Lab 1, Short Question

Contents


```
library(tidyverse)
library(stargazer)
library(finalfit)
library(nnet)
library(car)
library(ordinal)
theme_set(theme_bw()) # set the theme (theme_set is built inside ggplot2)
```

1 Political ideology (30 points)

These questions are based on Question 14 of Chapter 3 of the textbook "Analysis of Categorical Data with R" by Bilder and Loughin.

> An example from Section 4.2.5 examines data from the 1991 U.S. General Social Survey that cross-classifies people according to
>
> - Political ideology: Very liberal (VL), Slightly liberal (SL),Moderate (M), Slightly conservative (SC), and Very conservative (VC)
> - Political party: Democrat (D) or Republican (R)
> - Gender: Female (F) or Male (M).
>
> Consider political ideology to be a response variable, and political party and gender to be explanatory variables. The data are available in the file pol_ideol_data.csv.

1.1 Recode Data (2 points)

Use the factor() function with the ideology variable to ensure that R places the levels of the ideology variable in the correct order.

```
pol_ideol <- read.csv("pol_ideol_data.csv",
                      header=T,na.strings=c("","NA"))
```

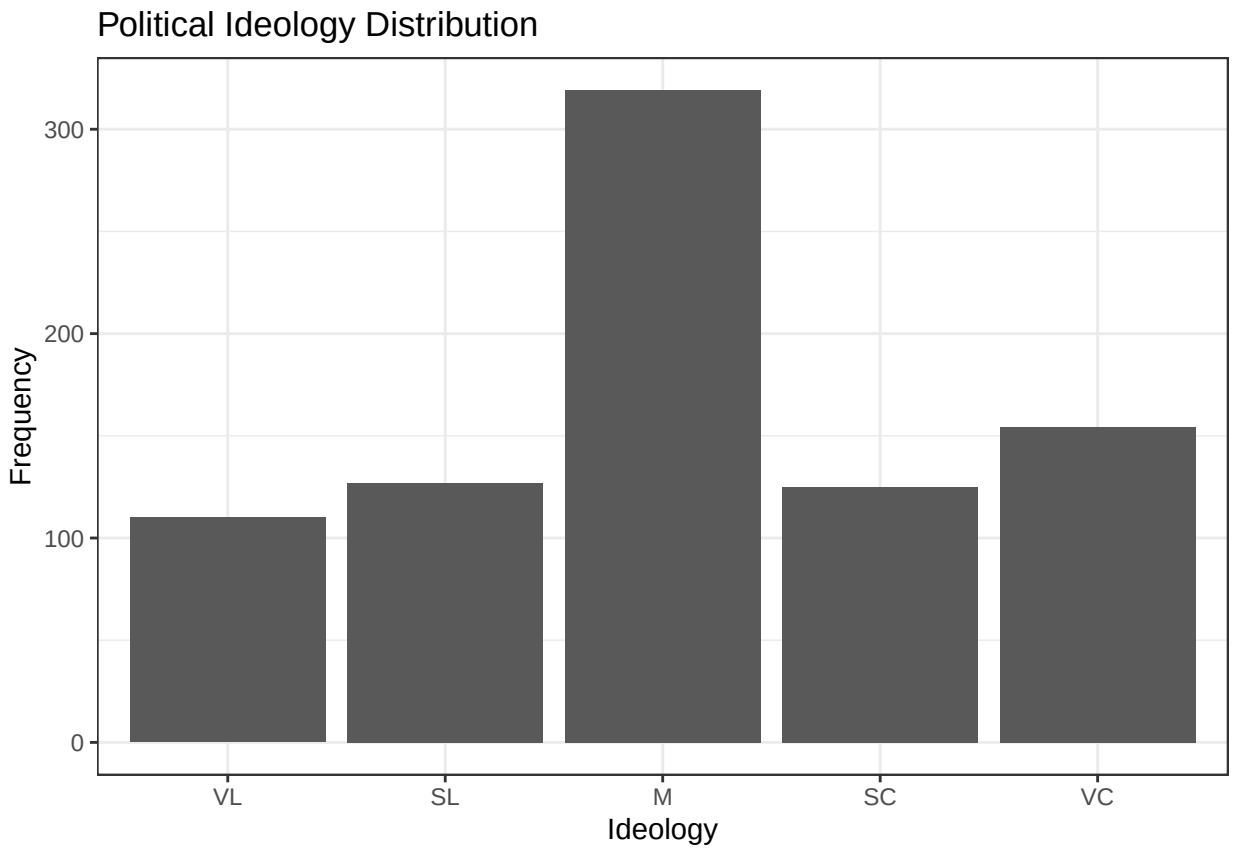
```r
pol_ideol$ideol_levels <- factor(pol_ideol$ideol, levels = c("VL", "SL", "M", "SC", "VC"))
pol_ideol$gender <- as.factor(pol_ideol$gender)
pol_ideol$party <- as.factor(pol_ideol$party)
head(pol_ideol)
```

```
##   gender party ideol count ideol_levels
## 1      F     D    VL    44           VL
## 2      F     D    SL    47           SL
## 3      F     D     M   118            M
## 4      F     D    SC    23           SC
## 5      F     D    VC    32           VC
## 6      F     R    VL    18           VL
```

