Lab 08

Rohit Kulkarni

3/9/2022

```
gss <- read_csv("data/gss2016.csv",
                na = c("", "Don't know", "No answer",
                       "Not applicable"),
                guess_max = 2867) %>%
  select(natmass, age, sex, sei10, region, polviews) %>%
  drop_na()
```

```
## Rows: 2867 Columns: 935

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (810): wrkstat, marital, martype, childs, age, degree, sex, race, born, ...
## dbl (106): year, id_, hrs2, sphrs2, sibs, agekdbrn, educ, emailmin, emailhr,...
## lgl  (19): bigbang1, spwrkgvt, where6, away8, where8, away9, where9, mar10, ...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exercises

Part I: Exploratory Data Analysis

See Reorder factor levels by hand for documentation about fct_relevel.

1. The variable `natmass` will be the response variable in the model, and you want to compare more opinionated views to the moderate position. Recode `natmass` so it is a factor variable with "About right" as the baseline.

```
gss$natmass <- as.factor(gss$natmass)

gss$natmass <- relevel(gss$natmass, ref = "About right")
```

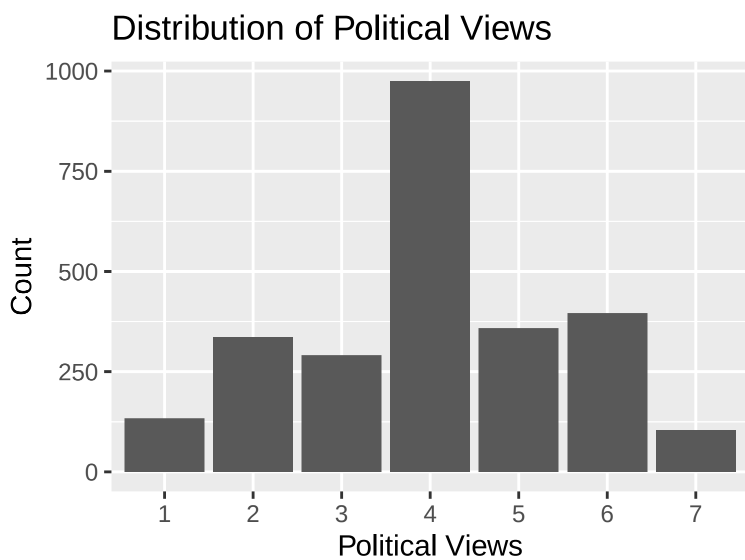
2. Recode `polviews` so it is a factor variable type with levels that are in an order that is consistent with question on the survey. Note how the categories are spelled in the data.

```r
gss$polviews <- fct_recode(gss$polviews,
                           "1"="Extremely liberal",
                           "2"="Liberal",
                           "3"="Slightly liberal",
                           "4"="Moderate",
                           "5"="Slghtly conservative",
                           "6"="Conservative",
                           "7"="Extrmly conservative")
gss$polviews <- fct_relevel(gss$polviews,
                            "1", "2", "3", "4", "5", "6", "7")
```

Make a plot of the distribution of `polviews`. Which political view occurs most frequently in this data set?

```r
ggplot(data=gss, aes(x = polviews)) + geom_histogram(stat="count") +
  labs(x = "Political Views",
       y = "Count",
       title = "Distribution of Political Views")
```

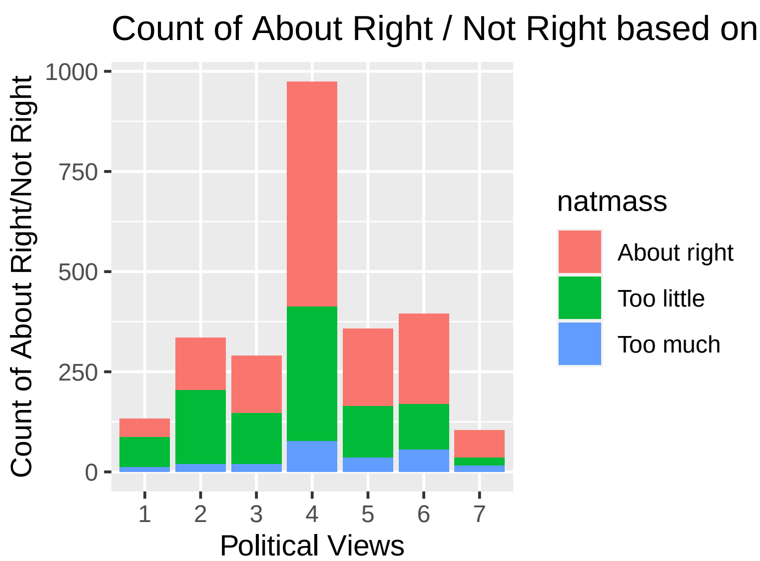
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



The political view occurs the most frequently is `Moderate`.

3. Make a plot displaying the relationship between `natmass` and `polviews`. Use the plot to describe the relationship between a person's political views and their views on mass transportation spending.

