Chapter 4 Guided 7

Professor George

5/5/2021

A

```
library(pscl)
library(tidyverse)
library(ggplot2)
```

We start by reading in our data and making a relative requency plot of the number of pages posted.

```
amb = read_csv("G:/My Drive/Cornell College/Cornell Classes/STA 355/Materials Used in Class/data/ambigu
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   name = col_character(),
##   ambiguity = col_double(),
##   distID = col_double(),
##   ideology = col_double(),
##   totalIssuePages = col_double(),
##   democrat = col_double(),
##   mismatch = col_double(),
##   incumbent = col_double(),
##   demHeterogeneity = col_double(),
##   attHeterogeneity = col_double(),
##   distLean = col_double()
## )
```

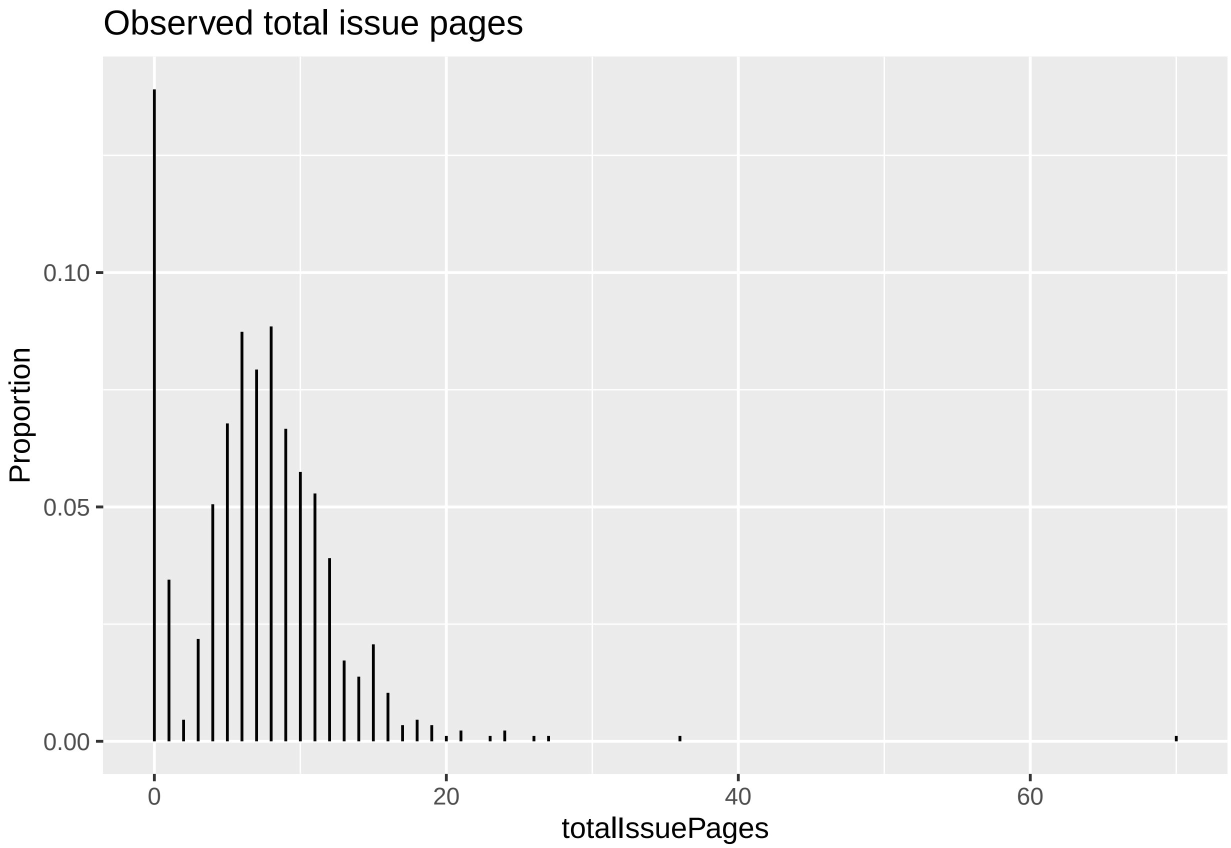
```
names(amb)
```

```
##  [1] "name"             "ambiguity"        "distID"           "ideology"
##  [5] "totalIssuePages"  "democrat"         "mismatch"         "incumbent"
##  [9] "demHeterogeneity" "attHeterogeneity" "distLean"
```

