Group8_Alcohol_use

Yichao Chen

12/3/2019

Introduction

Our group mainly focus on the relationship between eating habit and alcohol use of people. The relationship between some demographic variables and alcohol is also analyzed. Zero inflation negative binomial regression model is used in our analysis process. The reason of choosing this model is explained below and the final results of our model will be illuminate as well.

data

Three datasets from NHANES 2005-2006 are used:
1.Alcohol Use (ALQ_D.xpt)
2.Diet Behavior & Nutrition (DBQ_D.xpt)
3.Demographic Variables & Sample Weights (DEMO_D.xpt)

From these three datasets, we choose the following variable: ALQ130 - Avg # alcoholic drinks/day -past 12 mos (1-32:range of values)
DBQ700 - How healthy is the diet (1-5:Excellent-poor)
DBD091 - # of times/wk eat meals not from a home (1-21:range of values)
RIAGENDR - Gender (1:Male 2:Female)
RIDAGEYR - Age at Screening Adjudicated - Recode (0-84:range of values)
INDFMPIR - Family PIR (0-5:a ratio of family income to poverty threshold)

Here we only focus on adults and select samples with age>=21.

Methods

R (data.table,pscl)

Analysis

Data Cleaning

Firstly, the datasets are cleaned. All the missing values are deleted and three datasets are joined. Here, the package of data.table is used. Details could be seen in comments. In this process, we rename the variables for convenience: ALQ130:alq_drink, DBD091: meal_out, DBQ700:diet, RIAGENDR :gender, RIDAGEYR:age, INDFMPIR:pir.

```
library(foreign)
library(data.table)
AL=read.xport('./ALQ_D.xpt')
write.csv(AL,file='./ALQ_D.csv')
DBQ=read.xport('./DBQ_D.xpt')
write.csv(DBQ,file='./DBQ_D.csv')
```

```
DEMO=read.xport('./DEMO_D.xpt')
write.csv(DEMO,file='./DEMO_D.csv')
ALQ_D= fread('./ALQ_D.csv')
DBQ_D=fread('./DBQ_D.csv')
DEMO_D=fread('./DEMO_D.csv')
# delete missing values of ALQ30 ane rename ALQ30 as alq_drink
# In the original data <1 drink  are recorded as 1, we replace 1 with 0 here
A1=ALQ_D[ALQ130!=''&ALQ130!=999,.(SEQN,alq_drink=ALQ130)][alq_drink==1,alq_drink:=0]
# delete missing values of DBD091 and DBQ700
# DBD091:5555(representing >21) consider as 21; 6666(representing <1) consider as 0
# rename DBD091 as meal_out; rename DBQ700 as diet
DBQ1=DBQ_D[DBD091!=''&DBD091!=7777&DBD091!=9999&DBQ700!=''&DBQ700!=7&DBQ700!=9][
  DBD091==6666,DBD091:=0][DBD091==5555,DBD091:=21][
    ,.(SEQN,meal_out=DBD091,diet=DBQ700)]
# delete missing values of RIAGENDR,RIDAGEYR,INDFMPIR
# only focus on adults(age>=21)
# rename RIAGENDR as gender;RIDAGEYR as age; INDFMPIR as pir
DEMO1=DEMO_D[RIAGENDR!=''&RIDAGEYR!=''&RIDAGEYR>=21&INDFMPIR!=''][
  RIAGENDR==2,RIAGENDR:=0][,.(SEQN,gender=RIAGENDR,age=RIDAGEYR,pir=INDFMPIR)]
#join these three datasets together according to SEQN
data=A1[DBQ1,on='SEQN',nomatch=0L][DEMO1,on='SEQN',nomatch=0L]
```

