female, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=round(..prop..,2)), vjust=-1) +
  xlab("Gender") +
  ylab("Proportion") +
  ylim(0,1)

grid.arrange(p2, p3, nrow = 1, ncol = 2)
```



The majority of respondents in this sample are white and female.

- Use the following command to create a density plot of age. What do you think about the distribution of age?

```
p4 <- df2 %>%
  ggplot(aes(x = age)) +
  geom_density(aes(y = ..density.., color = presjob, fill = presjob),alpha=0.2) +
  ggtitle("Distribution of Respondents' Age") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Yaer") +
  ylab("Density")

p4
```

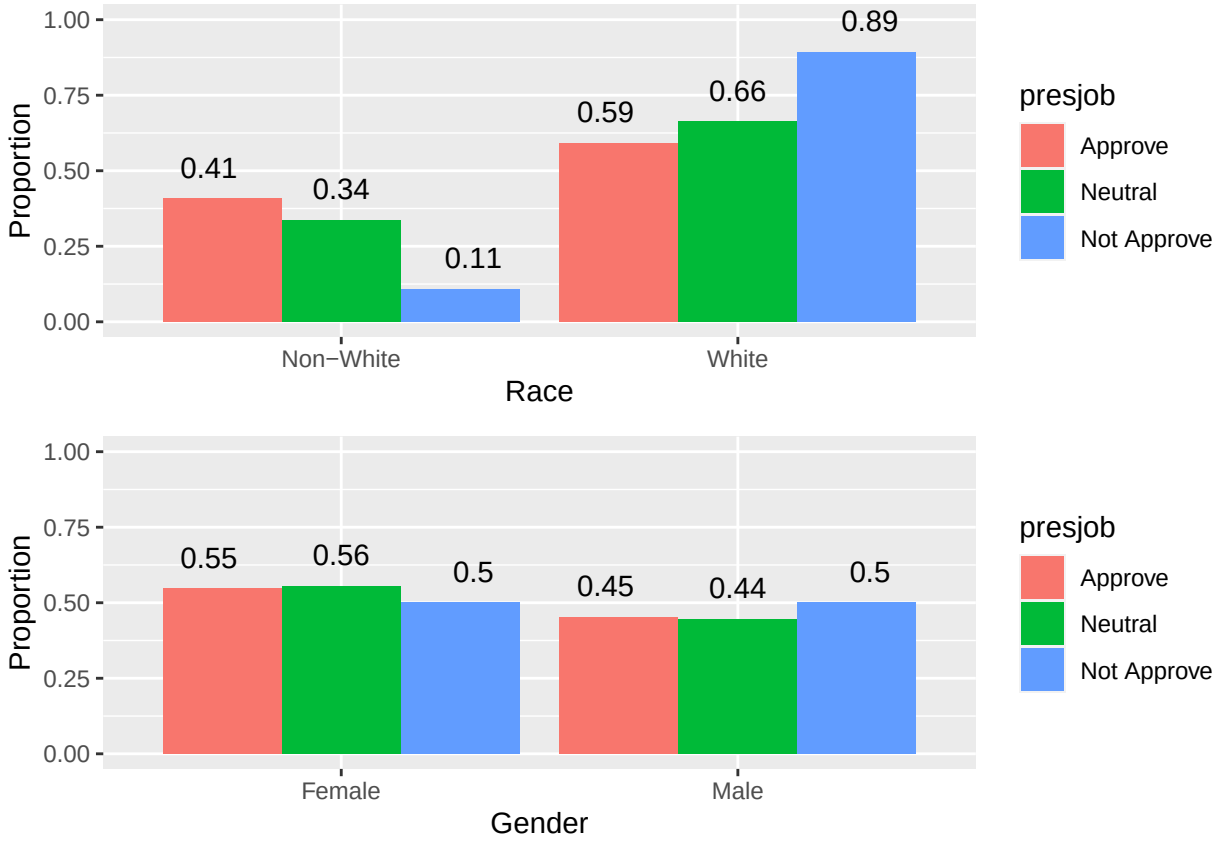


The distributions of age are different, and it seems that people who disapproved of president jobs are older on average, and people who were neutral are younger on average.

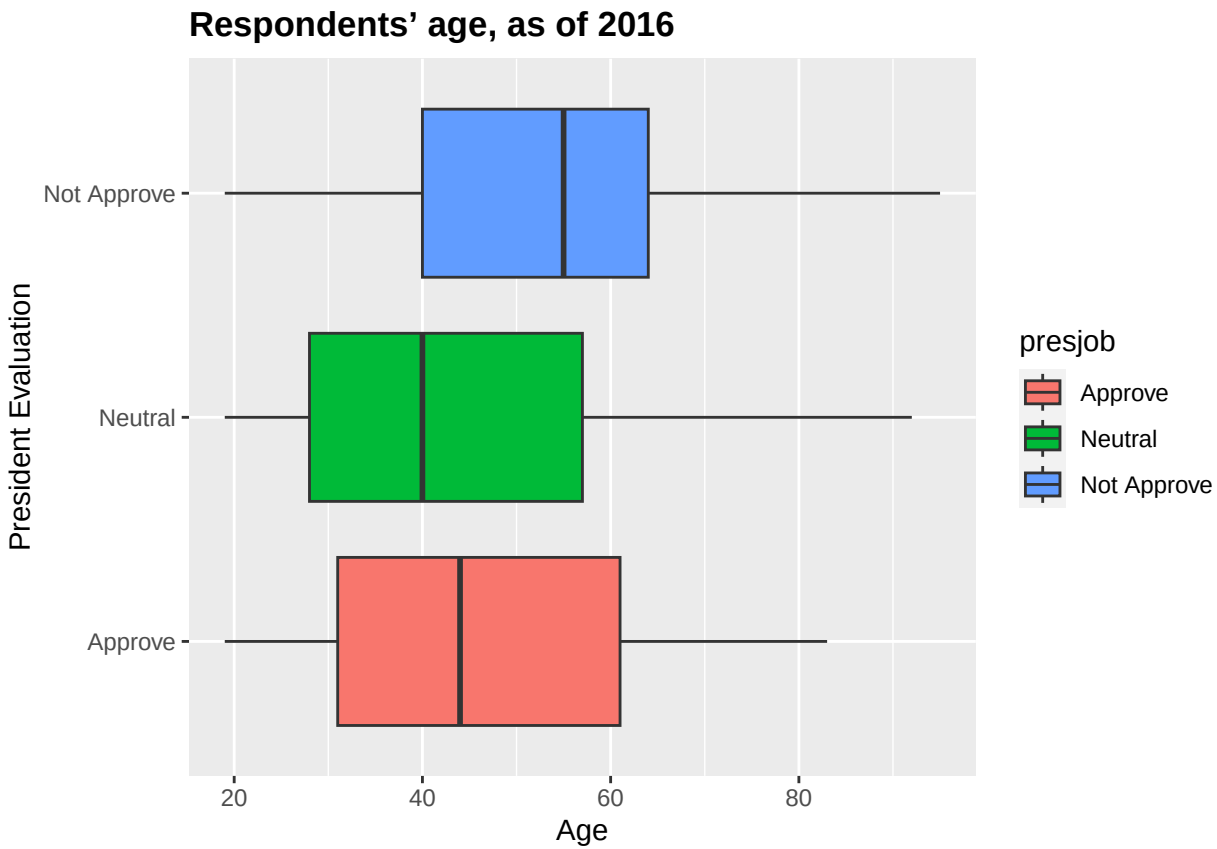
Bivariate Analysis

- Examine the simple associations between the president's evaluation and each explanatory variable. How are these variables correlated?

