W271 Group Lab

Bike share demand

Annie DeForge, Hannah Abraham, Mariah Ehmke, Nora Povejsil

Contents

1 Short Questions

2 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.

2.1 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.

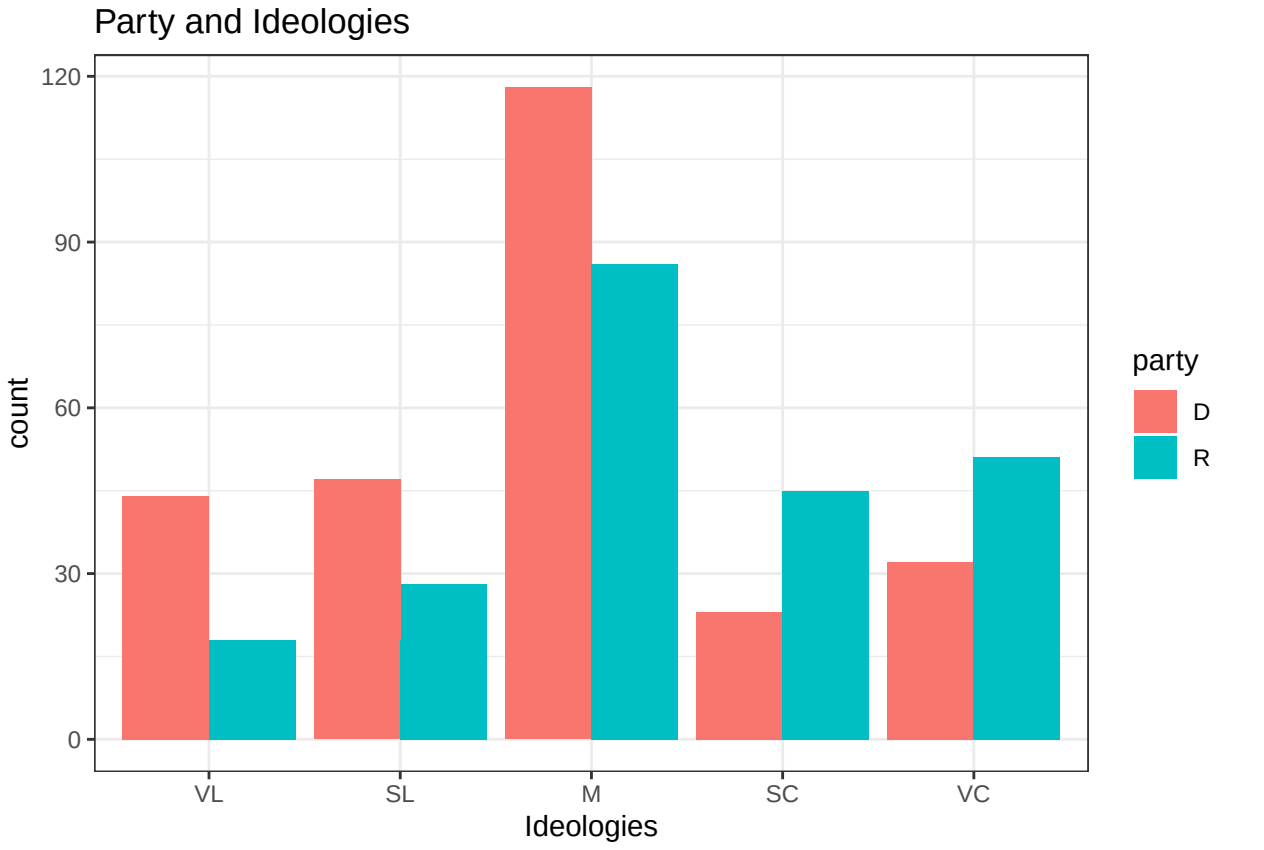
2.2 Test for Independence (5 points)

