Problem Set #2

Francisco Brady

2024-09-25

1.

For this question, use the dataset anes2016subet. The dependent variable, HomeOwnership, consists of four categories: "pay rent", "pay mortgage", "own home with no payments", and "some other arrangement". The independent variables are: Age in years; BAplus, which is a dummy variable in which 1 indicates the person has a degree from a four-year college; and Ideology, which is the person's self-reported ideology on a seven-point scale in which higher values indicate more conservative beliefs.

```
# haven package function to read in dta
# use as_factor to convert labelled values to R factors
anes <- read_dta('anes2016subset.dta') %>%
  mutate(HomeOwnership = as_factor(HomeOwnership)) %>%
  mutate(Health = as_factor(Health))
# check out variables
# glimpse(anes %>% select(HomeOwnership, Age, BAplus, Ideology) %>% head)
```

(a) Estimate a multinomial logit model with "pay mortgage" as the base category. Report the results and describe the basic substantive findings without calculating predicted or marginal effects. R users should convert the Ideology variable to a numeric variable starting at 0 and may want to do the same with BAplus.

```
# relevel to make Pay Mortgage the first level
anes <- anes %>% mutate(HomeOwnership = forcats::fct_relevel(HomeOwnership, 'Pay mortgage'))

model <- multinom(HomeOwnership ~ Age + BAplus + Ideology, data = anes)
```

```
## # weights:  20 (12 variable)
## initial  value 4096.499837
## iter  10 value 3534.125827
## iter  20 value 3200.116730
## final  value 3199.842631
## converged
```

```
#model
summary(model)
```

```
## Call:
## multinom(formula = HomeOwnership ~ Age + BAplus + Ideology, data = anes)
##
## Coefficients:
##                               (Intercept)         Age     BAplus     Ideology
## Pay rent                        2.1527137 -0.03485046 -0.6994160 -0.221239320
## Own home with no payments due  -4.0332395  0.06038201 -0.2468222 -0.003922784
## Some other arrangement          0.7939304 -0.04114991 -1.1340330 -0.095117317
##
## Std. Errors:
```

```
##                                 (Intercept)         Age     BAplus   Ideology
## Pay rent                          0.1767045 0.003043336 0.09598372 0.03121937
## Own home with no payments due     0.2603623 0.003760010 0.10566499 0.03410612
## Some other arrangement            0.2735955 0.005083074 0.16926123 0.05144428
##
## Residual Deviance: 6399.685
## AIC: 6423.685
```

```
# model %>%
#   z-scores
round(summary(model)$coefficients / summary(model)$standard.errors, 3)
```

```
##                                 (Intercept)     Age BAplus Ideology
## Pay rent                             12.183 -11.451 -7.287   -7.087
## Own home with no payments due       -15.491  16.059 -2.336   -0.115
## Some other arrangement                2.902  -8.095 -6.700   -1.849
```

- Age:
  - Pay Rent: As people get older, the probability that they are paying rent decreases relative to paying a mortgage. This estimate is statistically significant.
  - Own home with no payments due: As people age they are more likely to own their home without payments relative to paying a mortgage. This estimate is significant.

  - Some other arrangement: As people get older, the probability that a person has some sort of other arrangement for their home decreases, relative to paying a mortgage. This estimate is significant.
- BAPlus:
  - Pay rent/Own home without payments/Some other arrangement: All of the coefficients for this variable are negative, indicating that a person with a BA or more is more likely to be paying a mortgage, relative to all of the other living situations. Paying rent and some other arrangement are both highly significant, while owning a home with no payments is significant, but only at the p<.05 level.
- Ideology:
  - Pay rent: People who identify as more conservative are less likely to be paying rent as opposed to paying a mortgage. This estimate is highly significant.

  - Paid off home: The coefficient implies that relative to paying a mortgage, ideology has a very small negative effect, however it is not significant.

  - Some other arrangement: This estimate implies that people who are more conservative on the ideology scale are less likely to be in some other living arrangement, relative to paying a mortgage. The estimate is significant at the 10% level.

(b) Show how the χ^2 statistic from the Likelihood Ratio test is calculated. What is the substantive interpretation of this test?