```
p5 <- df2 %>%
  ggplot(aes(x=race_white,
             y = ..prop..,
             group = presjob,
             fill = presjob)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=round(..prop..,2)),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("Race") +
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "presjob")

p6 <- df2 %>%
  ggplot(aes(x=female,
             y = ..prop..,
             group = presjob,
             fill = presjob)) +
  geom_bar( position = 'dodge') +
  geom_text(stat='count',
            aes(label=round(..prop..,2)),
            vjust=-1,
            position = position_dodge(width = 1)) +
  xlab("Gender") +
  ylab("Proportion") +
  ylim(0,1) +
  labs(fill = "presjob")

grid.arrange(p5, p6, nrow = 2, ncol = 1)
```




```
p7 <- df2 %>%
 ggplot(aes(presjob, age)) +
  geom_boxplot(aes(fill = presjob)) +
  coord_flip() +
  ggtitle("Respondents' age, as of 2016") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  ylab("Age") +
  xlab("President Evaluation")

p7
```



From the first bar plot above, the non-white respondents have a higher probability of approving the president's job. In comparison, white respondents have a higher probability of not supporting the president's job.

In the second barplot above, both females and males have the almost same probability of approving, not approving, or being neutral, so gender is not a proper variable to classify these categories.

From the box plot for age, the respondents who disapproved of the president's job are on average older, and respondents who are neutral are on average younger.

- Use summary_factorlist() function from the finalfit package to tabulate data. What do you learn about the relationship between the president's evaluation and these variables?

```
dependent <- "presjob"
explanatory <- c("race_white", "female", "age")
df %>%
  summary_factorlist(dependent, explanatory, add_dependent_label = TRUE) %>%
  knitr::kable()
```

Dependent: presjob		Approve	Neutral	Not Approve
race_white	Non-White	183 (40.4)	84 (32.9)	58 (11.8)
	White	270 (59.6)	171 (67.1)	434 (88.2)
female	Female	246 (54.3)	140 (54.9)	244 (49.6)
	Male	207 (45.7)	115 (45.1)	248 (50.4)
age	Mean (SD)	46.2 (16.9)	42.8 (16.2)	52.5 (16.3)

$I * J$ Contigency table and Test for Independence

- In this section, we would like to determine if the president's evaluation is related to the following variables:
- Race
- Gender

Create a contingency table and test for independence to make this determination. Recall that the hypothesis for the test is:

$H_0 : \pi_{ij} = \pi i + * \pi_{+j}$

$H_a : \pi_{ij} \neq \pi i + * \pi_{+j}$

Effectively, we construct a table of the counts of approval rating x race and approval rating x gender and test whether the rows are independent of the columns i.e. approval rating is independent of race and gender.

The chi-squared test compares observed counts in the table to expected counts if the two attributes were independent, which in the particular form of $\Sigma_i (O_i - E_i)^2 / E_i$ follows a chi-squared distribution under the null hypothesis.

Note that one downside to this independence testing is that it examines relationships between variables without controlling for other variables as we would in a regression model. This means it is not as rigorous and could miss out on more nuanced relationships. Independence testing is good for exploratory analysis, but it should be confirmed with regression modeling as below.

```
## president evaluation and race
tab1 <- df2 %>%
  group_by(race_white,presjob) %>%
  count() %>%
  xtabs(formula = n ~ race_white + presjob)

tab1
```

```
##            presjob
## race_white  Approve Neutral Not Approve
##   Non-White     179      79          48
##   White         260     155         398
```

```
## chi-square test
test1 <- chisq.test(x = tab1, correct = FALSE)
```

Based on the p-value of less than 5%, we reject the null hypothesis that the president's job and race are independent, and there is strong evidence of dependence

```
## president evaluation and gender
tab2 <- df2 %>%
  group_by(female,presjob) %>%
```

```
  count() %>%
  xtabs(formula = n ~ female + presjob)

tab2
```

```
##         presjob
## female   Approve Neutral Not Approve
##   Female     240     130         223
##   Male       199     104         223
```

```
## chi-square test
test2 <- chisq.test(x = tab2, correct = FALSE)

## other useful function to test the independence
#summary(tab2)
#CrossTable(df2$female, df2$presjob, digits = 2, prop.c = FALSE, prop.t = FALSE, chisq = TRUE)
#assocstats(tab2)
```

The p-value is greater than 0.05, which supports the null hypothesis of independence, and we fail to reject it.

Multinomial logistic regression model

- In this part, we use regression modeling to assess the association between the variables more systematically.
- Estimate the following model and interpret the results. Note that here j indexes the level of Y that we are talking about (Approve, Neutral, or Not Approve).

We set our reference $j = 1$ to Approve and use multinomial logistic regression to model the log odds of other values of j i.e. $j = 2$ or Neutral and $j = 3$ or Not Approve each to the reference of Approve. This results in two separate sets of coefficients for each log odds ratio.

$$log(\frac{\pi_j}{\pi_1}) = \beta_{j0} + \beta_{j1} race + \beta_{j2} female + \beta_{j3} age + u$$

