Problem Set #2

Francisco Brady

2024-09-22

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))
# 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 %>%
#   tidy()
```

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

(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:

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

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

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

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

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

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

Some other arrangement :

$$Pr(y_i = 4|x_i) = \frac{\exp(x_i\beta_2)}{1 + \sum_{j=2}^{J} \exp(x_i\beta_j)}$$

$$Pr(y_i = 4|x_i) = \frac{.06}{1 + .322 + .37 + .06}$$

$$Pr(y_i = 4|x_i) = 0.03424658$$

[I THINK THIS MEANS PR(1) = 1 - PR(2) + PR(3) + PR(4)]

(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).

```
# set up data frame of medians
# keep only records with non-missing values
estimation_sample <- anes %>%
  select(HomeOwnership, Age, BAplus, Ideology) %>%
  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?

```
age_sd <- sd(estimation_sample$Age)
age_median <- median(estimation_sample$Age)
new_age <- age_median + age_sd
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
```

(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
```

(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?

```
# relevel to make Pay Mortgage the first level
anes <- anes %>% mutate(HomeOwnership = forcats::fct_relevel(HomeOwnership, 'Own home with no payments
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
```

```
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
```

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.

(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.

(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.

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

(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.

(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.

(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).

(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.

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

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