The likelihood ratio test is calculated by taking the log likelihood of the restricted model (0 explanatory variable), and comparing that to the final converged model (with all desired coefficients). Based on the model output, the restricted model log likelihood is 4096.499837, and the unrestricted log likelihood is 3199.842631:

$$LR = -2ln\frac{L_r}{L_ur} = -2(lnL_r - lnL_ur)$$

$$LR = -2(-4096.499837 - 3199.842631)$$

$$LR = 1793.314$$

The log likelihood ratio measures the explanatory power of an intercept-only model as compared to the unrestricted model which includes explanatory variables. The difference between the intercept only model and the model with our explanatory variables is significantly large that we can reject the null hypothesis that the intercept-only model has more explanatory power.

(c) Find the following by hand using the formulas. Suppose a person is 55, has a college degree, and is a moderate (3) on the ideology scale. With what predicted probabilities is the person in each home ownership category?

Find $exp(x_i\beta_j)$ for non-base categories:

$$\begin{aligned} &\text{Pay rent :}\\ x_i\beta_2 &= exp(2.153 - 0.035(55) - 0.699(1) - 0.22(3))\\ x_i\beta_2 &= exp(-1.131)\\ x_i\beta_2 &= 0.322 \end{aligned}$$

$$\begin{aligned} &\text{Own home with no payments due :}\\ x_i\beta_3 &= exp(-4.033 + 0.06(55) - 0.247(1) - 0.004(3))\\ x_i\beta_3 &= exp(-0.992)\\ x_i\beta_3 &= 0.37 \end{aligned}$$

$$\begin{aligned} &\text{Some other arrangement :}\\ x_i\beta_4 &= exp(0.79 - 0.041(55) - 1.134(1) - 0.095(3))\\ x_i\beta_4 &= exp(-2.884)\\ x_i\beta_4 &= 0.06 \end{aligned}$$

$$\begin{aligned} &\text{Pay mortgage :}\\ x_i\beta_1 &= 1 \end{aligned}$$

Then use $exp(x_i\beta_j)$'s to evaluate the probability for each category:

$$\begin{aligned} &\text{Pay rent :}\\ Pr(y_i = 2|x_i) &= \frac{\exp(x_i\beta_2)}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)}\\ Pr(y_i = 2|x_i) &= \frac{.322}{1 + .322 + .37 + .06}\\ Pr(y_i = 2|x_i) &= 0.18379 \end{aligned}$$

$$\begin{aligned} &\text{Own home with no payments due :}\\ Pr(y_i = 3|x_i) &= \frac{\exp(x_i\beta_2)}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)}\\ Pr(y_i = 3|x_i) &= \frac{.37}{1 + .322 + .37 + .06}\\ Pr(y_i = 3|x_i) &= 0.2111872 \end{aligned}$$

$$\begin{aligned} &\text{Some other arrangement :}\\ Pr(y_i = 4|x_i) &= \frac{\exp(x_i\beta_2)}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)}\\ Pr(y_i = 4|x_i) &= \frac{.06}{1 + .322 + .37 + .06}\\ Pr(y_i = 4|x_i) &= 0.03424658 \end{aligned}$$

$$\text{Pays Mortgage :}$$
$$Pr(y_i = 1|x_i) = \frac{1}{1 + .322 + .37 + .06}$$
$$Pr(y_i = 1|x_i) = 0.5707763$$

(d) Using your software, find the predicted probabilities for each category for a person with median values of all the independent variables (medians from the estimation sample, which is what the built-in functions use, not the full dataset medians).

```r
# set up data frame of medians
# keep only records with non-missing values
estimation_sample <- anes %>%
  dplyr::select(HomeOwnership, Age, BAplus, Ideology) %>%
  dplyr::filter(complete.cases(.))
median_values <- data.frame(Age = median(estimation_sample$Age),
                            BAplus = median(estimation_sample$BAplus),
                            Ideology = median(estimation_sample$Ideology))
predictions(model, type = 'probs', newdata = median_values)
```

```
##
##                          Group Estimate Std. Error    z Pr(>|z|)     S  2.5 %
##  Pay mortgage                    0.4284    0.01384 31.0   <0.001 696.7 0.4013
##  Pay rent                        0.3211    0.01337 24.0   <0.001 420.8 0.2949
##  Own home with no payments due   0.1631    0.01067 15.3   <0.001 172.8 0.1422
##  Some other arrangement          0.0874    0.00815 10.7   <0.001  86.6 0.0714
##  97.5 % Age BAplus Ideology
##   0.456  51      0        3
##   0.347  51      0        3
##   0.184  51      0        3
##   0.103  51      0        3
##
## Type:  probs
## Columns: rowid, group, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, Age, BA
```