```
mod.nomial <- multinom(formula = presjob ~ race_white + female + age, data = df2)
```

```
## # weights:  15 (8 variable)
## initial  value 1229.347151
## iter  10 value 1110.721234
## final  value 1105.086007
## converged
```

```
summary(mod.nomial)
```

```
## Call:
## multinom(formula = presjob ~ race_white + female + age, data = df2)
##
## Coefficients:
##             (Intercept) race_whiteWhite femaleMale         age
## Neutral     -0.09874428       0.3688984 -0.0591949 -0.01648380
## Not Approve -2.20367178       1.6544530  0.1969974  0.01728407
##
## Std. Errors:
##             (Intercept) race_whiteWhite femaleMale         age
## Neutral       0.2615224       0.1712941  0.1640140 0.005063962
## Not Approve   0.2708726       0.1827794  0.1427282 0.004288010
##
## Residual Deviance: 2210.172
## AIC: 2226.172
```

The model summary output has a block of coefficients and standard errors. Each block has one row of values corresponding to a model equation. Focusing on the block of coefficients, we can look at the first row comparing presjob= "Neutral" to our baseline, which is presjob = "Approve," and the second row comparing presjob= "Not Approve" to our baseline presjob = "Approve"

The log odds of being neutral vs. approve will increase by 0.37 if race change from non-withe to white. intuitively, higher probability of being neutral v.s approve among white respondents

A one-year increase in age is associated with the decrease in the log odds of being in neutral vs. approve by 0.16. A one-year increase in age increases the probability of being approve vs. neutral

In the second row, the log odds of not approving vs. approve will increase by 1.56 if race changes from not-withe to with, which means a higher probability of not approve among the white respondents

A one-year increase in age is associated with the rise in the log odds of not approve vs. approve. in the amount of 0.017 .

- Perform a Wald test to assess if the coefficients are statistically significant. Note that r indexes the predictor variable we are talking about (race, age, or gender). The null hypothesis we test here refers to selecting a particular variable in one of the log odds ratios to be zero.

$H_0 : \beta_{jr} = 0$

$H_a : \beta_{jr} \neq 0$

```
z_score <- summary(mod.nomial)$coefficients/summary(mod.nomial)$standard.errors

z_score
```

```
##             (Intercept) race_whiteWhite femaleMale       age
## Neutral      -0.3775748        2.153597 -0.3609137 -3.255119
## Not Approve  -8.1354550        9.051636  1.3802268  4.030791
```

```
# compute the p-values for a 2-tailed z test
p <- 2 * pnorm(abs(z_score), lower.tail = FALSE)
p
```

```
##              (Intercept) race_whiteWhite femaleMale          age
## Neutral     7.057465e-01    3.127181e-02  0.7181640 1.133447e-03
## Not Approve 4.103940e-16    1.408421e-19  0.1675168 5.558936e-05
```

Based on the p-values, coefficients of age and race are statistically significant, and gender is insignificant in both Neutral vs. approve and Not Approve vs. approve models

- Perform LRT using Anova() to explore if a given explanatory variable x_r is statistically significant over all response categories. Note that r indexes the predictor variable we are talking about (race, age, or gender). The null hypothesis we test here refers to inclusion of a predictor in the model or not, which means it involves testing whether all coefficients in the odds ratios for a particular predictor are zero or not.

$H_0 : \beta_{2r} = \beta_{3r} = 0$

$H_a : \beta_{2r} \neq 0 \quad or \quad \beta_{3r} \neq 0$

```
Anova(mod.nomial)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: presjob
##            LR Chisq Df Pr(>Chisq)
## race_white   97.555  2  < 2.2e-16 ***
## female        2.902  2     0.2343
## age          46.645  2  7.434e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the large p-value of the female coefficient, we fail to reject the null hypothesis that the female coefficient is zero. Hence, insufficient evidence indicates that the gender variable is not essential for overall response categories.

The small p-values indicate that age and race are statistically significant and affect the probability of approve, neutral, and not approve.