```r
ggplot(data = gss, aes(fill=natmass, x=polviews)) +
    geom_bar(position="stack") +
  labs(x = "Political Views",
       y = "Count of About Right/Not Right",
       title = "Count of About Right / Not Right based on Political Views")
```



Using the plot, I can tell that more liberal political views are associated with thinking that we do not spend enough on mass transportation and more conservative political views are associated with thinking that we spend a good amount or we need to spend less. Moderates mostly believe that we spend enough on mass transportation, while a good amount do believe we need to spend more and a small minority believe we must spend less.

4. You want to use `age` as a quantitative variable in your model; however, it is currently a character data type because some observations are coded as "89 or older". Recode `age` so that is a numeric variable. Note: Before making the variable numeric, you will need to replace the values "89 or older" with a single value.

```
gss$age <- as.factor(gss$age)
gss$age <- fct_recode(gss$age, "89" = "89 or older")
gss$age <- as.integer(gss$age)
```

Part II: Multinomial Logistic Regression Model

5. You plan to fit a model using `age`, `sex`, `sei10`, and `region` to understand variation in opinions about spending on mass transportation. Briefly explain why you should fit a multinomial logistic model.

Since the response variable, `natmass`, has three levels rather than just two, the model needs to be able to choose between one of these three options. This is why a multinomial logistic model must be used.

6. Fit the model described in the previous exercise and display the model output. Make any necessary adjustments to the variables so the intercept will have a meaningful interpretation. Be sure "About Right" is the baseline level. Be sure the full model displays in the knitted document.

```
model <- multinom(natmass ~ age + sex + sei10 + region,
             data=gss)
```

```
## # weights:  39 (24 variable)
## initial  value 2845.405828
```

```
## iter  10 value 2338.956207
## iter  20 value 2328.032754
## iter  30 value 2327.223304
## iter  30 value 2327.223281
## iter  30 value 2327.223281
## final  value 2327.223281
## converged
```

```
summary(model)
```

```
## Call:
## multinom(formula = natmass ~ age + sex + sei10 + region, data = gss)
##
## Coefficients:
##            (Intercept)         age   sexMale        sei10 regionE. sou. central
## Too little   -1.127003 0.003937463 0.1963228  0.009367978             0.2715622
## Too much     -2.141703 0.015981181 0.5532088 -0.009631888            -0.2852564
##            regionMiddle atlantic regionMountain regionNew england regionPacific
## Too little           -0.03021749     0.18270513         0.5948555     0.4086640
## Too much             -0.16241468    -0.02121511         0.8525949     0.2960679
##            regionSouth atlantic regionW. nor. central regionW. sou. central
## Too little            0.1230811             0.0297136           -0.08588673
## Too much             -0.2626466             0.1381563           -0.58273908
##
## Std. Errors:
##            (Intercept)         age    sexMale       sei10 regionE. sou. central
## Too little   0.1555957 0.002450067 0.08536084 0.001773727             0.1893107
## Too much     0.2572032 0.004051282 0.14525777 0.003191687             0.3495161
##            regionMiddle atlantic regionMountain regionNew england regionPacific
## Too little             0.1642448      0.1767700         0.2010660     0.1511281
## Too much               0.2777032      0.3034439         0.2893961     0.2423991
##            regionSouth atlantic regionW. nor. central regionW. sou. central
## Too little            0.1394087             0.1962766             0.1689158
## Too much              0.2419315             0.3019854             0.3103206
##
## Residual Deviance: 4654.447
## AIC: 4702.447
```

7. Interpret the intercept associated with odds of having an opinion of "Too much" versus "About right".

The intercept associated with "Too much" is -2.141703, which means that if all of the parameters are assumed to be the base case for categorical variables and average for quantitative variables, that the log likelihood of this case having an opinion of "Too much" is -2.141703.

8. Consider the relationship between age and one's opinion about spending on mass transportation. Interpret the coefficient of age in terms of the odds of having an opinion of "Too little" versus "About right".

The coefficient for age in "Too little" is 0.003937463, which is a very low coefficient. This means that an increase in a person's age, according to the model, has little correlation with whether or not they feel that we are spending too much on mass transportation.

9. Now that you have adjusted for some demographic factors, let's examine whether a person's political views has a significant impact on their attitude towards spending on mass transportation.

Conduct the appropriate test to determine if polviews is a significant predictor of attitude towards spending on mass transportation. State the null and alternative hypothesis, display all relevant code and output, and state your conclusion in the context of the problem.

H_0: A person's political views are not a significant predictor of their attitude towards spending on mass transportation. H_A: A person's political views are a significant predictor of their attitude towards spending on mass transportation.

To see the affect that `polviews` has on `natmass`, we can perform a regression analysis

