exercises_week9

Lindsey Greenhill

3/24/2021

Question 1

Part a

```
# have to include both women_alone and couples as treatments for the model

mod_a  <- glm(fupacts ~ women_alone + couples, data = df, family = "poisson")

# summary of the model

summary(mod_a)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + couples, family = "poisson",
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6153  -4.9615  -3.1895   0.9943  27.1999
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.08562    0.01905  162.02   <2e-16 ***
## women_alone -0.57535    0.03032  -18.98   <2e-16 ***
## couples     -0.32077    0.02741  -11.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13295  on 433  degrees of freedom
## Residual deviance: 12920  on 431  degrees of freedom
## AIC: 14248
##
## Number of Fisher Scoring iterations: 6
```

```r
# look at the deviance residuals. The min and the max should be relatively equal
# in absolute value. If they aren't, it hints at overdispersion.

# look at the deviances:  roughly 240 on 1 degree of freedom. We are way in the
# tail. So much so that we don't have to do the p chisq thing.

#ssr <- (df$fupacts - mod_a$fitted.values)^2/sqrt(mod_a$fitted.values)
#sum_ssr <- sum(ssr^2)

# k is the number of variables that you add over the mean model
# 1 df. Value is way into the tail, so there is a problem.



#over <- 1/(nrow(df) - mod_a$df.null -  mod_a$df.residual)*sum_ssr

# I'm going with the code from the book. It's more correct and makes more sense

yhat <- predict(mod_a, type = "response")
z <- (df$fupacts - yhat) / sqrt(yhat)

# cat("overdispersion ratio is", sum(z^2)/(nrow(df) - 3), "\n")

# cat("p-value of overdispersion test is", pchisq(sum(z^2), nrow(df) - 3), "\n")

# there is evidence of overdispersion
```

Model fit

- The treatment variables appears to be statistically significant
- There is a significant difference between the null deviance and the residual deviance, indicating that the model is better than the null model

Overdisperion check

There are a few ways to check for overdispersion. A quick way to check is to look at the minimum and maximum deviance residuans. If the model is a good fit, the min and max should be relatively equal in absolute value.

- The minimum deviance residual = -6.6 while the maximum deviance residual = 27.2. These two values differ substantially in absolute value and hints at overdispersion

Another way to check for overdispersion is to perform a chi squared test on the overdispersion ratio. If the overdisperion ratio is in the tail of the distribution, then there is likely overdispersion

- The overdispersion ratio is 44.2487021
- The p value of the overdispersion test is 1
- The p value is in the tail of the chi squared distribution, suggesting that there is overdispersion

Part b

In the code below I extend the model to include pre-treatment variables.

```
# creating model with pre treatment variables. It will still probably be
# overdispersed

mod_b <- glm(fupacts ~ women_alone  + sex + couples + bs_hiv + bupacts,
             data = df,
             family = "poisson")

summary(mod_b)
```

```
##
## Call:
## glm(formula = fupacts ~ women_alone + sex + couples + bs_hiv +
##     bupacts, family = "poisson", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.688   -4.287   -2.536    1.386   23.388
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.892428   0.023245 124.430  < 2e-16 ***
## women_alone    -0.666211   0.030993 -21.496  < 2e-16 ***
## sexman         -0.110992   0.023782  -4.667 3.05e-06 ***
## couples        -0.408762   0.028276 -14.456  < 2e-16 ***
## bs_hivpositive -0.439410   0.035470 -12.388  < 2e-16 ***
## bupacts         0.010799   0.000174  62.071  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13295  on 433  degrees of freedom
## Residual deviance: 10191  on 428  degrees of freedom
## AIC: 11525
##
## Number of Fisher Scoring iterations: 6
```

```
# look at the deviance residuals. The min and the max should be relatively equal
# in absolute value. If they aren't, it hints at overdispersion.

# doing the same overdispersion test as in part a

yhat_b <- predict(mod_b, type = "response")

z_b <- (df$fupacts - yhat_b) / sqrt(yhat_b)
```

Model fit

- All of the variables appear to be statistically significant

- There is a significant difference between the null deviance and the residual deviance, indicating that the model is improved from the model in part a

Overdisperion check

- The minimum and maximum deviance residuals are closer in absolute value than they were in mod_a, however, they differ by 5, hinting at overdispersion
- The overdispersion ratio is 30.0578574
- The p value of the overdispersion test is 1
- The p value is in the tail of the chi squared distribution, suggesting that there is overdispersion