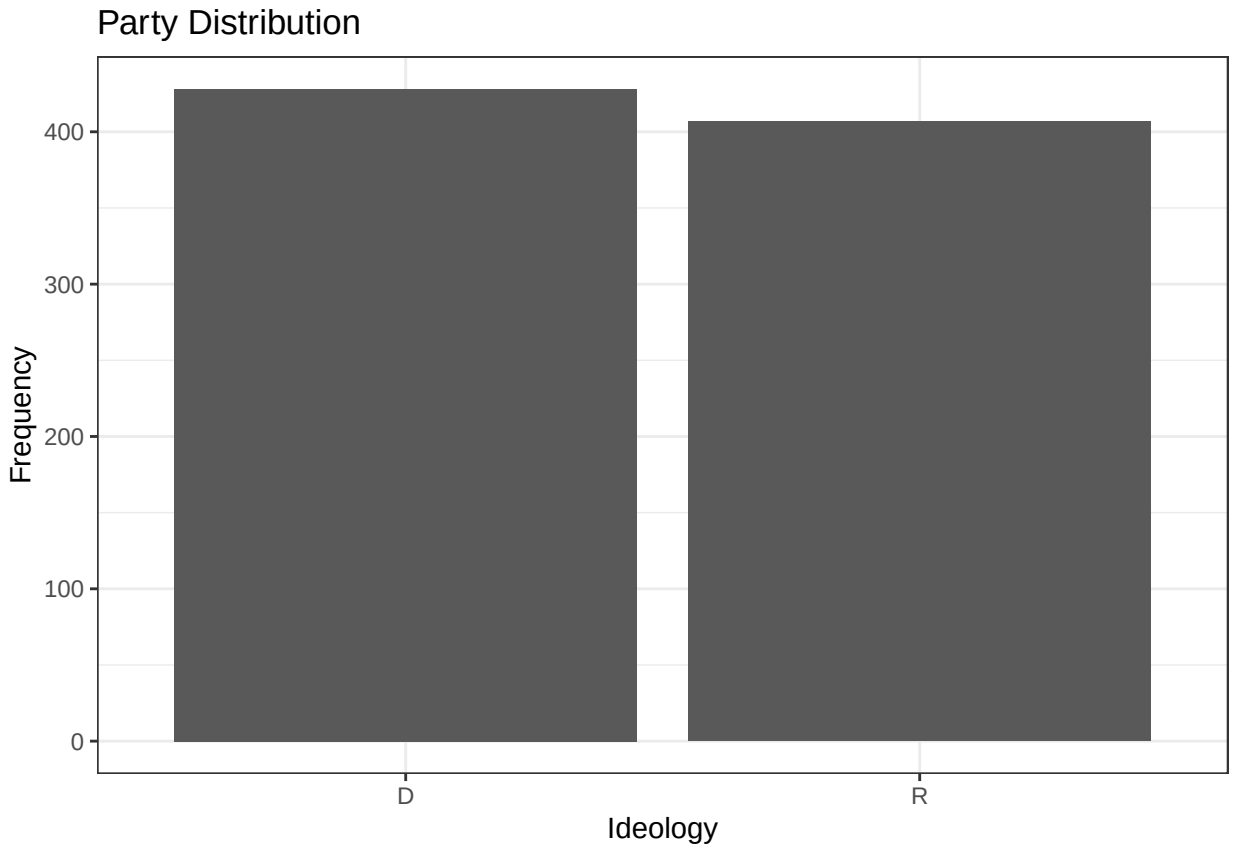
1.2 Test for Independence (5 points)

Analyze the relationships between political ideology and political party and gender using basic visualizations. Afterward, generate a contingency table and assess the independence of political ideology from political party and gender.