```
table(gss$natmass, gss$polviews)
```

```
##
##                  1   2   3   4   5   6   7
##   About right  46 132 143 562 194 226  69
##   Too little   75 185 128 335 128 114  20
##   Too much     12  19  19  77  36  55  15
```

```
chisq.test(gss$polviews, gss$natmass)
```

```
##
##  Pearson's Chi-squared test
##
## data:  gss$polviews and gss$natmass
## X-squared = 114.59, df = 12, p-value < 2.2e-16
```

Since the p-value is so small, we can assume that the alternative hypothesis is true. A person's political views are a significant predictor of their attitude towards spending on mass transportation.

10. Choose the appropriate model based on the results from the test. Use this model for the next part of the lab.

```
model <- multinom(natmass ~ polviews + age + sex + sei10 + region, data=gss)
```

```
## # weights:  57 (36 variable)
## initial  value 2845.405828
## iter  10 value 2318.201448
## iter  20 value 2277.727386
## iter  30 value 2276.064609
## iter  40 value 2275.923477
## final  value 2275.922613
## converged
```

```
summary(model)
```

```
## Call:
## multinom(formula = natmass ~ polviews + age + sex + sei10 + region,
##     data = gss)
```

```
##
## Coefficients:
##           (Intercept)  polviews2  polviews3  polviews4  polviews5   polviews6
## Too little  -0.3102171 -0.2015970 -0.5969486 -0.9694607 -0.9399259 -1.22066119
## Too much    -1.6062742 -0.6304096 -0.6706071 -0.6796402 -0.4010363 -0.07977087
##             polviews7         age   sexMale        sei10 regionE. sou. central
## Too little -1.6961661 0.006154276 0.2173925  0.008073511            0.3337970
## Too much   -0.3062883 0.014308729 0.5348721 -0.009947402           -0.3236125
##           regionMiddle atlantic regionMountain regionNew england regionPacific
## Too little            -0.0816385     0.13770904         0.4661417     0.3636141
## Too much              -0.1440201    -0.02476708         0.8790467     0.3401187
##           regionSouth atlantic regionW. nor. central regionW. sou. central
## Too little            0.1317356            0.03024385           -0.02755904
## Too much             -0.2740420            0.15818080           -0.60112344
##
## Std. Errors:
##           (Intercept) polviews2 polviews3 polviews4 polviews5 polviews6
## Too little   0.2429246 0.2226067 0.2266976 0.2026186 0.2224028 0.2237371
## Too much     0.4098348 0.4113259 0.4110834 0.3510331 0.3768073 0.3640041
##           polviews7         age    sexMale       sei10 regionE. sou. central
## Too little 0.3199111 0.002514309 0.08697107 0.001815830            0.1923246
## Too much   0.4428418 0.004111438 0.14615196 0.003231104            0.3508591
##           regionMiddle atlantic regionMountain regionNew england regionPacific
## Too little             0.1674120      0.1798141         0.2052694     0.1538893
## Too much               0.2790957      0.3047348         0.2921835     0.2438248
##           regionSouth atlantic regionW. nor. central regionW. sou. central
## Too little            0.1418407             0.1992524             0.1714862
## Too much              0.2428155             0.3038517             0.3112926
##
## Residual Deviance: 4551.845
## AIC: 4623.845
```

Part III: Model Fit

11. Calculate the predicted probabilities and residuals from your model.