Part c

In the code below I fit an overdispersed Poission model using the glm.nm() function.

```
#  fitting a negative binomial

mod_c  <- glm.nb(fupacts ~ women_alone  + sex + couples + bs_hiv + bupacts, data = df)

summary(mod_c)
```

```
##
## Call:
## glm.nb(formula = fupacts ~ women_alone + sex + couples + bs_hiv +
##     bupacts, data = df, init.theta = 0.417040419, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1476  -1.4831  -0.4510   0.1979   2.8180
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.432377   0.172386  14.110  < 2e-16 ***
## women_alone    -0.725712   0.192408  -3.772 0.000162 ***
## sexman          0.024765   0.151855   0.163 0.870455
## couples        -0.349085   0.188969  -1.847 0.064701 .
## bs_hivpositive -0.554562   0.186073  -2.980 0.002879 **
## bupacts         0.022482   0.002362   9.516  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.417) family taken to be 1)
##
##     Null deviance: 582.52  on 433  degrees of freedom
## Residual deviance: 487.52  on 428  degrees of freedom
## AIC: 2964.7
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##                Theta:  0.4170
##            Std. Err.:  0.0313
##
##  2 x log-likelihood:  -2950.6560
```

```
# really changes the  deviance residuals range. the min and max are pretty much
# equal now (abs value). Sex and couples become not statistically significant.
```

Effectiveness of intervention

We assume that the glm.nm function fixes the overdispersion problem from the poisson models above

- The coefficient estimates change from mod_a/mod_b to modc, suggesting that the new model may have been an effective intervention
- Two variables that were statistically significant in the previous models are now not statistically significant (sexman and couples), suggesting that this model may have more accurately calculated the standard errors and was thus an effective intervention

Interpretation of some coefficients

We can interprety the coefficients in the model using the formula 1/(exp(beta)), where an increase in 1 in the variables leads to a 1/(exp(beta)) change in the outcome variable

- Women_alone: Using the above formula, an increase in the women_along variables (or switching from non treatment to treatment) results in a 1/(exp(-.72)), or 2.0544332 increase in unprotected sex acts on average, holding all else constant
- bs_hivpositive: those who have a positive baseline hiv status on average engaged in 1/(exp(.02248), or 0.9777708, more unprotected sex acts, holding all else constant

Part d

I have concerns about the IID Arrivals assumption (the assumption that the observations are independent and identically distributed). I would expect the sexual activities of couples are not independent and are highly correlated. Because the data includes responses from both men and women in the same couple, the observations are probably not completely independent.

Question 2

Part a

```
# using the method from class

mod_2 <- multinom(partyid3 ~ age + educ1 + income + ideo7,
                  data = nes_clean_simple)
```

```
## # weights:  18 (10 variable)
## initial  value 616.321494
## iter  10 value 459.083499
## final  value 427.339632
## converged
```