Basic Analysis

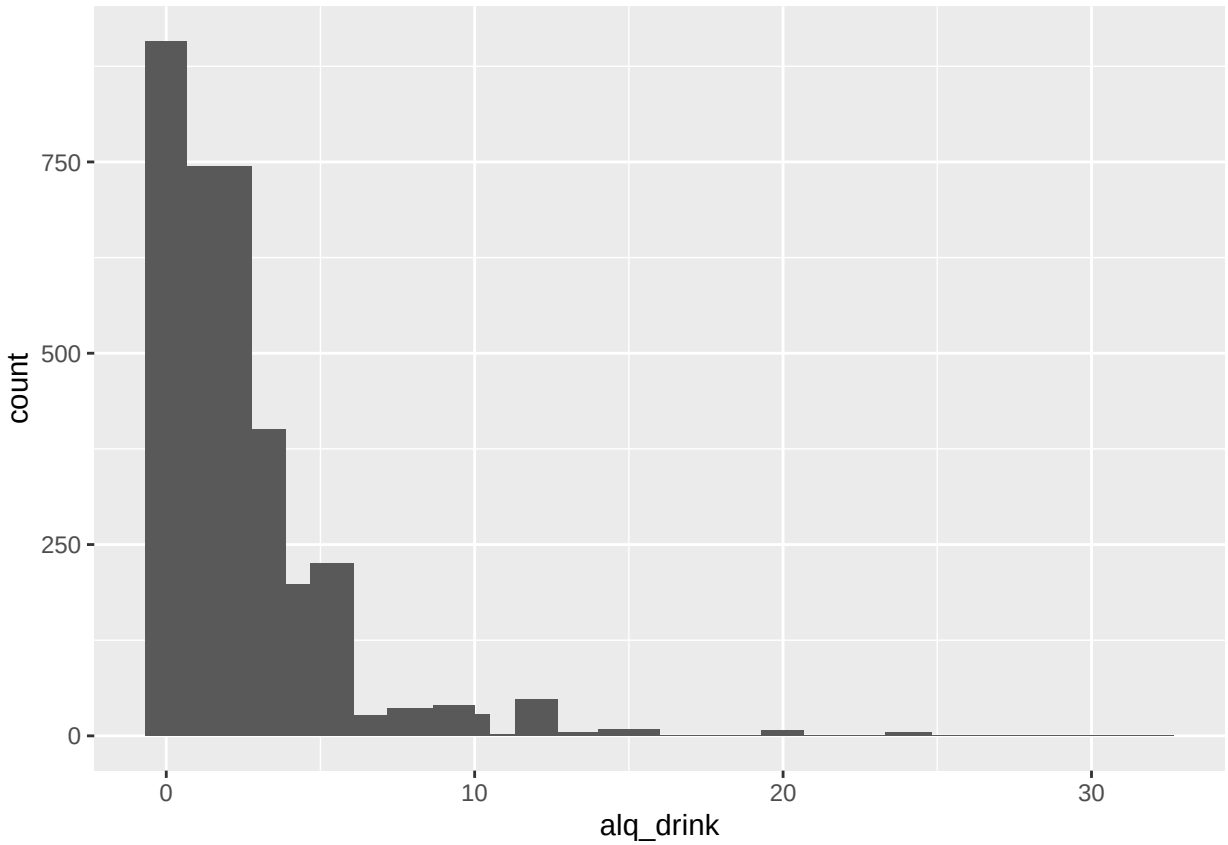
As observed, alq_drink is count variable, this implys that we may use possion regression or negative binomial regression, we then do some basic analysis of our response:alq_drink.

```
library(ggplot2)
#calculate mean and variance
sprintf("Mean and Variance = %1.2f and %1.2f",
        mean(data$alq_drink), var(data$alq_drink))
```

```
## [1] "Mean and Variance = 2.51 and 8.89"
```

```
# histogram with alq_drink
ggplot(data,aes(alq_drink))+geom_histogram()+stat_bin(bins=25)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the result, alq_drink is overdispersed as its variance is not basiclly equal to its mean.So, the negative binomial regression is more suitable in our data. From the plot, we could see that a large part of people drink alcohol less than 1. The excess zeros may be generated by a separate process from the count value alq_drink and can be modeled independently.Some people may never drink for some other reasons. So, zero-inflation model is considered.

Build Model

The package of pscl is used for zero-inflated negative binomial regression. The variables of diet, gender, age and pir are used in the part of negative binomial model and the variable of meal_out is used in the logit part of the model.

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
m1=zeroinfl(alq_drink~diet+gender+age+pir|meal_out,data=data,dist='negbin',EM=TRUE)
summary(m1)
```

```
##
## Call:
## zeroinfl(formula = alq_drink ~ diet + gender + age + pir | meal_out,
##     data = data, dist = "negbin", EM = TRUE)
```

```
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.3608 -0.8713 -0.1232  0.5076  9.0189
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.547094   0.089043  17.375  < 2e-16 ***
## diet         0.079589   0.017896   4.447 8.69e-06 ***
## gender       0.543883   0.037670  14.438  < 2e-16 ***
## age         -0.016147   0.001229 -13.136  < 2e-16 ***
## pir         -0.097166   0.010993  -8.839  < 2e-16 ***
## Log(theta)   1.497940   0.090054  16.634  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.05652    0.09302 -11.358  < 2e-16 ***
## meal_out    -0.05516    0.01886  -2.924  0.00345 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 4.4725
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -5172 on 8 Df
```

```
m0 <- update(m1, . ~ 1)
pchisq(2 * (logLik(m1) - logLik(m0)), df = 5, lower.tail=FALSE)
```

```
## 'log Lik.' 1.224649e-107 (df=8)
```

From the output of chi-squared test, we know that our overall model is statistically significant.
From the result of m1, the variables of diet, gender, age and pir in the part of negative binomial are all significant predictors. The variable of meal_out in the part of the logit model predicting excessive zero is also statistically significant.
The expeceted change in log(alq_drink) for one-unit increase in diet is 0.079589 holding other variable constant. From the codebook, the larger diet factor indicate poorer diet behavior. The model shows that poorer diet may related to more alcohol use.
When gender change from 0 to 1, the change in log(alq_drink) is 0.544, men tends to drink more than women.
The expeceted change in log(alq_drink) for one-unit increase in age is -0.016 holding other variable constant. This means when age increase, people tend to drink alcohol less.
The expeceted change in log(alq_drink) for one-unit increase in age is -0.097 holding other variable constant, which means family with better finacial situation might use alcohol less.
The log odds of being an excessive zero will decrease by 0.05516 for every one more meal eating outside. This means when the frequency of eating out of home is larger, the zero are less likely comes from the part of people who never use alcohol. In other words, more meals eating out of home implies more alcohol use.