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. *Comment*

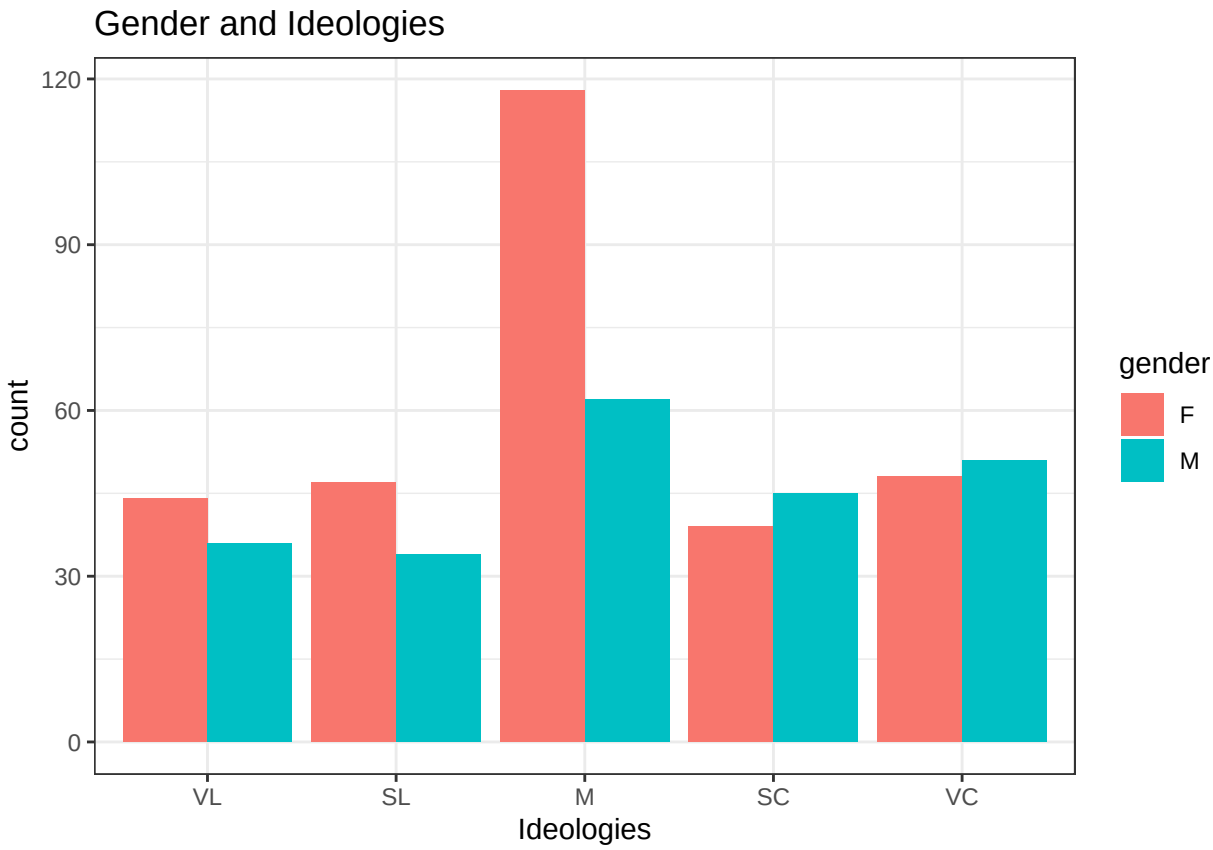
The histogram of ideology is presented first by party and then by gender. Both histograms show the largest concentration of voters in the "neutral" political ideology. Democrats are especially likely to be "neutral" as are women. Republicans are more prominent in the "slightly conservative" and "very conservative" categories. Men also are more likely to be in these categories.