(e) Now suppose that Age is one standard deviation higher, while all other variables remain at their medians. What is the predicted change in probabilities for each category?

```r
age_sd <- sd(estimation_sample$Age)
age_median <- median(estimation_sample$Age)
new_age <- age_median + age_sd
cat('Age Median:',
    age_median,
    '\nAge SD:',
    age_sd,
    '\nNew age:',
    new_age
    )
```

```
## Age Median: 51
## Age SD: 17.41993
## New age: 68.41993
```

```r
new_values <- data.frame(Age = new_age,
                            BAplus = median(estimation_sample$BAplus),
                            Ideology = median(estimation_sample$Ideology))
predictions(model, type = 'probs', newdata = new_values)
```

```
##
##                           Group Estimate Std. Error     z Pr(>|z|)     S  2.5 %
##  Pay mortgage                    0.3849     0.0168 22.97   <0.001 385.5 0.3521
##  Pay rent                        0.1572     0.0119 13.26   <0.001 130.9 0.1340
##  Own home with no payments due   0.4196     0.0185 22.65   <0.001 374.8 0.3833
##  Some other arrangement          0.0383     0.0060  6.39   <0.001  32.5 0.0266
##  97.5 %  Age BAplus Ideology
##  0.4177 68.4      0        3
##  0.1804 68.4      0        3
##  0.4559 68.4      0        3
##  0.0501 68.4      0        3
##
## Type:  probs
## Columns: rowid, group, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, Age, BA
```

Change in predicted probabilities (Initial - New Age)

- Pays Mortgage: $0.5707763 - 0.3849 = 0.1858763$
- Pays Rent: $0.18379 - 0.1572 = 0.02659$
- Own home with no payments due: $0.2111872 - 0.4196 = -0.2084128$
- Some other arrangement: $0.03424658 - 0.0383 = -0.00405342$

(f) Using your software, what is the average marginal effect of the variable BAplus?

```
ame_mlogit <- avg_slopes(model,
                         variables = 'BAplus',
                         type = 'probs',
                         slope = 'dydx')
ame_mlogit
```

```
##
##                           Group  Estimate Std. Error        z Pr(>|z|)    S
##  Pay mortgage                   1.36e-01    0.01811  7.52601   <0.001 44.1
##  Pay rent                      -8.77e-02    0.01560 -5.62007   <0.001 25.6
##  Own home with no payments due  2.05e-05    0.01375  0.00149    0.999  0.0
##  Some other arrangement        -4.87e-02    0.00909 -5.35551   <0.001 23.5
##    2.5 %  97.5 %
##   0.1008  0.1718
##  -0.1182 -0.0571
##  -0.0269  0.0270
##  -0.0665 -0.0309
##
## Term: BAplus
## Type:  probs
## Comparison: mean(1) - mean(0)
## Columns: term, group, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
```

According to the output:
- Pay mortgage: Having a BA+ is on average associated with a marginal effect of .136. The p-value for this estimate is significant.

- Pay rent: Having a BA+ is on average associated with a marginal effect of -.08. The p-value for this estimate is significant.
- Own home with no payments due: Having a BA+ is on average associated with a marginal effect of 0.0000205. The p-value for this estimate is not significant.
- Some other arrangement: Having a BA+ is associated with an average marginal effect of -0.0487. The p-value for this estimate is significant.

(g) Change the base category to "own home with no payments" and re-run the model. Examine the coefficients and compare them to the earlier estimation. How do you interpret the differences? [Everything is also the same statistical significance.]

```r
# relevel to make Pay Mortgage the first level
anes <- anes %>% mutate(HomeOwnership = forcats::fct_relevel(HomeOwnership,
                                                              'Own home with no payments due'))
model_2 <- multinom(HomeOwnership ~ Age + BAplus + Ideology, data = anes)
```

```
## # weights:  20 (12 variable)
## initial  value 4096.499837
## iter  10 value 3350.010798
## iter  20 value 3199.842681
## final  value 3199.842631
## converged
```

```r
summary(model_2)
```

```
## Call:
## multinom(formula = HomeOwnership ~ Age + BAplus + Ideology, data = anes)
##
## Coefficients:
##                        (Intercept)         Age     BAplus     Ideology
## Pay mortgage              4.033324 -0.06038228  0.2468275  0.003905196
## Pay rent                  6.186049 -0.09523291 -0.4525885 -0.217335822
## Some other arrangement    4.827216 -0.10153109 -0.8872081 -0.091214758
##
## Std. Errors:
##                        (Intercept)         Age    BAplus   Ideology
## Pay mortgage             0.2603646 0.003760028 0.1056653 0.03410626
## Pay rent                 0.2842282 0.004333512 0.1235039 0.03975748
## Some other arrangement   0.3530895 0.005964348 0.1866670 0.05713950
##
## Residual Deviance: 6399.685
## AIC: 6423.685
```