- Estimate and interpret the following odd ratios in the estimated model for each explanatory variable ($\widehat{OR} = exp(\beta_{jr})$)
  - Approve vs. Neutral
  - Approve vs. Not Approve

```
#Approve vs. Neutral
beta.hat2 <- coefficients(mod.nomial)[1,2:4]
round(exp(beta.hat2),2)
```

```
## race_whiteWhite      femaleMale             age
##            1.45            0.94            0.98
```

```
#Approve vs. Not Approve
beta.hat3 <- coefficients(mod.nomial)[2,2:4]
round(exp(beta.hat3),2)
```

```
## race_whiteWhite      femaleMale             age
##            5.23            1.22            1.02
```

The estimated odds of a neutral vs. approve change by 1.45 times switching from non-white to white, holding the other variables constant.

The estimated odds of a neutral vs. approve for a one-year increase in age change by 0.98 times, holding the other variables constant. This estimated odd of age is not practically significant

The estimated odds of a not approve vs. approve change by 5.23 times switching from non-white to white, holding the other variables constant.

- Construct confidence intervals for the odds ratios using the confint() function.

```
## compute CI for the coefficients
conf.beta <- confint(object = mod.nomial, level = 0.95)

## construct CI for OR
exp(conf.beta)
```

```
## , , Neutral
##
##                       2.5 %    97.5 %
## (Intercept)     0.5426348 1.5126004
## race_whiteWhite 1.0337243 2.0230952
## femaleMale      0.6834126 1.2998731
## age             0.9739367 0.9934628
##
## , , Not Approve
##
##                        2.5 %    97.5 %
## (Intercept)     0.06492177 0.1877261
## race_whiteWhite 3.65542314 7.4834514
## femaleMale      0.92058598 1.6108139
## age             1.00891927 1.0260212
```

With 95% confidence, the odds of neutral instead of approve change by 1.03 to 2.02 times when race changes from non-white to white, holding the other variables constant.

With 95% confidence, the odds of not approve instead of approve change by 3.65 to 7.48 times when race changes from non-white to white, holding the other variables constant. The probability of not approving Obama's performance is higher among white respondents.

Ordinal response regression models (1) (classic formulation)

- The high approval rate before the election is a critical factor in any election. So the 'Not approve" is less desirable than "Neutral" and "Approve."
- Estimate and interpret following model using the ordering of "Not Approve"(Y=1) < "Neutral" (Y=2) < "Approve" (Y=3).

$$logit(P(Y \leq j)) = \beta_{j0} + \beta_{j1} race_{white} + \beta_{j2} female + \beta_{j3} age + u$$

Based on the this definition, our two cases are the following (since $P(Y > 3)$ is not defined):

$$\frac{P(Y \leq 2)}{P(Y > 2)} = \frac{P(Y = not\ approve\ or\ neutral)}{P(Y = approve)}$$

$$\frac{P(Y \leq 1)}{P(Y > 1)} = \frac{P(Y = not\ approve)}{P(Y = neutral\ or\ approve)}$$

Ordinal regression also has another assumption known as the proportional odds assumption which means that the coefficients we fit in the model will apply to both odds ratios above. Effectively, we assume the impact of race, age, and gender in our context is the same for both odds ratios and only return 1 set of values for each coefficient. Note the difference from multinomial logistic regression that had different coefficients for each odds ratio.

The proportional odds assumption if satisfied results is a higher power model, which makes ordinal regression better than multinomial regression if it is true.

**Note that because of notational differences in polr() and the model above, we need to reverse the sign of all the estimated coefficients before analysis below.

```
## create required ordering
df2$presjob.order <- factor(presjob, levels = c("Not Approve", "Neutral", "Approve"))
attach(df2)
```

```
## The following objects are masked from df2 (pos = 3):
##
##     age, female, party, presjob, race_white, srv_spend
```

```
levels(presjob.order)
```

```
## [1] "Not Approve" "Neutral"     "Approve"
```

```
### estimate the model
mod.ord <- polr(formula = presjob.order ~ race_white + female + age, data = df2, method = "logistic")
summary(mod.ord)
```

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

```
## Call:
## polr(formula = presjob.order ~ race_white + female + age, data = df2,
##     method = "logistic")
##
## Coefficients:
##                     Value Std. Error t value
## race_whiteWhite -1.19144    0.13199  -9.027
## femaleMale      -0.15800    0.11491  -1.375
## age             -0.01356    0.00342  -3.966
##
## Intercepts:
##                      Value    Std. Error t value
## Not Approve|Neutral  -2.0242   0.2018   -10.0293
## Neutral|Approve      -1.0949   0.1950    -5.6140
##
## Residual Deviance: 2256.851
## AIC: 2266.851
```

In this model, we use log-odds of cumulative probabilities instead of log-odds of probabilities.

