# Unit 4 Live Session

## 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
- Nominal response models
- Odds ratios in the context of nominal response 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

```
# Insert the function to *tidy up* the code when they are printed out
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)


# 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
library(tidyverse)
## provide useful functions to facilitate the application and interpretation of regression analysis.
library(car)
## provides  many functions useful for data analysis, high-level graphics, utility operations like describe()
library(Hmisc)
## to work with "grid" graphics
library(gridExtra)
## provides function to for Visualization techniques, summary and inference procedures such as assocstats()
library(vcd)
## for multinomial log-linear models.
library(nnet)
## To use plor()
library(MASS)
## To generate regression results tables and plots
library(finalfit)
## To produce LaTeX code, HTML/CSS code and ASCII text for well-formatted tables
library(stargazer)
```

## 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**, which conducted a survey several months prior to 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 |

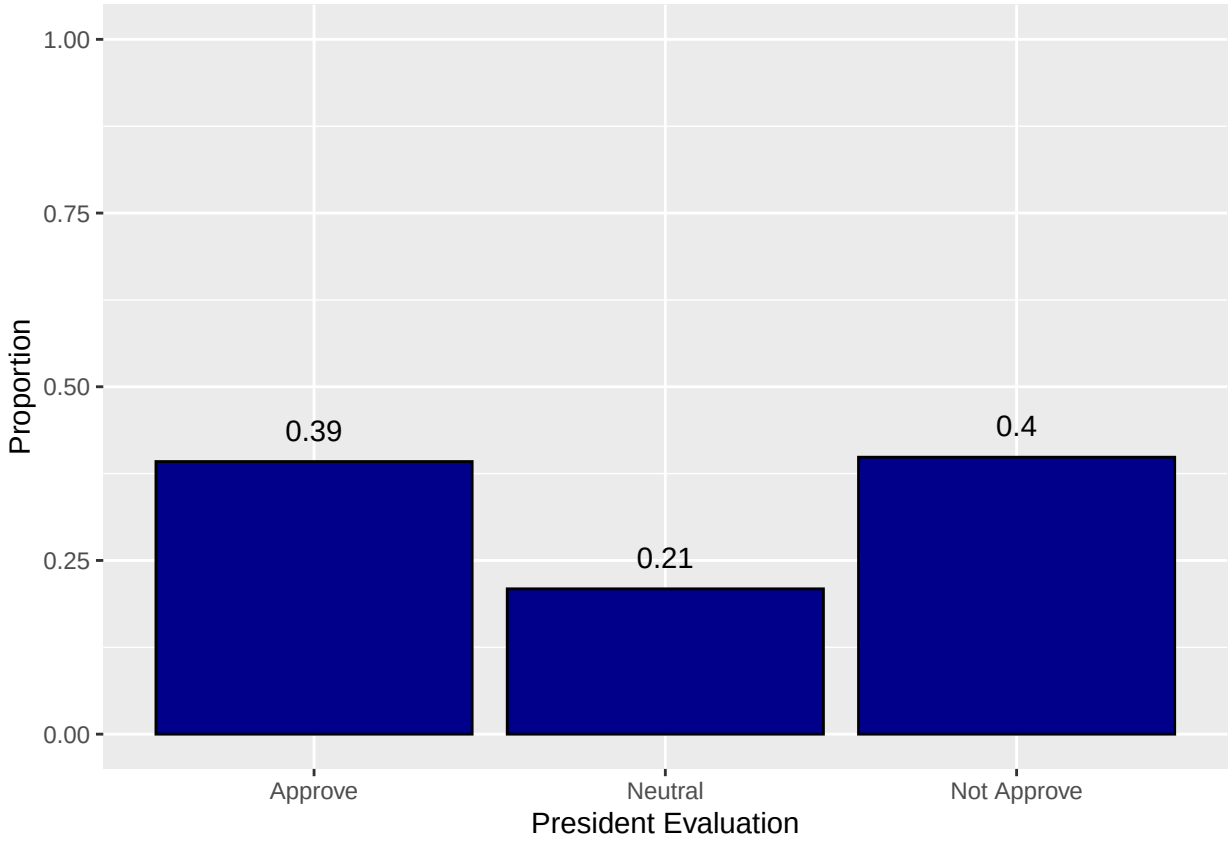
```
#str(df)
#describe(df)

## 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
```