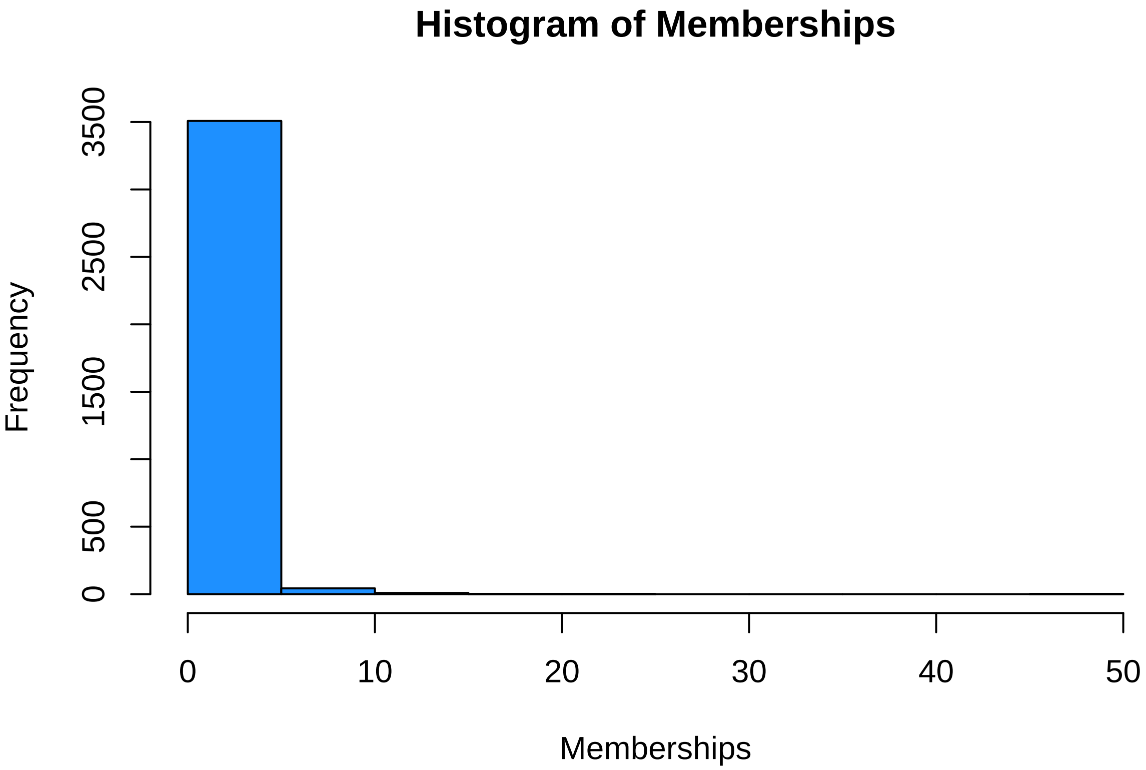
When you change the base category for a model, none of the predicted probabilities or marginal effects change. The only thing that changes is the coefficients, because they are in reference to a different base category.

2.

This question will use the dataset `anes2016subset`. The dependent variable is `Memberships`.

(a) Produce a histogram of the dependent variable. Explain why OLS might be not be the best estimation method for these data.

```r
hist(anes$Memberships, main = 'Histogram of Memberships',
     xlab = 'Memberships', col = 'dodgerblue'
     )
```



The histogram shows that the count of memberships is heavily concentrated around values less than 10, with a large proportion of observations at 0. The data are not evenly distributed, and using OLS can result in biased standard errors and incorrect significance tests.

