Group8: Zero-inflated Negative Binomial Regression

Analysis of the relationship between diet habit and alcohol use

Amanda Lee Ketner, Yichao Chen, Xirui Wu

12/9/2019

Contents

Introduction

What is the relationship between eating habits and alcohol use? Are certain types of eating habits associated with higher alcohol consumption?

Our group mainly focuses on the relatinship between diet habit (specifically, the self-reported healthfulness of a person's overall diet as well as the average number of meals eaten outside the home weekly) and alcohol use (operationalized as the average number of alocholic drinks consumed weekly). We also use demographic variables (gender, family povery income ratio, and age) as additional covariates.

Based on our data, we chose to use zero-inflated negative binomial regression to fit the model. The choice of this regression is further illustrated our method section. Our group members used Stata,R and SAS.

Data

Data resource

We used three datasets in this analysis:

1.Alcohol Use (ALQ_D.xpt)

2.Diet Behavior & Nutrition (DBQ_D.xpt)

3.Demographic Variables & Sample Weights (DEMO_D.xpt)

They are downloaded from NHANES (National Health and Nutrition Examination Survey) from year 2005 - 2006.

As we are making analysis on the alcohol use, so we will only focus on adults (age $\geqslant$ 21).

Selected Variables

Variales	Origin_name	Description	Dataset
alq_frink	ALQ130	Avg number of alcoholic drinks/day -past 12 mos (1-32:range of values)	ALQ_D
diet	DBQ700	How healthy is the diet (1-5:Excellent-poor)	DBQ_D
meal_out	DBD091	Number of times/wk eat meals not from a home (1-21:range of values)	DBQ_D
gender	RIAGENDR	Gender (1:Male 2:Female)	DEMO_D
age	RIDAGEYR	Age at Screening Adjudicated - Recode (0-84:range of values)	DEMO_D
pir	INDFMPIR	Family Poverty Income Ratio (0-5:a ratio of family income to poverty threshold)	DEMO_D

Methods

The zero-inflated negative binomial regression is used for count data that exhibit overdispersion and excess zeros. The data distribution combines the negative binomial distribution and the logit distribution.

Suppose for each observation, there are two possible cases. If case 1 occurs, the count is zero. If case 2 occurs, counts(including zeros) are generated according to the negative binomial model. We focus on our analysis here. Case is that this person never drink because of some reasons such as alcohol allergy. Case 2 is that this person do have the habit of drink alcohol but the counts of drink may differ according to some related factors. Suppose that case 1 occurs with probalility π and case 2 occurs with probability $1-\pi$. Therefore, the probability distribution of zero-inflated negative binomial regression random variable y_i could be written:

$$\Pr\left(y_i=j\right)=\begin{cases}\pi_i+\left(1-\pi_i\right)g\left(y_i=0\right) & \text{if } j=0\\ \left(1-\pi_i\right)g\left(y_i\right) & \text{if } j>0\end{cases}$$

where π_i is the logistic link function defined below and $g\left(y_i\right)$ is the negative binomial distribution given by

$$g\left(y_i\right)=\Pr\left(Y=y_i|\mu_i,\alpha\right)=\frac{\Gamma\left(y_i+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Why we choose this zero-inflated negative binomial regression model to fit our data?

As observed, alq_drink is a count variable, this implys that the choice of possion regression or negative binomial regression may be more suitable. We then do some basic analysis of our response:alq_drink.It is overdispersed as its variance is not basicly equal to its mean. In this case, negative binomial regression might be a better choice. Also, there are excessive zeros in alq_drink. As a result, zero-inflation should be taken into consideration as well. We will also make comparisions to figure out whether the consideration of zero-inflation could contribute to a better model. Details could be found in the core analysis below.

Core Analysis

R

Data cleaning

The cleaning process is done using the package of data.table. Three diffrent datasets are merged by the variable of SEQN. Observations with missing values are all dropped.

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

As discussed above, whether negative binomial regression is more suitable than poisson regression? Should the zero-inflation be considered? To figure this out, the basic analysis should be made on the response: alq_drink.