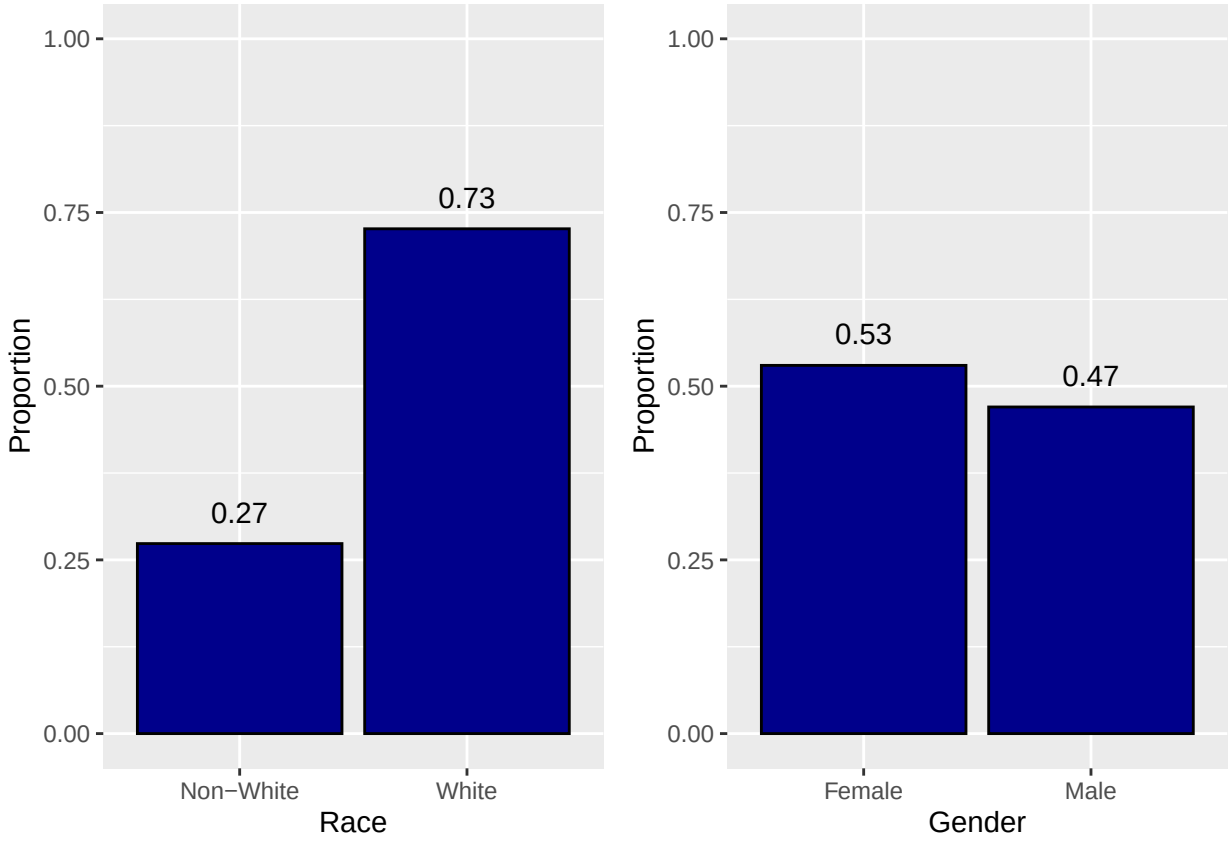
```
## Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(prop)` instead.
```