```
p.table = group_by(amb,totalIssuePages)%>%
  summarise(n=n())%>%
  mutate(prop = n/sum(n))

plot1 = ggplot(p.table, aes(x = totalIssuePages, xend = totalIssuePages,
          y = 0, yend = prop)) +
          geom_segment() +
          labs(y = "Proportion", title = "Observed total issue pages")
plot1
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```



Here we can see that there is still a large number of counts at 0 (people who did not post) but also a large number at 1. After 1, then it looks about poisson.

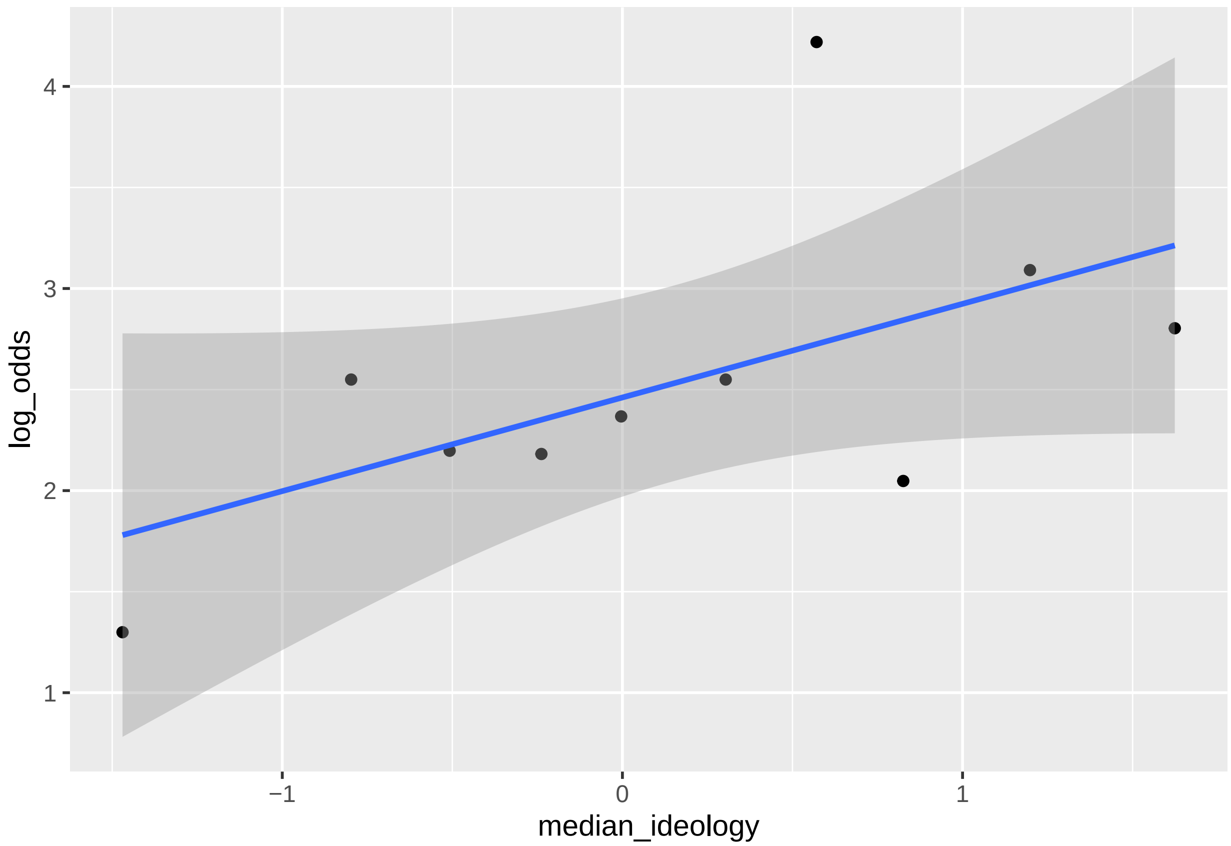
B

We are checking if using a hurdle model, that will will be essentially a poisson model to explain those posting more 1 or more pages, and a logistic model to model the odds of someone posting, is reasonable by checking the linearity assumption of logistic regression. Recall, this is checked by graphing the log(odds) vs a continuous predictor and looking for a linear relationship.