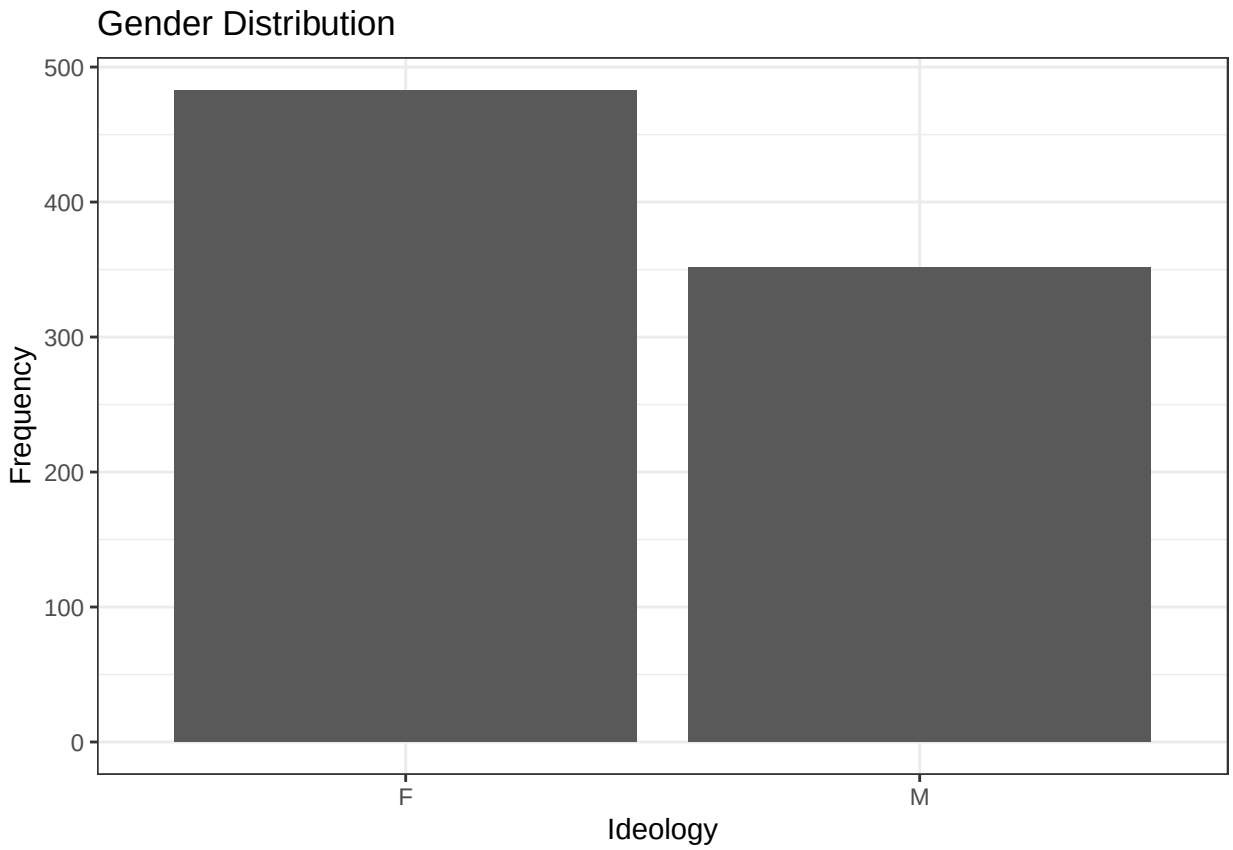
```r
p <- ggplot(data=pol_ideol, aes(x=ideol_levels, y=count))+
  geom_bar(stat='identity') +
  ggtitle("Political Ideology Distribution") +
  xlab("Ideology") + ylab("Frequency")
p
```