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

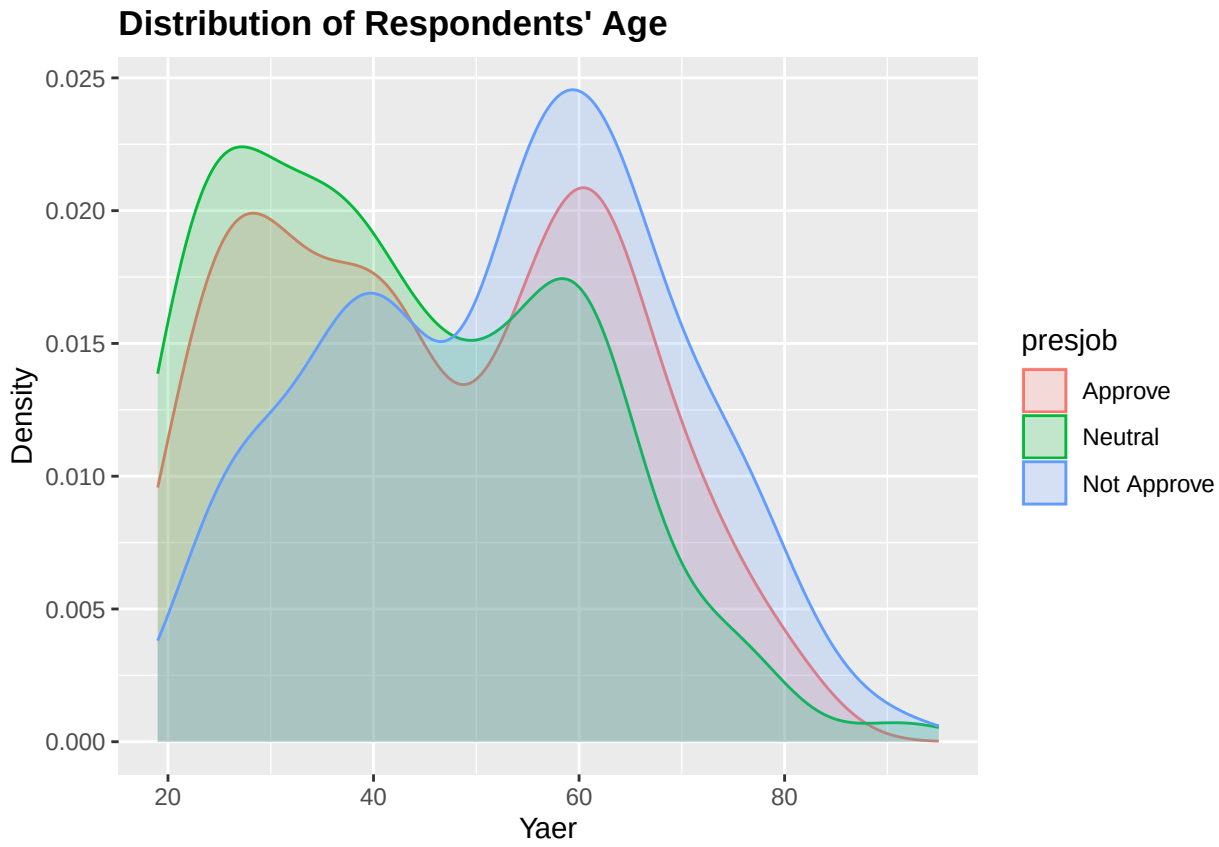
```
p2<- df2 %>%
  ggplot(aes(x= race_white, y = ..prop.., group = 1)) +
  geom_bar(fill = 'DarkBlue', color = 'black') +
  geom_text(stat='count', aes(label=round(..prop..,2)), vjust=-1) +
  xlab("Race") +
  ylab("Proportion") +
  ylim(0,1)

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)
```