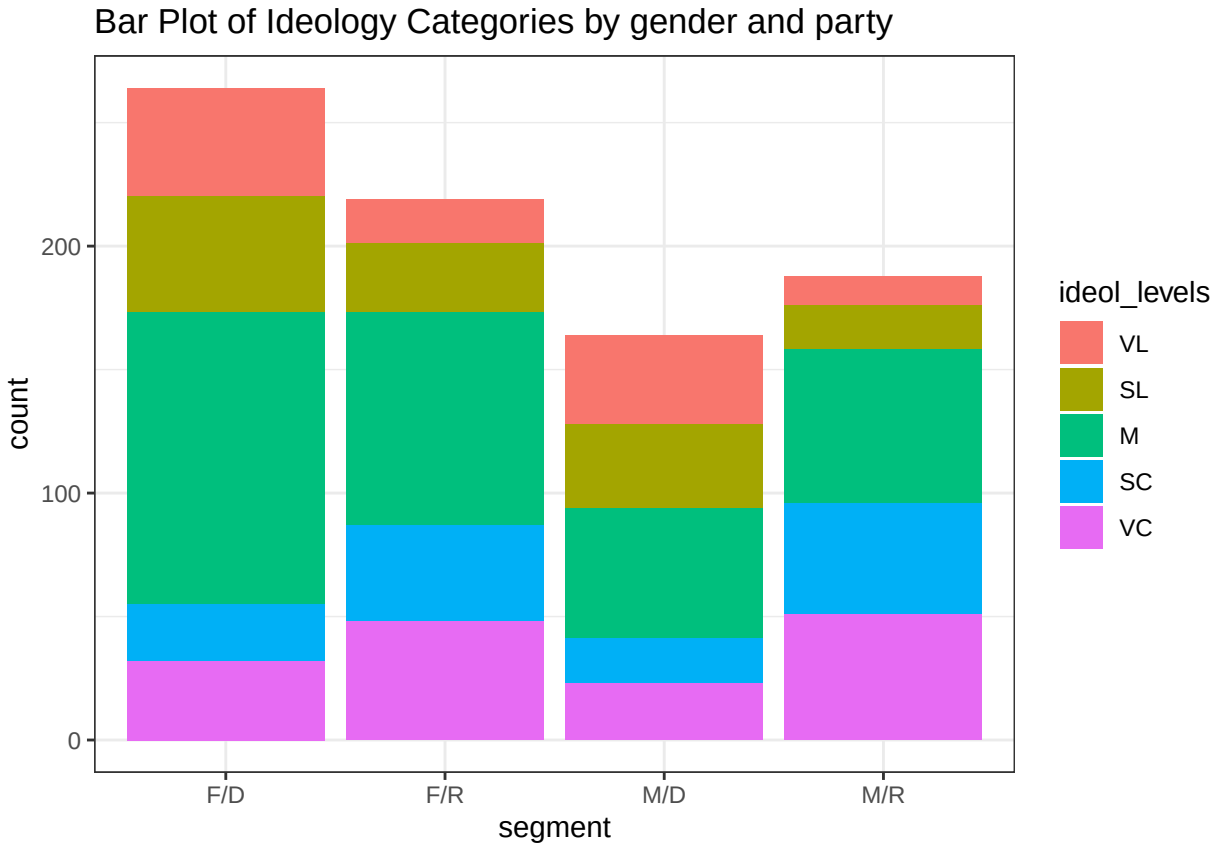
The following bar plot suggests the data were collected from an equal number of republicans and democrats, and men and women. Finally, the bar plot "Bar Plot of Ideology Categorically by Gender and Party" also suggests female democrats are more likely to fall in a liberal ideology category while male republicans will be more likely to fall in a conservative ideological category.

```
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")
```


Gender and Ideologies
120
90
60
30
0
count
VL
SL
M
SC
VC
Ideologies
gender
F
M


Bar Plot of Ideology Categories by gender and party
200
100
0
count
F/D
F/R
M/D
M/R
segment
ideol_levels
VL
SL
M
SC
VC

From the visualizations we can see that of the conservative categories, there are more men. For the liberal categories there are more women. The greatest proportion of the observations are moderate, and women are more likely to be moderate.