```
logodds_table <- amb %>%
  # Remove rows with either an NA for ideology or an NA for totalIssuePages
    filter(!is.na(ideology), !is.na(totalIssuePages)) %>%
  # Create an indicator for if they had 1 or more
    mutate(one_or_more_issues = totalIssuePages > 0,
           # make bins of ideology (it is on a continuous skill)
    ideology_groups = cut_number(ideology, 10)) %>%
           # group by our new bins
    group_by(ideology_groups) %>%
          # Now we calculate the mean, odds and log odds
          # of one_or_more_issues we defined above, essentially
          # allowing us to investigate better "of those who post"
    summarise(p = mean(one_or_more_issues),
              odds = p / (1-p),
              log_odds = log(odds),
              median_ideology = median(ideology))

ggplot(data = logodds_table, aes(x = median_ideology,
                y = log_odds)) +
                geom_point() +
                geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In logistic regression we assume that the log of the odds is linearly related to our predictors. Thus on this plot we are looking for a linear line. This looks good.

C

Next we are checking the assumption for our poisson part of the model that ideology (again on a contnous scale) is linearly realted to the log of the number of pages. But, recall for this part of the model we are only modeling the non-zero part, i.e. only those that posted at least once. We also compare that assumption within each party.

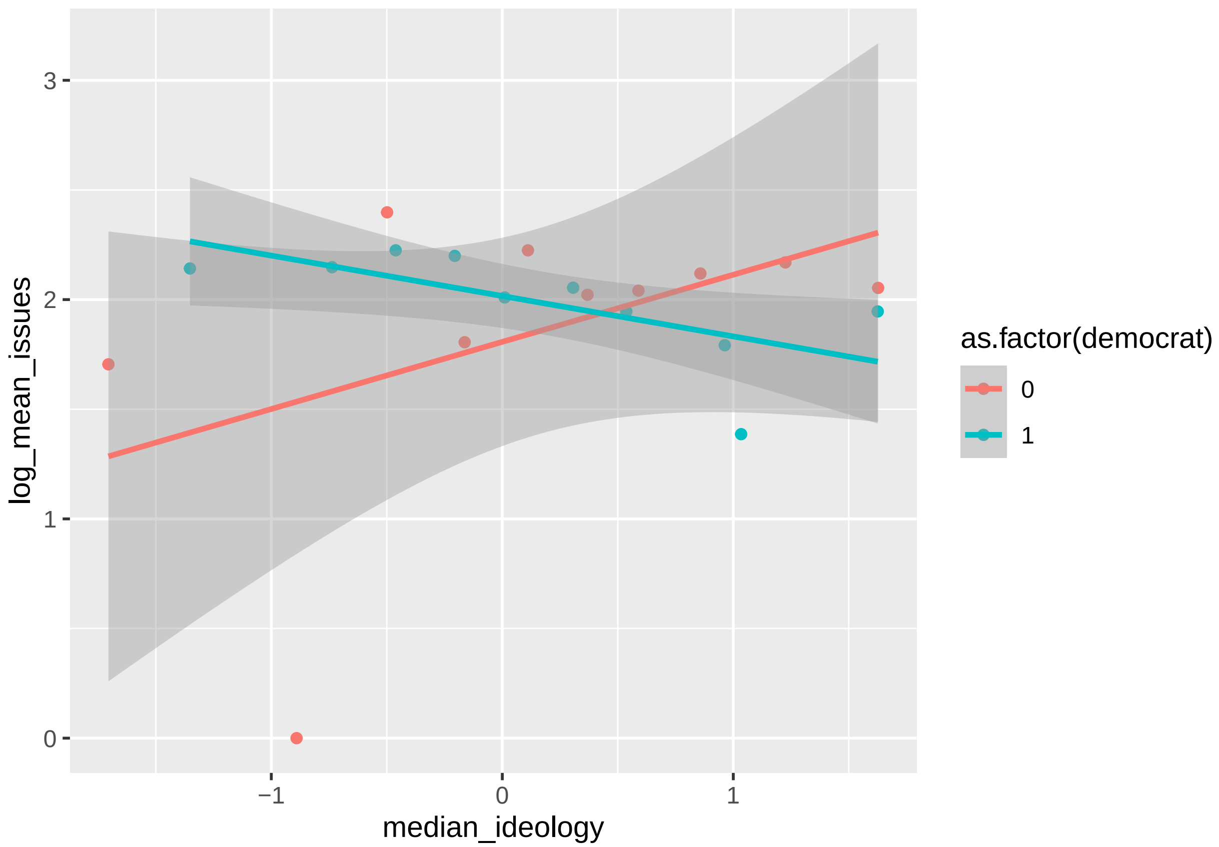
```
logissue_table <- amb %>%
  # First 3 lines same as above, look there
    filter(!is.na(ideology), !is.na(totalIssuePages),
    totalIssuePages > 0) %>%
    mutate(ideology_groups = cut_number(ideology, 10)) %>%
  # Now we want to split by both party and our new bins of ideology,
  #   that we called ideology groups
    group_by(democrat, ideology_groups) %>%