(b) Estimate a Poisson model using the following independent variables: `Age`, `BAplus`, `Ideology`, and `NewsDays`. Report the results and describe the basic substantive findings without calculating predicted or marginal effects. Note: R users should make Ideology a numeric variable starting at 0.

```
poisson_model <- glm(
  Memberships ~ Age + BAplus + Ideology + NewsDays,
  family = "poisson",
  data = anes)
poisson_sum <- summary(poisson_model)
poisson_sum
```

```
##
## Call:
## glm(formula = Memberships ~ Age + BAplus + Ideology + NewsDays,
##     family = "poisson", data = anes)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.447499   0.078043  -5.734 9.81e-09 ***
## Age          0.002594   0.001050   2.470  0.01351 *
## BAplus       0.528994   0.034798  15.202  < 2e-16 ***
## Ideology    -0.032647   0.010825  -3.016  0.00256 **
## NewsDays     0.057457   0.010717   5.361 8.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6287.5  on 2904  degrees of freedom
## Residual deviance: 5947.5  on 2900  degrees of freedom
##   (1366 observations deleted due to missingness)
## AIC: 9847.1
##
## Number of Fisher Scoring iterations: 6
```

- Age: As a person ages, they are more likely to have a higher count of memberships. This estimate is significant at the .05 level.
- BAPlus: Having a BA or more is associated with a higher count of memberships. This estimate is highly significant.

- Ideology: An increased score on the ideology scale (more conservative) is associated with a lower count of memberships. This estimate is significant.

- A higher number of days reported spent watching the news is associated with a higher count of memberships. This estimate is highly significant.

(c) Using Stata or R, calculate the predicted number of memberships for a person who is 35 years-old, has a college degree, is liberal (1) on the ideology scale, and who watches/reads the news 5 days a week.

```
value_list <- data.frame(Age = 35,
                         BAplus = 1,
                         Ideology = 1,
                         NewsDays = 5)
prediction_1 <- predictions(poisson_model,
                            newdata = value_list,
                            type = 'response')
prediction_1
```

```
##
##  Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 % Age BAplus Ideology
##      1.53     0.0506 30.3   <0.001 666.8  1.43   1.63  35      1        1
##  NewsDays
##         5
##
## Type:  response
## Columns: rowid, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, Age, BAplus,
```

According to the estimates, a person with these characteristics is predicted to have 1.53 memberships. This is slightly higher than the overall average number of memberships for the sample.

(d) By hand, calculate the probability that this person belongs to 2 groups.

$$Pr(y_i = 2) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

$$Pr(y_i = 2) = \frac{e^{-1.53}1.53^2}{2!}$$

The probability that the person above belongs to 2 groups is: 0.253.

(e) Run the same model using a negative binomial regression. Using Stata or R, calculate the predicted number of memberships for a person with the same characteristics as those described in part (c). Compare this result to the previous prediction.

```r
mod_nbreg <-
  glm.nb(Memberships ~ Age + BAplus + Ideology + NewsDays,
         data = anes)
summary(mod_nbreg)
```

```
##
## Call:
## glm.nb(formula = Memberships ~ Age + BAplus + Ideology + NewsDays,
##     data = anes, init.theta = 0.9333956938, link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.417042   0.113547  -3.673  0.00024 ***
## Age          0.001364   0.001609   0.848  0.39667
## BAplus       0.525007   0.052467  10.006  < 2e-16 ***
## Ideology    -0.028655   0.016787  -1.707  0.08783 .
## NewsDays     0.061182   0.015580   3.927  8.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9334) family taken to be 1)
##
##     Null deviance: 3064.0  on 2904  degrees of freedom
## Residual deviance: 2918.2  on 2900  degrees of freedom
##   (1366 observations deleted due to missingness)
## AIC: 8725.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9334
##           Std. Err.:  0.0545
##
##  2 x log-likelihood:  -8713.1180
```

```r
predictions(mod_nbreg, value_list)
```

```
##
##  Estimate Pr(>|z|)    S 2.5 % 97.5 % Age BAplus Ideology NewsDays
##      1.54   <0.001 53.5  1.39   1.71  35      1        1        5
##
## Type:  invlink(link)
## Columns: rowid, estimate, p.value, s.value, conf.low, conf.high, Age, BAplus, Ideology, NewsDays, Mer
```

The negative binomial did not change the predicted estimate by very much. The predicted value for the characteristics from the previous question is 1.54.