Comment
The results of the Chi-Square test for independence tests the null hypothesis $H_0 : \pi_{ij} = \pi_i \pi_j$. The alternative hypothesis is $H_a : \pi_{ij} \neq \pi_i \pi_j$. We reject the null hypothesis for the test of independence between gender and ideology ($\chi^2 = 10.73, p < 0.05$). Gender and ideology are not independent according to this test. The chi-square value is 60.905 and p<0.001 for the test of independence between ideology and party. We reject the null. The party and ideology levels are not independent.

```
##             gender
## ideol_levels   F   M
##           VL  62  48
##           SL  75  52
##           M  204 115
##           SC  62  63
##           VC  80  74
```

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

```
##             party
## ideol_levels   D   R
##           VL  80  30
##           SL  81  46
##           M  171 148
##           SC  41  84
##           VC  55  99
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab2
## 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.

2.3 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?

Comment
We estimate the multinomial logit and ordinal regression models. The AIC for the multinomial logit model is 2491.087. It is 2484.15 for the ordinal model.

Using Anova to test each of the explanatory variables using a χ^2 test. In the multinomial logit model, only one variable, party, is significant–only as an isolated variable and not in the interaction term. In the multinomial logit model, party is positively related to a voter being in any one of the non-neutral categories. Gender does not achieve significance with a p-value of 0.06.

In the ordinal model, both party and gender, and their interaction term is statistically significant ($p<0.001$). 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.

We then use the proportional odds test to determine whether the odds model is appropriate for this task. While the ordinal model performed slightly better than the multinomial by AIC and number of significant coefficients, the hypothesis that the probability of the outcome increases with each level is rejected. The p-value of their significance in explaining the likelihood improvement is does not support the null. The multinomial model is our model of choice for this exercise.

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

```
## 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
```

```
## formula: ideol_levels ~ party + gender + party:gender
## data:    pol_ideol
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
```

```
##  logit flexible  835  -1235.08 2484.15 5(0)  2.58e-08 6.0e+01
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## partyR           0.7562     0.1659   4.559 5.13e-06 ***
## genderM         -0.1431     0.1820  -0.786   0.4318
## partyR:genderM   0.5091     0.2550   1.996   0.0459 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##       Estimate Std. Error z value
## VL|SL  -1.5521     0.1332 -11.656
## SL|M   -0.5550     0.1157  -4.796
## M|SC    1.1647     0.1226   9.501
## SC|VC   2.0012     0.1364  14.667
```

Having printed the multinomial and ordinal model, we can run an ANOVA test on both.

```
## 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
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: ideol_levels
##              Df   Chisq Pr(>Chisq)
## party         1 474.276  < 2.2e-16 ***
## gender        1 317.454  < 2.2e-16 ***
## party:gender  1  90.268  < 2.2e-16 ***
## ---
## 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.

```
## Tests of nominal effects
##
## formula: ideol_levels ~ party + gender + party:gender
##              Df  logLik    AIC     LRT Pr(>Chi)
```

```
## <none>          -1235.1 2484.2
## party         3 -1233.1 2486.3  3.8711   0.2757
## gender        3 -1232.3 2484.6  5.5831   0.1338
## party:gender  9 -1229.5 2491.1 11.0634   0.2714
```

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.

2.4 Estimated probabilities (5 points)

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

```
##                            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
```

2.5 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.

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

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

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

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

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

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

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

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

```
## [1] "VC contingency table"
##    F  M
## D 32 23
## R 48 51
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vl.dat
## X-squared = 0.065069, df = 1, p-value = 0.7987
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sl.dat
## X-squared = 0.015786, df = 1, p-value = 0.9
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  m.dat
## X-squared = 3.6279, df = 1, p-value = 0.05682
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sc.dat
## X-squared = 0.67991, df = 1, p-value = 0.4096
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  vc.dat
## X-squared = 0.97175, df = 1, p-value = 0.3242
```

Comment 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. The estimated counts are the same as the observed counts.