  # Now we find the log of the mean of the number of pages posted
  #  we use the median on ideology instead of mean to avoid skewness issues
    summarise(log_mean_issues = log(mean(totalIssuePages)),
        median_ideology = median(ideology))
```

```
## `summarise()` has grouped output by 'democrat'. You can override using the `.groups` argument.
```

```
ggplot(data = logissue_table, aes(x = median_ideology,
          y = log_mean_issues, color = as.factor(democrat))) +
        geom_point() +
        geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



This plot is concerning for republicans. Their points no where near fall on a line. A poisson model may work for democrats here but not republicans. We can attempt to handle this issue in our model by including an interaction term between party and ideology.

D

We now want to fit the hurdle model using the *hurdle* function.

```
hm = hurdle(totalIssuePages ~ democrat+ideology|democrat+ideology,
            data = amb,dist="poisson",zero="binomial")
summary(hm)
```

```
##
## Call:
## hurdle(formula = totalIssuePages ~ democrat + ideology | democrat + ideology,
##     data = amb, dist = "poisson", zero.dist = "binomial")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.5897 -0.7846 -0.1034  0.7202 16.9177
##
## Count model coefficients (truncated poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.093167   0.026736  78.289   <2e-16 ***
## democrat     0.041379   0.040008   1.034    0.301
## ideology    -0.005902   0.020820  -0.283    0.777
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1266     0.2247   9.465  < 2e-16 ***
## democrat      0.4279     0.3494   1.225 0.220683
## ideology      0.5746     0.1667   3.446 0.000568 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 8
## Log-likelihood: -2131 on 6 Df
```

```
exp(coef(summary(hm))$count)
```

```
##              Estimate Std. Error      z value Pr(>|z|)
## (Intercept) 8.1105622   1.027097 1.001095e+34 1.000000
## democrat    1.0422470   1.040820 2.813012e+00 1.351232
## ideology    0.9941155   1.021039 7.531697e-01 2.174547
```