The model assumes that the effects of the explanatory variables are the same regardless of which cumulative probabilities are used to form the log odds. This is known as the proportional odds assumption.

- How could we perform a Wald test to assess if the coefficients are statistically significant?

```
## store table
ctable <- coef(summary(mod.ord))
```

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

```
## calculate and store p values
p <- 2 * pnorm(abs(ctable[, "t value"]), lower.tail = FALSE)

## combined table
ctable <- cbind(ctable, "p value" = p)
ctable
```

```
##                         Value  Std. Error    t value      p value
## race_whiteWhite     -1.19143631 0.131986008  -9.026990 1.764593e-19
```

```
## femaleMale          -0.15799513 0.114906045  -1.374994 1.691333e-01
## age                 -0.01356396 0.003419798  -3.966304 7.299582e-05
## Not Approve|Neutral -2.02416421 0.201825656 -10.029271 1.133503e-23
## Neutral|Approve     -1.09485527 0.195023422  -5.613968 1.977387e-08
```

All coefficients except for gender are statistically significant, indicating age and race are correlated with the president's approval rate.

- Construct and interpret the odds ratios for each explanatory variable ($\widehat{OR} = exp(\beta_{jr})$).

```
round(exp(-mod.ord$coefficients),2)
```

```
## race_whiteWhite      femaleMale             age
##            3.29            1.17            1.01
```

The estimated odds of not approve vs. neutral or apporve is 3.29 times as large for whites compared to non-whites, holding the other variables constant.

The estimated odds of not approve vs. neutral or approve change by 1.01 times for a year increase in the age, holding the other variables constant. But it's not practically significant

Because of the proportional odds assumption here, each of these statements also applies to the odds of a not approve or neutral vs. approve.

For this reason, it is common to interpret odds ratios, such as for race, by saying: The estimated odds of president approval being below a particular level change by 3.29 times for changing from non-white to the white, holding the other variables constant.

- Construct LR confidence interval for the odds ratios.

```
# first compute the coefficients CI
coef.beta <- confint(object = mod.ord, level = 0.95)
```

```
## Waiting for profiling to be done...
```

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

```
## construct OR CI
ci <- exp((-coef.beta))

round(data.frame(lwr = ci[,2], upr = ci[,1]),2)
```

```
##                   lwr  upr
## race_whiteWhite 2.55 4.27
## femaleMale      0.94 1.47
```

```
## age             1.01 1.02
```

With 95% confidence, the odds of president approval being below a particular level change by 2.55 to 4.27 times for white respondents v.s non-white respondents, holding the other variables constant.

Ordinal response regression models (2) (alternative formulation)

- If the idea of flipping the negative sign in the R output / interpreting the cumulative probabilities of being less than or equal to a category is confusing, we can instead use the R output directly and just flip our odds ratio numerator and denominator. We flip the odds ratio since a negative in the context of exponentiation flips the numerator and denominator.

Here we define η as the coefficient of the R output so that $\beta = -\eta$ in the model above:

$$logit(P(Y > j)) = \eta_{j0} - \eta_{j1} race_{white} - \eta_{j2} female - \eta_{j3} age + u$$

Based on the this definition, our two cases are the following (since $P(Y > 3)$ is not defined):

$$\frac{P(Y > 2)}{P(Y \leq 2)} = \frac{P(Y = approve)}{P(Y = not\ approve\ or\ neutral)}$$

$$\frac{P(Y > 1)}{P(Y \leq 1)} = \frac{P(Y = neutral\ or\ approve)}{P(Y = not\ approve)}$$

Odds ratios are found by direct exponentiation of the coefficients from the R output:

```
round(exp(mod.ord$coefficients), 2)
```

```
## race_whiteWhite      femaleMale             age
##            0.30            0.85            0.99
```

For example, White voters increase the odds of approve compared to not approve or netural by 0.3 times compared to non white voters.