2.6 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.

The odds ratios for a given variable, depend on the category of comparison (e.g., VL, SL, N, SC, or VL). In the multinomial logit model, the left-out gender category was female and the left-out party category was democrat. The comparison level for ideology was 'Neutral.' We compare the odds of the different coefficients for the coefficients from each VL, SL, SC, and VL level.
Assuming the model specification $\log(\pi_j/\pi_1) = \beta_{j0} + \beta_{jGender}x_{Gender} + \beta_{jParty} + \beta_{jGender*Party}$

where j = SC, SL, VC, VL.

The Odds Ratio for male versus female for the each jth level is equal to $\frac{exp(\beta_{j0}+\beta_{j1}*(gender+c)+\beta_{j3}*party*(gender+c))}{exp(\beta_{j0}+\beta_{j1}*gender+\beta_{j3}*party*gender)} = exp(\beta_{j1} * c + \beta_{j3} * party * c)$ where c = 1 for a categorical variable. Further, gender will be equal to one as it is also a categorical variable.

The Odds Ratio for republican versus democrat for each jth level is equal to $\frac{exp(\beta_{j0}+\beta_{j2}*(party+c)+\beta_{j3}*gender*(party+c))}{exp(\beta_{j0}+\beta_{j2}*party+\beta_{j3}*gender*party)} = exp(\beta_{j2} * c + \beta_{j3} * gender * c)$ where c = 1.

‘

```
## NULL
```

```
## [1] "model summary"
```

```
## 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
```

```
## [1] "beta_hats"
```

```
##    (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
```

Table Odds Ratios for Gender

Generally, there are higher odds men will be rated as either slightly or very conservative rather than moderate compared to women. Men are more likely to fall in the conservative categories instead

of the moderate category if they are republicans, rather than democrats. Men were much more likely than women to be in the liberal categories instead of the neutral category conditional if they were classified as democrat. Generally, republican women than democratic women to be in the the conservative categories rather than moderate.

```
ideology_labels <- c('SL', 'SL', 'SL', 'M', 'M', 'M', 'SC', 'SC', 'SC', 'VC', 'VC', 'VC')
gender_labels <- c('Male','Male', 'Female', 'Male','Male', 'Female', 'Male','Male', 'Female',
party_labels <- c('Rep', 'Dem', 'Rep', 'Rep', 'Dem', 'Rep', 'Rep', 'Dem', 'Rep', 'Rep', 'Dem',
odds_ratio_gender <- c(sl.beta.male.rep, sl.beta.male.dem, sl.beta.fem.rep, m.beta.male.rep, m

gender_odds <- data.frame(Ideology = ideology_labels, Gender = gender_labels, Party = party_lab
gender_odds
```

```
##    Ideology Gender Party OR.hat
## 1        SL   Male   Rep 2.3295
## 2        SL   Male   Dem 0.8841
## 3        SL Female   Rep 1.4562
## 4         M   Male   Rep 1.9265
## 5         M   Male   Dem 0.5489
## 6         M Female   Rep 1.7815
## 7        SC   Male   Rep 7.1737
## 8        SC   Male   Dem 0.9565
## 9        SC Female   Rep 4.1450
## 10       VC   Male   Rep 5.8267
## 11       VC   Male   Dem 0.8784
## 12       VC Female   Rep 3.6664
```

Looking at this table, we can see that the largest odds ratios occurred for men or women Republicans who identified as slightly conservative they were 7.17 and 4.14 times, respectively, more likely to identify as slightly conservative than very liberal compared to the base comparison of a female Democrat. The smallest odds ratios were for male democrats who were less likely to identify with an ideology compared to very liberal.