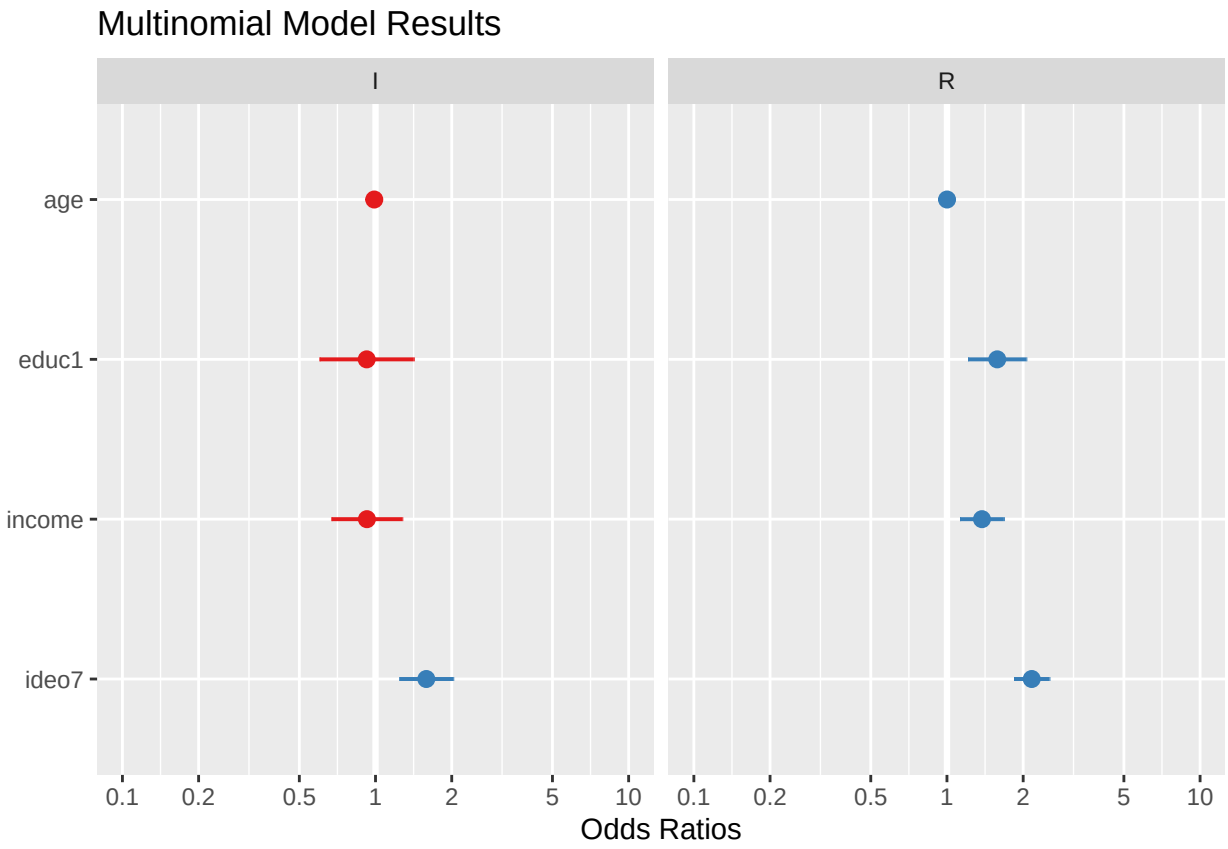
```
# using stargazer to display the model

stargazer(mod_2, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                            I              R
##                           (1)            (2)
## -----------------------------------------------
## age                      -0.012          0.001
##                         (0.010)        (0.007)
##
## educ1                    -0.080        0.457***
##                         (0.219)        (0.135)
##
## income                   -0.078        0.318***
##                         (0.164)        (0.102)
##
## ideo7                   0.462***       0.771***
##                         (0.125)        (0.082)
##
## Constant               -2.950***      -6.124***
##                         (1.028)        (0.708)
##
## -----------------------------------------------
## Akaike Inf. Crit.      874.679        874.679
## ===============================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# using the plot model function to show the coefficients for the model

plot_model(mod_2,
           type = "est",
           title = "Multinomial Model Results")
```




```
# interpret the results using the divide by 4 rule!
```

Part b

The variables have different levels of statistical significance for I and R. I will interpret the statistically significant coefficients below using the divide by 4 rule. Although this doesn't give an exact interpretation of the coefficients, it is much quicker to calculate.

Coefficient Interpretations: I

The coefficients for age, educ1, and income are not statistically significant for the Independent coefficients. Ideology is the only statistically significant variable.

- Using the divide by 4 rule, a 1 point increase on the ideology scale is associated with roughly an 11% increase on average in the likelihood of being an Independent compared to a Democrat, holding all else constant.

Coefficient Interpretations: R

The coefficients for educ1, income, and ideo7 are statistically significant for the Republican coefficients. Age is the only non statistically significant variable. In summary, the effects of education, income, and ideology have greater impacts on the likelihood of being a Republican compared to a Democrat verses the likihood of being an Independent compared to a Democrat.

- Using the divide by 4 rule, a 1 point increase on the ideology scale is associated with roughly an 19% increase on average in the likelihood of being a Republican compared to a Democrat, holding all else constant.
- Using the divide by 4 rule, a 1 point increase on the education scale is associated with roughly an 12% increase on average in the likelihood of being a Republican compared to a Democrat, holding all else constant.
- Using the divide by 4 rule, a 1 point increase on the income scale is associated with roughly a 8% increase on average in the likelihood of being a Republican compared to a Democrat, holding all else constant.

Part c

For part c, I chose two test cases from the data frame to make predictions off of

```
# this row is someone who is 63, has some college, 34 to 67th percentile income, and is slightly conser

pred_df <- nes_clean_simple %>%
  slice(1) %>%
  select(-partyid3)

# The model incorrectly predicts this person to be R

predict(mod_2, pred_df)
```

```
## [1] R
## Levels: D I R
```

```
# case 2. This person is 40, has some college, 34 to 67th percentile income, and is conservative. They

pred_df_2 <- nes_clean_simple %>%
  slice(2)  %>%
  select(-partyid3)

# The model correctly predicts this person is R

predict(mod_2, pred_df_2)
```

```
## [1] R
## Levels: D I R
```

Test case 1

- This prediction is for someone who is 63, has some college education, is in the 34 to 67th percentile income, and is slightly conservative.
- The model predicts that thiis person is a Republican
- The model is incorrect – this person was actually a Democrat

Test case 2

- This prediction is for someone who is 40, has some college education, is in the 34 to 67th percentile income, and is conservative.
- The model predicts that this person is a Republican
- The model is correct – this person was actually a Republican