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

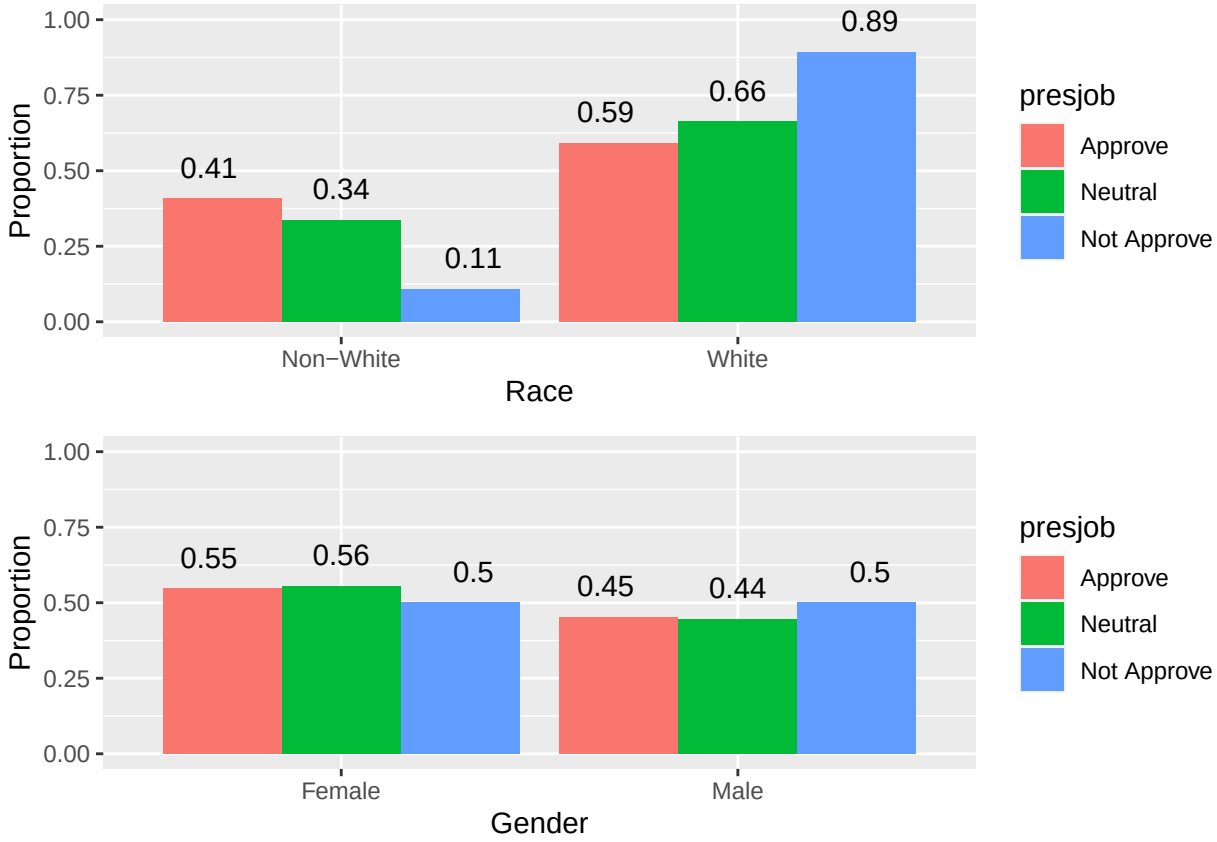
```
p4 <- df2 %>%
  ggplot(aes(x = age)) +
  geom_density(aes(y = ..density.., color = presjob, fill = presjob),alpha=0.2) +
  ggtitle("Distribution of Respondents' Age") +
  theme(plot.title = element_text(lineheight=1, face="bold")) +
  xlab("Yaer") +
  ylab("Density")

p4
```