We could also build a negative bionomial regression model and compares its result with the result above. The package of MASS is used for building negative bionomial regression model.

```
library(MASS)
m2=glm.nb(alq_drink~diet+gender+age+pir,data=data)
vuong(m1, m2)
```

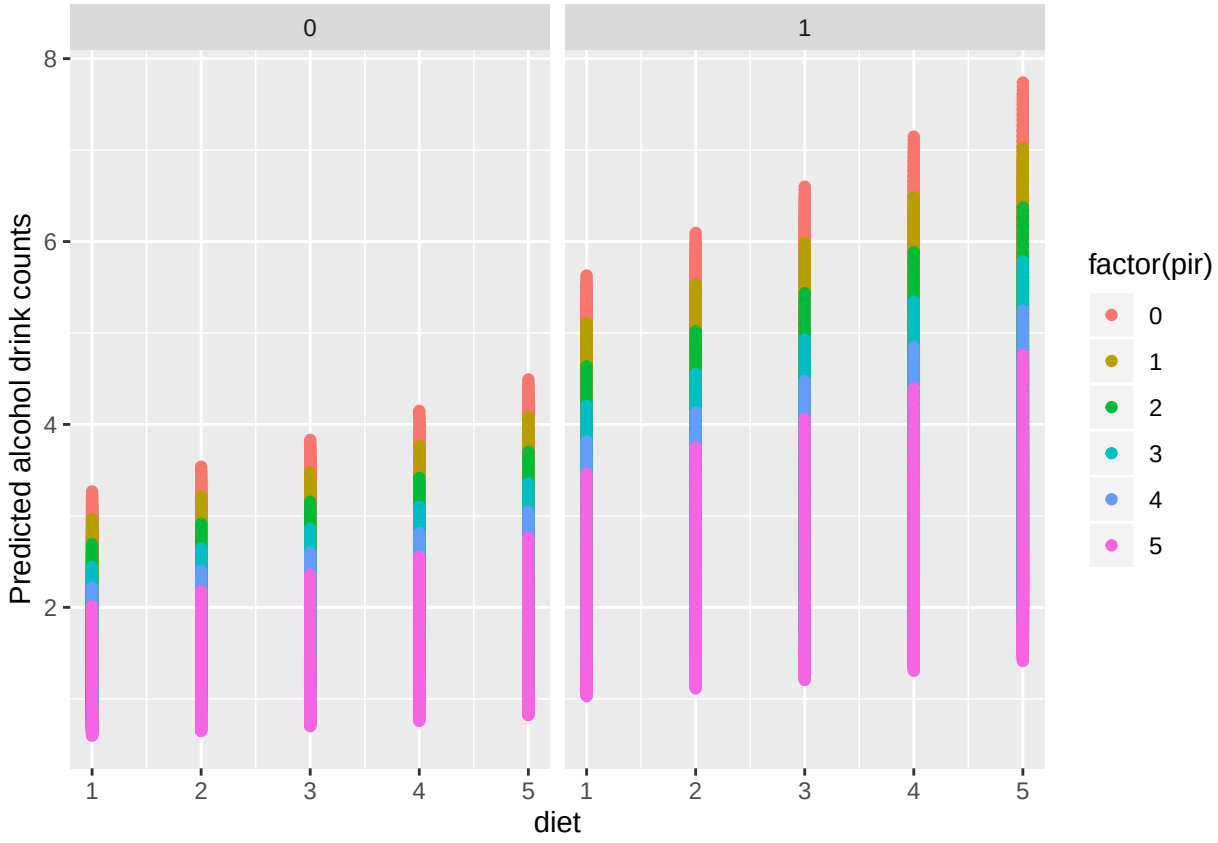
```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
```

```
## ------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                    7.589214 model1 > model2 1.6098e-14
## AIC-corrected          7.450189 model1 > model2 4.6629e-14
## BIC-corrected          7.041163 model1 > model2 9.5324e-13
```

From the result of Vuong test, we could know that zero-inflated negative binomial regression have better performance and its improvement is significant.

Finally, we can compute the predicted number of alcohol use for diffrent combinations of our predictors

```
newdata1 <- expand.grid(1:5, factor(0:1),21:85,0:5,1:21)
colnames(newdata1)<-c('diet','gender','age','pir','meal_out')
newdata1$alqpre<-predict(m1,newdata1)
ggplot(newdata1, aes(x = diet, y = alqpre, colour = factor(pir))) +
  geom_point() +
  facet_wrap(~gender) +
  labs(x = "diet", y = "Predicted alcohol drink counts")
```



From the plot, we could see directly that male use alcohol more, poor family pir use alcohol more and unheathier diet use alcohol more