##Confidence Intervals for Odds Ratios

```
conf.beta <- confint(object = mod.nomial, level = 0.95)
conf.beta #Results are in 3-D array
```

```
## , , SL
##
##                     2.5 %    97.5 %
## (Intercept)    -0.3451603 0.4771323
## partyR         -0.3450053 1.0967328
## genderM        -0.7466342 0.5003328
## partyR:genderM -1.0414601 1.2149705
##
## , , M
##
##                      2.5 %      97.5 %
## (Intercept)     0.64031215  1.33273648
```

```
## partyR          -0.03730709  1.19224165
## genderM         -1.14661502 -0.05290614
## partyR:genderM -0.29114977  1.64710534
##
## , , SC
##
##                      2.5 %     97.5 %
## (Intercept)    -1.1530064 -0.1443793
## partyR          0.6694162  2.1744030
## genderM        -0.8023489  0.7134949
## partyR:genderM -0.5436594  1.7295247
##
## , , VC
##
##                      2.5 %    97.5 %
## (Intercept)    -0.7737402 0.1369710
## partyR          0.5915336 2.0068746
## genderM        -0.8232827 0.5639174
## partyR:genderM -0.4858887 1.6774119
```

```
#Outcomes for Gender SC
varcov <- vcov(mod.nomial)
# varcov
genderlevels1 <- c(1,1,0,1,1,0,1,1,0,1,1,0)
partylevels1 <- c(1,0,1,1,0,1,1,0,1,1,0,1)
interactlevels1 <- c(1,0,0,1,0,0,1,0,0,1,0,0)

sl.beta.male.rep.ci <- sl.beta.male.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[2,2]+varcov[3
sl.beta.male.dem.ci <- sl.beta.male.dem + qnorm(p= c(0.025, 0.975))*sqrt((varcov[3,3]))
sl.beta.fem.rep.ci <- sl.beta.fem.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[2,2]))
```

```
#Outcomes for Gender M
m.beta.male.rep.ci <- m.beta.male.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[6,6]+varcov[7,7]
m.beta.male.dem.ci <- m.beta.male.dem + qnorm(p= c(0.025, 0.975))*sqrt((varcov[7,7]))
m.beta.fem.rep.ci <- m.beta.fem.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[6,6]))
```

```
#Outcomes for Gender SC
sc.beta.male.rep.ci <- sc.beta.male.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[10,10]+varcov
sc.beta.male.dem.ci <- sc.beta.male.dem + qnorm(p= c(0.025, 0.975))*sqrt((varcov[11,11]))
sc.beta.fem.rep.ci <- sc.beta.fem.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[10,10]))
```

```
#Outcomes for Gender VC
vc.beta.male.rep.ci <- vc.beta.male.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[14,14]+varcov
vc.beta.male.dem.ci <- vc.beta.male.dem + qnorm(p= c(0.025, 0.975))*sqrt((varcov[15,15]))
vc.beta.fem.rep.ci <- vc.beta.fem.rep + qnorm(p= c(0.025, 0.975))*sqrt((varcov[14,14]))
```

```
gender_odds_cis <- data.frame(Ideology = ideology_labels, Gender = gender_labels, Party = party

gender_odds_cis
```

```
##    Ideology Gender Party  Wald.lower Wald.upper
## 1        SL   Male   Rep 1.574298367   3.084664
## 2        SL   Male   Dem 0.260646894   1.507614
## 3        SL Female   Rep 0.735379636   2.177118
## 4         M   Male   Rep 1.265195565   2.587726
## 5         M   Male   Dem 0.002088609   1.095797
## 6         M Female   Rep 1.166746253   2.396295
## 7        SC   Male   Rep 6.378652772   7.968656
## 8        SC   Male   Dem 0.198623510   1.714467
## 9        SC Female   Rep 3.392534741   4.897522
## 10       VC   Male   Rep 5.073605127   6.579834
## 11       VC   Male   Dem 0.184774130   1.571974
## 12       VC Female   Rep 2.958707089   4.374048
```

For all of the OR confidence intervals, the interval is not inclusive of 0 except for female republicans who were slightly conservative, meaning that for all of the other categories, we can conclude that the difference in likelihood is significantly different than 0.