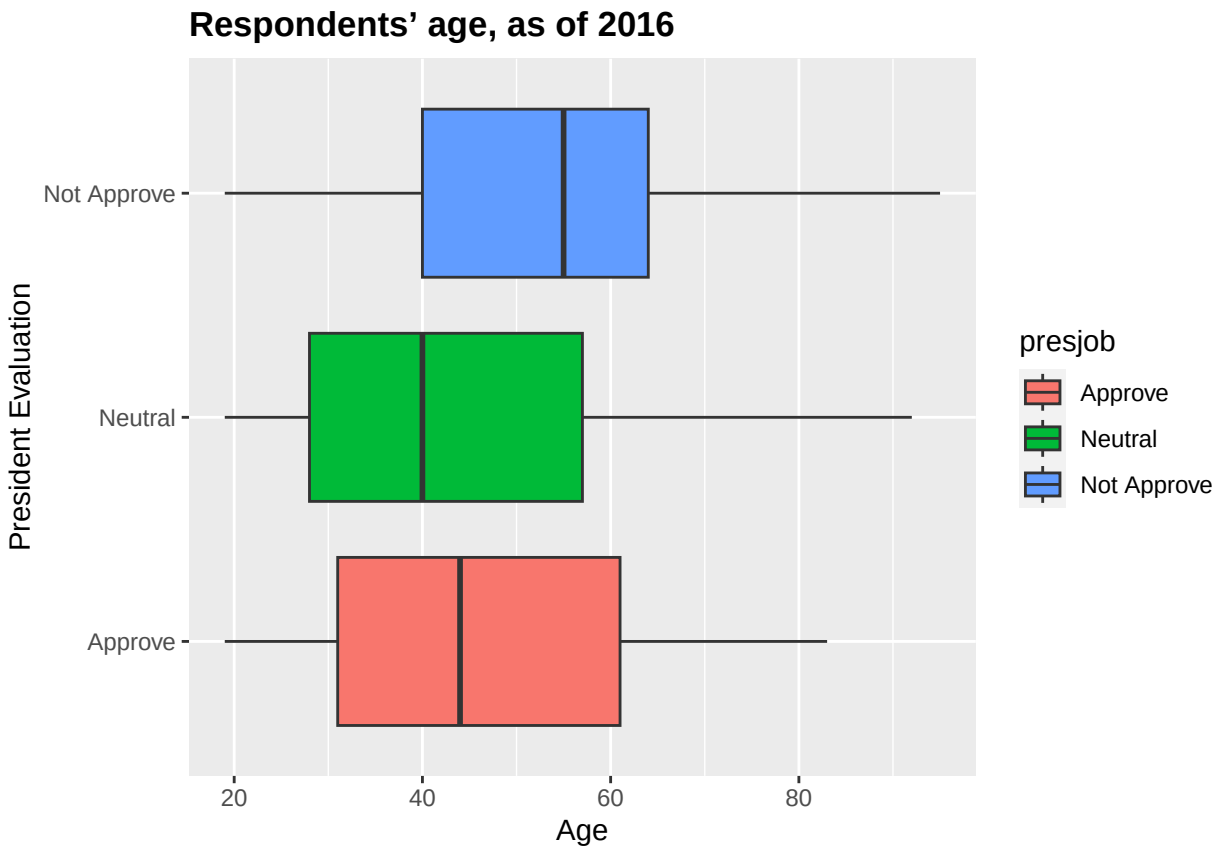
**Bivariate Analysis**

- 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
```


Respondents' age, as of 2016
President Evaluation
Not Approve
Neutral
Approve
20
40
60
80
Age
presjob
Approve
Neutral
Not Approve

- 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) |

### Model Development

### $I * J$ 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:

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

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

- When independence is rejected, examine the association between the president's evaluation and the explanatory variable using the model residual.

```
## president evaluation and race

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

## chi-square test
#test1<- # uncomment and replace with your code
#test1$stdres
```

```
## 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
```

```r
## chi-square test
#test2 <- # uncomment and replace with your code
#test2$stdres  # uncomment
```

**Nominal response 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.

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

```
#mod.nomial <- # uncomment and replace with your code
#summary(mod.nomial)
```

- Perform a Wald test to assess if the coefficients are statistically significant.

$H_0 : \beta_{jr} = 0$

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

```
#z_score <- # uncomment and replace with your code
#z_score # uncomment
```

- Perform LRT to explore if a given explanatory variable $x_r$ is statistically significant over all response categories.

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

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

```
# Replace with your code
```

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

```
#Approve vs. Neutral

# Replace with your code

#Approve vs. Not Approve

# Replace with your code
```

- Construct Wald confidence intervals for the odds ratios.

```
## compute CI for the coefficients

# Replace with your code

## construct CI for OR

# Replace with your code
```

**Ordinal response regression models**

- 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).

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

```
## 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 <- # uncomment and replace with your code
#summary(mod.ord) # uncomment
```

- How could we perform a Wald test to assess if the coefficients are statistically significant?
- Construct and interpret the odds ratios for each explanatory variable ($\widehat{OR} = exp(c * \beta_{jr})$).

```
# Replace with your code
```

- Construct LR confidence interval for the odds ratios.

```
# first compute the coefficients CI

# Replace with your code

## construct OR CI

# Replace with your code
```

## Reminders

1. Complete all videos and reading for unit 5