(f) Now estimate the same model using a zero-inflated Poisson regression. In this model, use the variables `Voted2016` and `Health` as variables that predict whether a person is in the "always zero" category. Interpret the substantive meaning and statistical significance of the coefficients on these two variables. Note: R users should treat Health as a numeric variable starting at 0.

```r
anes <- anes %>% mutate(Health = as.numeric(Health))
anes <- anes %>% mutate(Voted2016 = as.numeric(Voted2016))

zip_model <- zeroinfl(Memberships ~ Age + BAplus + Ideology + NewsDays | Voted2016 + Health,
                      data = anes, dist = 'negbin')
summary(zip_model)
```

```
##
## Call:
## zeroinfl(formula = Memberships ~ Age + BAplus + Ideology + NewsDays |
##     Voted2016 + Health, data = anes, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.8701 -0.7292 -0.3563  0.4623 38.2535
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0257636  0.1241547  -0.208  0.83561
## Age          0.0008435  0.0015908   0.530  0.59598
## BAplus       0.4126152  0.0533431   7.735 1.03e-14 ***
## Ideology    -0.0387630  0.0159673  -2.428  0.01520 *
## NewsDays     0.0495931  0.0157911   3.141  0.00169 **
## Log(theta)   0.4428282  0.1426223   3.105  0.00190 **
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5865     0.3446   1.702   0.0887 .
## Voted2016    -1.6358     0.2756  -5.935 2.94e-09 ***
## Health       -0.2362     0.1004  -2.352   0.0187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 1.5571
## Number of iterations in BFGS optimization: 20
## Log-likelihood: -4162 on 9 Df
```

The negative coefficients on `Voted2016` and `Health` indicates that people who voted in 2016 and that are in better health are less likely to be in the "always zero" category (0 memberships). Both coefficients are statistically significant.

(g) Using Stata or R, calculate the predicted number of memberships for a person with the same characteristics as those described in part (c). Assume also that the person voted in 2016 and is in very good health (3).

```r
value_list <- data.frame(Age = 35,
                         BAplus = 1,
                         Ideology = 1,
                         NewsDays = 5,
                         Voted2016 = 1,
                         Health = 3)
prediction_2 <- predictions(zip_model,
                            newdata = value_list,
                            type = 'response')
```

```
prediction_2 # %>% t %>% knitr::kable()
```

```
##
##  Estimate Std. Error    z Pr(>|z|)     S 2.5 % 97.5 % Age BAplus Ideology
##      1.59     0.0847 18.8   <0.001 260.3  1.43   1.76  35      1        1
##  NewsDays Voted2016 Health
##         5         1      3
##
## Type:  response
## Columns: rowid, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high, Age, BAplus, I
```

Using the zero-inflated model yields a prediction of 1.59, which is slightly higher than the poisson model.

(h) Now estimate a Hurdle model in which the first stage is a logit and the second stage is a truncated Poisson. In Stata, this will require suest.

```
# define model
hurdle_mod <- hurdle(Memberships ~ Age + BAplus + Ideology + NewsDays |
                       Voted2016 + Health, dist = 'poisson',
                     data = anes)
summary(hurdle_mod)
```

```
##
## Call:
## hurdle(formula = Memberships ~ Age + BAplus + Ideology + NewsDays | Voted2016 +
##     Health, data = anes, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.9445 -0.8422 -0.3201  0.4940 37.7243
##
## Count model coefficients (truncated poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.458618   0.098999   4.633 3.61e-06 ***
## Age         -0.000373   0.001282  -0.291 0.771136
## BAplus       0.222293   0.042944   5.176 2.26e-07 ***
## Ideology    -0.050504   0.012775  -3.953 7.71e-05 ***
## NewsDays     0.051344   0.013773   3.728 0.000193 ***
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.13714    0.15926  -7.140 9.33e-13 ***
## Voted2016    0.90795    0.10269   8.842  < 2e-16 ***
## Health       0.12770    0.03782   3.377 0.000734 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -4348 on 8 Df
```

(i) Calculate the predicted number of memberships for a person with the same characteristics as described in part (g).

```
# use same value list as above
hurdle_pred <- predict(hurdle_mod,
                       newdata = value_list,
                              type = 'response', se.fit = TRUE)
hurdle_pred
```

```
##       1
## 1.41958
```

A person with these characteristics is expected to have 1.42 memberships.

(j) When trying to decide which of these models is most appropriate/best for this scenario, what factors do you consider?

- Consider the distribution of the data. If the outcome variable is concentrated in the 0 range, you may not be able to reliably use OLS.

- Consider whether the data represents a situation where there are two classes of outcomes: "always-zero" and cases in which 0 happened.

- Consider whether there is over- or under-dispersion in the number of events The generalized poisson model assumes an equal mean and variance of the distribution. If there is over- or under-dispersion present, one solution is to use a negative binomial distribution model.