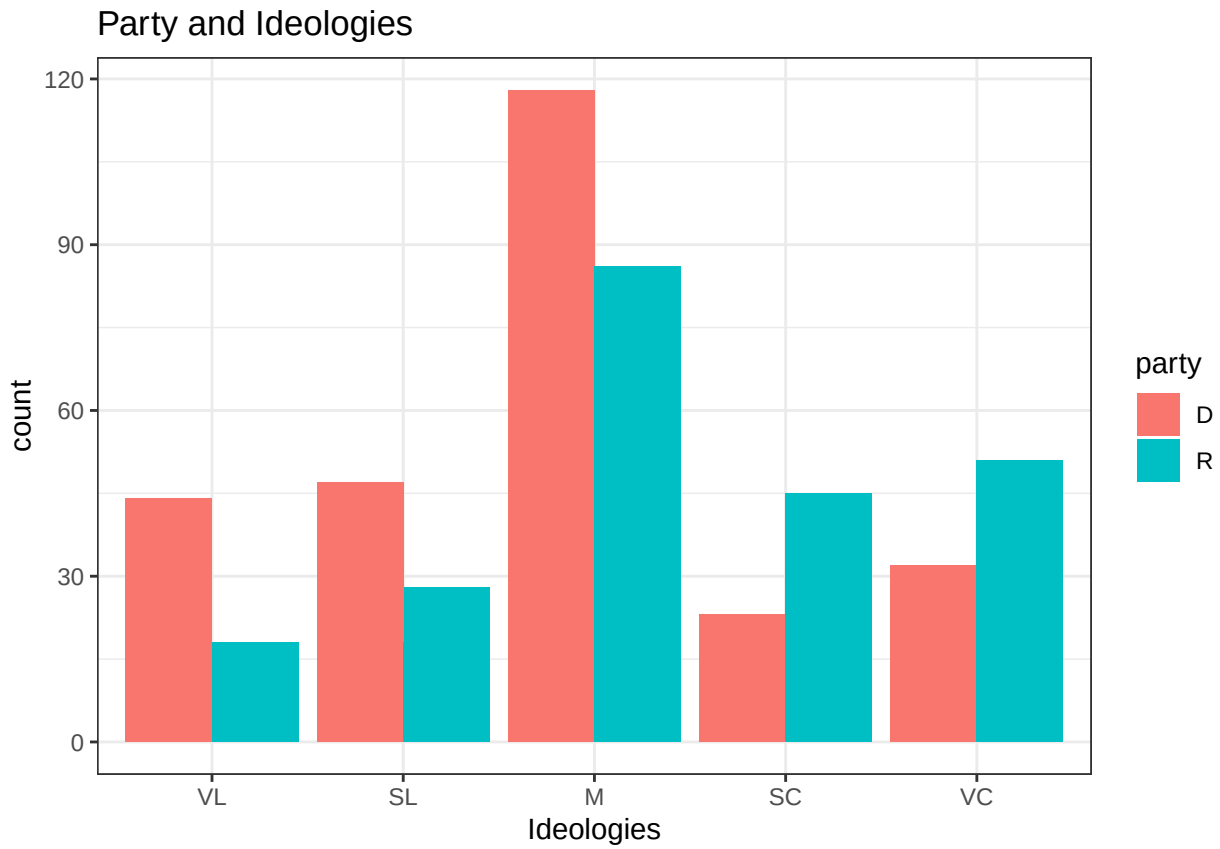
```
p <- ggplot(data=pol_ideol, aes(x=party, y=count))+
  geom_bar(stat='identity') +
  ggtitle("Party Distribution") +
  xlab("Ideology") + ylab("Frequency")
p
```