```
exp(coef(summary(hm))$zero)
```

```
##             Estimate Std. Error      z value Pr(>|z|)
## (Intercept) 8.386104   1.251933 12894.171722 1.000000
## democrat    1.534086   1.418239     3.403189 1.246929
## ideology    1.776350   1.181414    31.387915 1.000568
```

Interpret Ideology Poisson Part

After exponentiation, .9941155

For each unit increase in ideology (going toward conservatism), and for candidates with at least one page posted, we see a .01% decrease in the number of pages posted on average, holding party constant.

Interpret Ideology Logistic Part

After exponentiation, 1.7763

For each unit increase in ideology (going toward conservatism), and for candidates with at least one page posted, the odds a candidate has at least one page posted increased by 77.6%, holding party constant.

E

Now we add the interaction term.

```
hm2 = hurdle(totalIssuePages ~
            democrat+ideology+democrat:ideology|
              democrat+ideology+democrat:ideology,
            data = amb,dist="poisson",zero="binomial")
summary(hm2)
```

```
##
## Call:
## hurdle(formula = totalIssuePages ~ democrat + ideology + democrat:ideology |
##     democrat + ideology + democrat:ideology, data = amb, dist = "poisson",
##     zero.dist = "binomial")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.90849 -0.73294 -0.07367  0.66102 16.65303
##
## Count model coefficients (truncated poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.05817    0.03315  62.087   <2e-16 ***
## democrat           0.05736    0.04135   1.387   0.1654
## ideology           0.03387    0.03006   1.127   0.2598
## democrat:ideology -0.07592    0.04141  -1.833   0.0668 .
## Zero hurdle model coefficients (binomial with logit link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.8129     0.2319   7.819 5.34e-15 ***
## democrat            0.3581     0.3175   1.128 0.259249
## ideology            1.3667     0.2795   4.890 1.01e-06 ***
## democrat:ideology  -1.3995     0.3731  -3.751 0.000176 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Number of iterations in BFGS optimization: 9
## Log-likelihood: -2122 on 8 Df
```

```
coef(hm2)
```

```
##       count_(Intercept)          count_democrat          count_ideology
##              2.05817000              0.05735848              0.03387103
## count_democrat:ideology        zero_(Intercept)           zero_democrat
##             -0.07592323              1.81293396              0.35814972
##           zero_ideology  zero_democrat:ideology
##              1.36674512             -1.39947172
```

```
exp(coef(hm2))
```

```
##       count_(Intercept)          count_democrat          count_ideology
##               7.8316248               1.0590354               1.0344512
## count_democrat:ideology        zero_(Intercept)           zero_democrat
##               0.9268874               6.1284016               1.4306798
##           zero_ideology  zero_democrat:ideology
##               3.9225624               0.2467273
```

Interpret interaction in zero part of model

After exponentiation, .24672 coefficient to the interaction

Note: For republicans both the variable democrat and the interaction are zero so the estimated change in the response comes from increaseing only the ideology by 1.

For each 1 unit increase in ideology (more conservative), the odds a Republican candidate has at least one issue page is 3.92 times greater.

To interpret for democrats we have to consider how being a democrats (thus zero_democrat is a 1 and is not changing the response), what is the expected change in the response. We can get this by seeing that if democrat is 1, then we get

$$logit(\hat{\alpha}) = 1.813 + 0.358 * democrat + 1.367 \cdot ideology - 1.399 \cdot ideology * democrat$$

$$logit(\hat{\alpha}) = 1.813 + .0358 + 1.367 \cdot ideology - 1.399 \cdot ideology$$

Thus if you are a democrat then the change in the log of odds will be 1.367-1.99 = -.623, the combination of the coefficients.

Then the change in the odds is

```
exp(-1.3995+1.3667)
```

```
## [1] 0.9677321
```

For each 1 unit increase in ideology (more conservative), the odds a democratic candidate candidate has at least one paged issued is changed by a factor of .96. Note that the odds are then decreasing rather than increasing.

F

This part you will work with a group and try and find the best model in terms of AIC and BIC (functions AIC and BIC in R). Try to table these after each model so at the end you have a dataframe with 3 columns: model name, AIC, and BIC