```
summary(model$fitted.values)
```

```
##   About right       Too little        Too much
##  Min.   :0.1710   Min.   :0.1234   Min.   :0.01393
##  1st Qu.:0.4629   1st Qu.:0.2980   1st Qu.:0.05317
##  Median :0.5458   Median :0.3593   Median :0.07728
##  Mean   :0.5297   Mean   :0.3803   Mean   :0.08996
##  3rd Qu.:0.6062   3rd Qu.:0.4490   3rd Qu.:0.11314
##  Max.   :0.8159   Max.   :0.7186   Max.   :0.41346
```

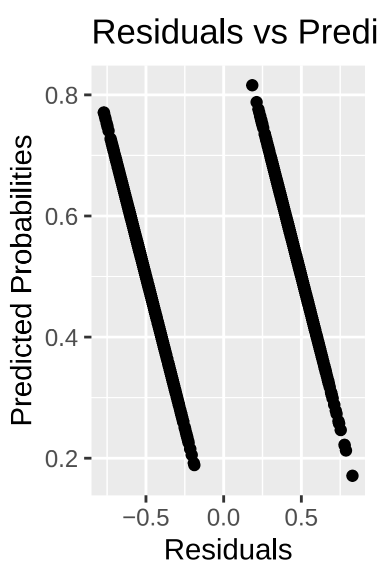
```
summary(model$residuals)
```

```
##   About right        Too little           Too much
##  Min.   :-0.7709   Min.   :-0.7028578   Min.   :-0.38907
##  1st Qu.:-0.5130   1st Qu.:-0.3670106   1st Qu.:-0.10541
##  Median : 0.3171   Median :-0.2702866   Median :-0.07098
```

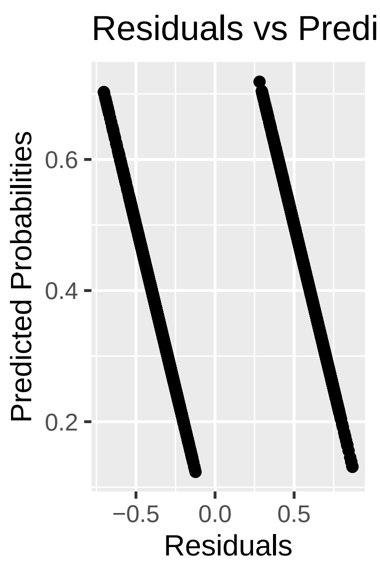
```
##  Mean    : 0.0000    Mean    : 0.0000001    Mean    : 0.00000
##  3rd Qu.: 0.4409    3rd Qu.: 0.5396926    3rd Qu.:-0.04516
##  Max.    : 0.8290    Max.    : 0.8689031    Max.    : 0.98126
```

12. Let's make some of the plots and tables you use to check the linearity assumption for multinomial logistic regression. Plot the binned residuals versus the predicted probabilities for each category of `natmass`. You will have three plots.

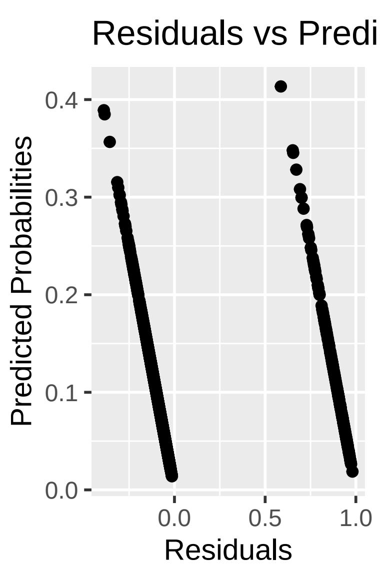
```r
residuals <- data.frame(model$residuals)
fitted <- data.frame(model$fitted.values)

gss$about_right_resid <- residuals$About.right
gss$about_right_fitted <- fitted$About.right


ggplot(data=gss, aes(x = about_right_resid, y = about_right_fitted)) + geom_point() +
  labs(x = "Residuals",
       y = "Predicted Probabilities",
       title = "Residuals vs Predicted Probabilities of About Right")
```




```r
gss$too_little_resid <- residuals$Too.little
gss$too_little_fitted <- fitted$Too.little

ggplot(data=gss, aes(x = too_little_resid, y = too_little_fitted)) + geom_point() +
  labs(x = "Residuals",
       y = "Predicted Probabilities",
       title = "Residuals vs Predicted Probabilities of Too Little")
```




```
gss$too_much_resid <- residuals$Too.much
gss$too_much_fitted <- fitted$Too.much

ggplot(data=gss, aes(x = too_much_resid, y = too_much_fitted)) + geom_point() +
  labs(x = "Residuals",
       y = "Predicted Probabilities",
       title = "Residuals vs Predicted Probabilities of Too Much")
```



You can change the size of your plots, so you can fit multiple plots on a single page. Include the arguments fig.height = and fig.width = in the header of the code chunk to change the plot size. See Using R Markdown for an example.

13. To examine the residuals versus each categorical predictor, you will look at the average residuals for each each category of the categorical variables.

- For each category of natmass, calculate the average residuals across categories of region.