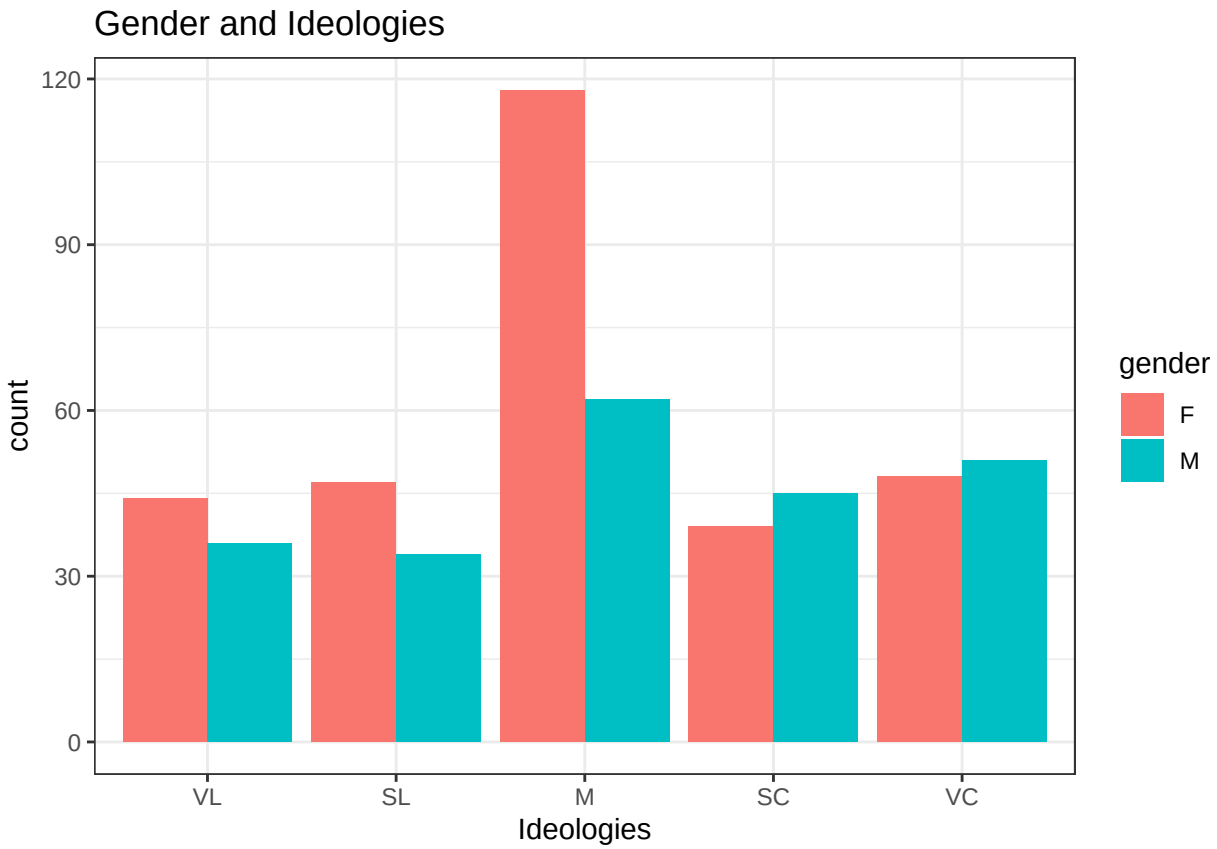

```
p <- ggplot(data=pol_ideol, aes(x=gender, y=count))+
  geom_bar(stat='identity') +
  ggtitle("Gender Distribution") +
  xlab("Ideology") + ylab("Frequency")
p
```




```
ggplot(pol_ideol, aes(fill=party, y=count, x=ideol_levels)) +
    geom_bar(position="dodge", stat="identity") +
    ggtitle("Party and Ideologies") +
    xlab("Ideologies")
```




```
ggplot(pol_ideol, aes(fill=gender, y=count, x=ideol_levels)) +
    geom_bar(position="dodge", stat="identity") +
    ggtitle("Gender and Ideologies") +
    xlab("Ideologies")
```




```
# independence of political ideology and gender
tab1 <- xtabs(count ~ ideol_levels + gender, data=pol_ideol)

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

```
##             gender
## ideol_levels          F          M
##           VL -0.3374985  0.3374985
##           SL  0.3000854 -0.3000854
##           M   2.8091751 -2.8091751
##           SC -2.0242541  2.0242541
##           VC -1.6407719  1.6407719
```

```
test1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 10.732, df = 4, p-value = 0.02975
```

```
# independence of political ideology and party
tab1 <- xtabs(count ~ ideol_levels + party, data=pol_ideol)
```

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

```
##             party
## ideol_levels         D         R
##           VL  4.834658 -4.834658
##           SL  3.065992 -3.065992
##           M   1.067069 -1.067069
##           SC -4.477226  4.477226
##           VC -4.273049  4.273049
```

```r
test1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 60.905, df = 4, p-value = 1.872e-12
```

The p-values for gender and party are both small, so the null hypothesis is that party and gender are independent from ideology is rejected, and we can conclude that ideology is dependent on these two variables.

1.3 Regression analysis (5 points)

Estimate a multinomial regression model and ordinal (proportional odds) regression model that both include party, gender, and their interaction. Perform Likelihood Ratio Tests (LRTs) to test the importance of each explanatory variable.

Also, test whether the proportional odds assumption in the ordinal model is satisfied. Based on this test and other results, which model do you think is more valid?