And confidence intervals are found by directly using confint and exponentiating the resulting bounds on the log scale.

```
coef.beta <- confint(object = mod.ord, level = 0.95)
```

```
## Waiting for profiling to be done...
```

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

```
ci <- exp((coef.beta))
round(data.frame(lwr = ci[, 1], upr = ci[, 2]), 2)
```

```
##                   lwr  upr
## race_whiteWhite 0.23 0.39
## femaleMale      0.68 1.07
## age             0.98 0.99
```

For example, with 95% confidence we believe that White voters increase the odds of approve compared to not approve or neutral by 0.2 to 0.4 times compared to non white voters.

Testing Proportional Odds

- There are two ways we can test the proportional odds assumption in ordinal regression.
- The first uses a likelihood ratio test that compares the likelihood function of a null hypothesis where we assume the proportional odds assumption for a given variable and an alternative hypothesis of relaxing the proportional odds assumption for just that variable.
- If we reject the null hypothesis, then we reject the proportional odds assumption for that variable.

Run this test using the ordinal package and clm() function to fit the ordinal logistic regression function (this is the same as polr but is from a different package).

The nominal_test() function runs the LRT for relaxing proprotional odds of each variable in the model and is similar to Anova().

We see that we reject the null hypothesis of proportional odds for both race and age but not for gender. The ordinal logistic regression as constructed is not valid for this data set.

```
#refit prop odds model using ordinal package to allow use of nominal_test function
mod.ord2 <- clm(presjob.order ~ race_white + female + age, data =df2)

nominal_test(mod.ord2)
```

```
## Tests of nominal effects
##
## formula: presjob.order ~ race_white + female + age
##            Df  logLik    AIC     LRT  Pr(>Chi)
## <none>        -1128.4 2266.8
## race_white  1 -1121.2 2254.3 14.5333 0.0001377 ***
## female      1 -1128.0 2268.0  0.8559 0.3548754
## age         1 -1112.7 2237.3 31.5228 1.971e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also test the proportional odds assumption, roughly speaking, by using the results from the multinomial model fit. The proportional odds assumption implies that the coefficients for the same variable are equal across regression models i.e. the coefficient for race_white in not approve vs. approve is the same as the coefficient for race_white in neutral vs. not approve

We can compute the test statistic for this using the Wald form and the variance covariance matrix of the model fit. We can see that the results are similar to using the nomial_test function above.

```
#can do hypothesis test for equality of coefficients across categories for race and/or age
coef(mod.nomial)
```

```
##             (Intercept) race_whiteWhite femaleMale         age
## Neutral     -0.09874428       0.3688984 -0.0591949 -0.01648380
## Not Approve -2.20367178       1.6544530  0.1969974  0.01728407
```

```
diff.race <- abs((coef(mod.nomial)[1,2] - coef(mod.nomial)[2,2]))
diff.race.se <- sqrt(vcov(mod.nomial)[2,2] + vcov(mod.nomial)[6,6] - 2*vcov(mod.nomial)[2,6])
diff.race / diff.race.se
```

```
## [1] 6.122508
```

```
diff.age <- abs((coef(mod.nomial)[1,4] - coef(mod.nomial)[2,4]))
diff.age.se <- sqrt(vcov(mod.nomial)[4,4] + vcov(mod.nomial)[8,8] - 2*vcov(mod.nomial)[4,8])
diff.age / diff.age.se
```

```
## [1] 6.524581
```

```
diff.gender <- abs((coef(mod.nomial)[1,3] - coef(mod.nomial)[2,3]))
diff.gender.se <- sqrt(vcov(mod.nomial)[3,3] + vcov(mod.nomial)[7,7] - 2*vcov(mod.nomial)[3,7])
diff.gender / diff.gender.se
```

```
## [1] 1.516264
```

Note that the proportional odds assumption implies that across categories in the multionomial model we have parallel logit regression lines since the coefficients are the same with different intercepts. Violating the proportional odds assumption means that the different logistic regression lines are not parallel.

Reminders

1. Complete all videos and reading for unit 5