```r
table(gss$region)
```

```
##
## E. nor. central E. sou. central Middle atlantic        Mountain     New england
##             458             174             283             218             161
##         Pacific  South atlantic W. nor. central W. sou. central
##             365             499             171             261
```

```r
regions <- c("Pacific", "South atlantic", "W. nor. central", "W. sou. central",
             "E. nor. central", "E. sou. central", "Middle atlantic", "Mountain",
             "New england")

for (val in regions) {
  temp <- filter(gss, region == val)
  print(val)
  print("About Right:")
  print(mean(temp$about_right_resid))
  print("Too Little:")
  print(mean(temp$too_little_resid))
  print("Too much:")
  print(mean(temp$too_much_resid))
}
```

```
## [1] "Pacific"
## [1] "About Right:"
## [1] -4.33347e-08
## [1] "Too Little:"
## [1] 2.79303e-08
## [1] "Too much:"
## [1] 1.54044e-08
## [1] "South atlantic"
## [1] "About Right:"
## [1] 3.773418e-08
## [1] "Too Little:"
## [1] -2.89951e-08
## [1] "Too much:"
## [1] -8.739078e-09
## [1] "W. nor. central"
## [1] "About Right:"
## [1] 5.655719e-08
## [1] "Too Little:"
## [1] 2.364564e-07
## [1] "Too much:"
## [1] -2.930136e-07
## [1] "W. sou. central"
## [1] "About Right:"
## [1] -3.390278e-07
## [1] "Too Little:"
## [1] 1.689337e-07
## [1] "Too much:"
## [1] 1.700941e-07
## [1] "E. nor. central"
## [1] "About Right:"
```

```
## [1] -2.293689e-08
## [1] "Too Little:"
## [1] -1.813678e-09
## [1] "Too much:"
## [1] 2.475057e-08
## [1] "E. sou. central"
## [1] "About Right:"
## [1] 1.171416e-07
## [1] "Too Little:"
## [1] -1.519115e-08
## [1] "Too much:"
## [1] -1.019505e-07
## [1] "Middle atlantic"
## [1] "About Right:"
## [1] 5.421018e-09
## [1] "Too Little:"
## [1] 8.466981e-08
## [1] "Too much:"
## [1] -9.009082e-08
## [1] "Mountain"
## [1] "About Right:"
## [1] -8.594236e-08
## [1] "Too Little:"
## [1] 1.811766e-07
## [1] "Too much:"
## [1] -9.523425e-08
## [1] "New england"
## [1] "About Right:"
## [1] 1.006193e-07
## [1] "Too Little:"
## [1] 5.611203e-08
## [1] "Too much:"
## [1] -1.567313e-07
```

Based on the plot and table above, discuss with your group whether there are any obvious violations of the linearity assumption. Note that we haven't examined all of the plots and tables of the residuals needed to make an assessment about the linearity assumption.

The other assumptions are randomness and independence. Discuss with your group whether these assumptions are satisfied for this analysis.

Part IV: Using the Model

16. Use your model to describe the relationship between one's political views and their attitude towards spending on mass transportation.

Using my model, I can see that when someone views are more liberal, they are more likely to want to spend more on mass transportation, while someone who is conservative is more likely to either think we spend enough on transportation or too much.

17. Use your model to predict the category of `natmass` for each observation in your dataset. Display a table of the actual versus the predicted `natmass`. What is the misclassification rate?

```r
pred <- predict(model, newdata = gss)

confusionMatrix(gss$natmass, pred)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    About right Too little Too much
##   About right        1151        219        2
##   Too little          645        340        0
##   Too much            196         36        1
##
## Overall Statistics
##
##                Accuracy : 0.5761
##                  95% CI : (0.5568, 0.5952)
##     No Information Rate : 0.7691
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1607
##
##  Mcnemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##                      Class: About right Class: Too little Class: Too much
## Sensitivity                      0.5778            0.5714       0.3333333
## Specificity                      0.6304            0.6767       0.9103208
## Pos Pred Value                   0.8389            0.3452       0.0042918
## Neg Pred Value                   0.3095            0.8411       0.9991515
## Prevalence                       0.7691            0.2297       0.0011583
## Detection Rate                   0.4444            0.1313       0.0003861
## Detection Prevalence             0.5297            0.3803       0.0899614
## Balanced Accuracy                0.6041            0.6241       0.6218271
```

The misclassification rate of the model is 1 - 0.5761 = 0.4293, or 42.93%