```r
mod.nomial <- multinom(ideol_levels ~ party + gender + party:gender, data=pol_ideol, weights =
```

```
## # weights:  25 (16 variable)
## initial  value 1343.880657
## iter  10 value 1231.244704
## iter  20 value 1229.548447
## final  value 1229.543342
## converged
```

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

```
## Call:
## multinom(formula = ideol_levels ~ party + gender + party:gender,
##     data = pol_ideol, weights = count)
##
## Coefficients:
##    (Intercept)    partyR     genderM partyR:genderM
## SL  0.06598601 0.3758637 -0.12315074      0.0867552
```

```
## M   0.98652431 0.5774673 -0.59976058      0.6779778
## SC -0.64869284 1.4219096 -0.04442702      0.5929326
## VC -0.31838463 1.2992041 -0.12968265      0.5957616
##
## Std. Errors:
##    (Intercept)    partyR   genderM partyR:genderM
## SL   0.2097724 0.3677971 0.3181097      0.5756306
## M    0.1766421 0.3136662 0.2790125      0.4944619
## SC   0.2573076 0.3839323 0.3867020      0.5799046
## VC   0.2323285 0.3610630 0.3538841      0.5518725
##
## Residual Deviance: 2459.087
## AIC: 2491.087
```

```
mod.ord <- clm(ideol_levels ~ party + gender, data=pol_ideol, weights = count)
summary(mod.ord)
```

```
## formula: ideol_levels ~ party + gender
## data:    pol_ideol
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  835  -1237.07 2486.14 5(0)  2.28e-08 2.6e+01
##
## Coefficients:
##         Estimate Std. Error z value Pr(>|z|)
## partyR    0.9636     0.1297   7.431 1.08e-13 ***
## genderM   0.1169     0.1273   0.918    0.359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##       Estimate Std. Error z value
## VL|SL  -1.4518     0.1226 -11.837
## SL|M   -0.4583     0.1048  -4.375
## M|SC    1.2550     0.1142  10.987
## SC|VC   2.0890     0.1293  16.159
```

```
Anova(mod.nomial, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ideol_levels
##              LR Chisq Df Pr(>Chisq)
## party          60.555  4  2.218e-12 ***
## gender          8.965  4    0.06198 .
## party:gender    3.245  4    0.51763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(mod.ord, test = "Chisq")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ideol_levels
##        Df   Chisq Pr(>Chisq)
## party   1 140.112  < 2.2e-16 ***
## gender  1  19.137  1.217e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using Anova to test each of the explanatory variables in the multinomial model, the p-value is small for party, so we can conclude that is has an effect on ideology. The p-values were large for gender and the party-gender interaction effect, so for these variables we are unable to reject the null hypothesis.

For the ordinal model, the p-value for party, gender and its interaction was small, so these variables do have an affect on ideology under this model.

```r
nominal_test(mod.ord)
```

```
## Tests of nominal effects
##
## formula: ideol_levels ~ party + gender
##        Df  logLik    AIC    LRT Pr(>Chi)
## <none>    -1237.1 2486.1
## party   3 -1235.2 2488.4 3.7115  0.29435
## gender  3 -1233.4 2484.8 7.2981  0.06298 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values for all of the variables are large, we cannot reject the null hypothesis for proportional odds, so the ordinal model is not valid and we should use the multinomial model.

1.4 Estimated probabilities (5 points)

Compute the estimated probabilities for each ideology level given all possible combinations of the party and gender levels.

```r
data.f.d <- data.frame(party = 'D', gender = 'F')
pred.f.d <- predict(mod.nomial, newdata = data.f.d, type="probs", se=TRUE)

data.f.r <- data.frame(party = 'R', gender = 'F')
pred.f.r <- predict(mod.nomial, newdata = data.f.r, type="probs", se=TRUE)

data.m.d <- data.frame(party = 'D', gender = 'M')
pred.m.d <- predict(mod.nomial, newdata = data.m.d, type="probs", se=TRUE)

data.m.r <- data.frame(party = 'R', gender = 'M')
pred.m.r <- predict(mod.nomial, newdata = data.m.r, type="probs", se=TRUE)
```

```
tab_dat <- rbind(pred.f.d, pred.f.r, pred.m.d, pred.m.r)
colnames(tab_dat) <- c("VL", "SL", "M", "SC", "VC")
rownames(tab_dat) <- c("Female, Democrat", "Female, Republican", "Male, Democrat", "Male, Repul
tab <- as.table(tab_dat)
tab
```

```
##                            VL         SL          M         SC         VC
## Female, Democrat   0.16666222 0.17803054 0.44697087 0.08711911 0.12121726
## Female, Republican 0.08219087 0.12785463 0.39269600 0.17808499 0.21917351
## Male, Democrat     0.21951379 0.20731727 0.32317009 0.10975989 0.14023895
## Male, Republican   0.06383112 0.09574563 0.32978787 0.23935868 0.27127670
```

1.5 Contingency table of estimated counts (5 points)

Construct a contingency table with estimated counts from the model. These estimated counts are found by taking the estimated probability for each ideology level multiplied by their corresponding number of observations for a party and gender combination.

For example, there are 264 observations for gender = "F" and party = "D". Because the multinomial regression model results in $\hat{\pi}_{VL} = 0.1667$, this model's estimated count is $0.1667 \times 264 = 44$.

- Are the estimated counts the same as the observed? Conduct a goodness of fit test for this and explain the results.

```
fd.sum <- sum(pol_ideol[which(pol_ideol$gender == "F" & pol_ideol$party == 'D'), ]$count)
fr.sum <- sum(pol_ideol[which(pol_ideol$gender == "F" & pol_ideol$party == 'R'), ]$count)
md.sum <- sum(pol_ideol[which(pol_ideol$gender == "M" & pol_ideol$party == 'D'), ]$count)
mr.sum <- sum(pol_ideol[which(pol_ideol$gender == "M" & pol_ideol$party == 'R'), ]$count)

# VL contingency
vl.prob <- unname(tab[, 1])
vl.dat <- matrix(round(c(vl.prob[1]*fd.sum, vl.prob[2]*fr.sum, vl.prob[3]*md.sum, vl.prob[4]*m
colnames(vl.dat) <- c("F", "M")
rownames(vl.dat) <- c("D", "R")
print("VL contingency table")
```

```
## [1] "VL contingency table"
```

```
vl.dat
```

```
##    F  M
## D 44 36
## R 18 12
```

```
# SL contingency
sl.prob <- unname(tab[, 2])
sl.dat <- matrix(round(c(sl.prob[1]*fd.sum, sl.prob[2]*fr.sum, sl.prob[3]*md.sum, sl.prob[4]*m
colnames(sl.dat) <- c("F", "M")
rownames(sl.dat) <- c("D", "R")
print("SL contingency table")
```

```
## [1] "SL contingency table"
```

```
sl.dat
```

```
##    F  M
## D 47 34
## R 28 18
```

```
# M contingency
m.prob <- unname(tab[, 3])
m.dat <- matrix(round(c(m.prob[1]*fd.sum, m.prob[2]*fr.sum, m.prob[3]*md.sum, m.prob[4]*mr.sum
colnames(m.dat) <- c("F", "M")
rownames(m.dat) <- c("D", "R")
print("M contingency table")
```

```
## [1] "M contingency table"
```

```
m.dat
```

```
##     F  M
## D 118 53
## R  86 62
```

```
# SC contingency
sc.prob <- unname(tab[, 4])
sc.dat <- matrix(round(c(sc.prob[1]*fd.sum, sc.prob[2]*fr.sum, sc.prob[3]*md.sum, sc.prob[4]*m
colnames(sc.dat) <- c("F", "M")
rownames(sc.dat) <- c("D", "R")
print("SC contingency table")
```

```
## [1] "SC contingency table"
```

```
sc.dat
```

```
##    F  M
## D 23 18
## R 39 45
```

```
# VC contingency
vc.prob <- unname(tab[, 5])
vc.dat <- matrix(round(c(vc.prob[1]*fd.sum, vc.prob[2]*fr.sum, vc.prob[3]*md.sum, vc.prob[4]*m
colnames(vc.dat) <- c("F", "M")
rownames(vc.dat) <- c("D", "R")
print("VC contingency table")
```

```
## [1] "VC contingency table"
```

```
vc.dat
```

```
##    F  M
## D 32 23
## R 48 51
```

```
vl.actual <- matrix(pol_ideol[which(pol_ideol$ideol_levels == 'VL'), ]$count, nrow=2, ncol=2)
chisq.test(vl.dat, vl.actual)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vl.dat
## X-squared = 0.065069, df = 1, p-value = 0.7987
```

```
sl.actual <- matrix(pol_ideol[which(pol_ideol$ideol_levels == 'SL'), ]$count, nrow=2, ncol=2)
chisq.test(sl.dat, sl.actual)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sl.dat
## X-squared = 0.015786, df = 1, p-value = 0.9
```

```
m.actual <- matrix(pol_ideol[which(pol_ideol$ideol_levels == 'M'), ]$count, nrow=2, ncol=2)
chisq.test(m.dat, m.actual)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m.dat
## X-squared = 3.6279, df = 1, p-value = 0.05682
```

```
sc.actual <- matrix(pol_ideol[which(pol_ideol$ideol_levels == 'SC'), ]$count, nrow=2, ncol=2)
chisq.test(sc.dat, sc.actual)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sc.dat
## X-squared = 0.67991, df = 1, p-value = 0.4096
```

```
vc.actual <- matrix(pol_ideol[which(pol_ideol$ideol_levels == 'VC'), ]$count, nrow=2, ncol=2)
chisq.test(vc.dat, vc.actual)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vc.dat
## X-squared = 0.97175, df = 1, p-value = 0.3242
```

The p-values are large which means that there is not enough evidence to reject the null hypothesis and conclude that the distribution of the expected and actual counts are different from each other.

1.6 Odds ratios and confidence intervals (8 points)

To better understand relationships between the explanatory variables and the response, compute odds ratios and their confidence intervals from the estimated models and interpret them.

```r
beta.hat2 <- coefficients(mod.nomial)[1, 2:4]
beta.hat3 <- coefficients(mod.nomial)[2, 2:4]
beta.hat4 <- coefficients(mod.nomial)[3, 2:4]
beta.hat5 <- coefficients(mod.nomial)[4, 2:4]

# Sl vs VL OR
round(exp(beta.hat2), 2)
```

```
##         partyR        genderM partyR:genderM
##           1.46           0.88           1.09
```

```r
# Sl vs VL OR
round(exp(beta.hat3), 2)
```

```
##         partyR        genderM partyR:genderM
##           1.78           0.55           1.97
```

```r
# Sl vs VL OR
round(exp(beta.hat4), 2)
```

```
##         partyR        genderM partyR:genderM
##           4.15           0.96           1.81
```

```r
# Sl vs VL OR
round(exp(beta.hat5), 2)
```

```
##         partyR        genderM partyR:genderM
##           3.67           0.88           1.81
```

```r
# beta confidence interval
conf.beta <- confint(object = mod.nomial, level=0.95)

## construct CI for OR

exp(conf.beta)
```

```
## , , SL
##
##                     2.5 %   97.5 %
## (Intercept)    0.7081069 1.611447
## partyR         0.7082166 2.994367
## genderM        0.4739591 1.649270
## partyR:genderM 0.3529390 3.370195
##
## , , M
##
##                     2.5 %   97.5 %
## (Intercept)    1.8970730 3.791404
## partyR         0.9633802 3.294458
## genderM        0.3177104 0.948469
## partyR:genderM 0.7474037 5.191929
```

```
##
## , , SC
##
##                     2.5 %    97.5 %
## (Intercept)    0.3156863 0.8655594
## partyR         1.9530967 8.7969317
## genderM        0.4482748 2.0411122
## partyR:genderM 0.5806196 5.6379734
##
## , , VC
##
##                     2.5 %   97.5 %
## (Intercept)    0.4612845 1.146795
## partyR         1.8067572 7.440028
## genderM        0.4389882 1.757544
## partyR:genderM 0.6151503 5.351688
```