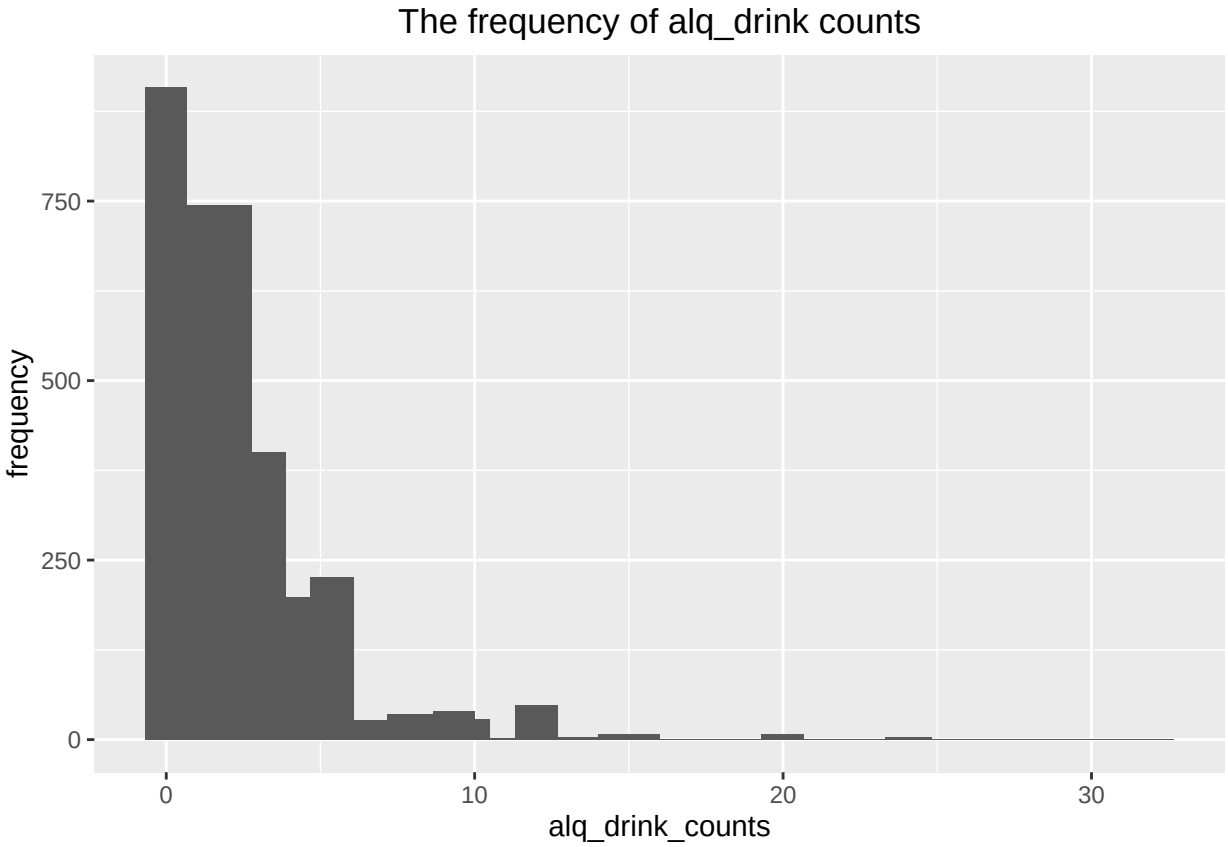
The mean and variance are calculated:

```
#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"
```

The plot of alq_drink are shown as below:

```
# histogram with alq_drink
ggplot(data,aes(alq_drink))+geom_histogram()+stat_bin(bins=25)+xlab('alq_drink_counts')+ylab('frequency
```



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 than poisson regression in our data. From the plot, we could see that a large part of observation has 0 alcohol drink. Taking excess zeros of alq_drink into consideration, zero-inflated regression is used.

Zero-Inflated Negative Binomial Regression

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.

```
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.

Negative Binomial Regression

Has the consideration of zero-inflation improved our regression model? We could fit the data with negative binomial regression and make comparisions with the former model.

The package of MASS is used for building negative bionomial regression model.

```
m2=glm.nb(alq_drink~diet+gender+age+pir,data=data)
summary(m2)
```

```
##
## Call:
## glm.nb(formula = alq_drink ~ diet + gender + age + pir, data = data,
##     init.theta = 1.567262371, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3020  -1.3625  -0.1694   0.4394   3.9882
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.474816   0.104339  14.135  < 2e-16 ***
## diet         0.084925   0.020937   4.056 4.99e-05 ***
## gender       0.652531   0.041715  15.643  < 2e-16 ***
## age         -0.021419   0.001284 -16.686  < 2e-16 ***
## pir         -0.104340   0.012822  -8.137 4.04e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.5673) family taken to be 1)
##
##     Null deviance: 3771.7  on 2654  degrees of freedom
## Residual deviance: 3103.4  on 2650  degrees of freedom
## AIC: 10574
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.5673
##           Std. Err.:  0.0862
##
##  2 x log-likelihood:  -10562.3840
```

```
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
```

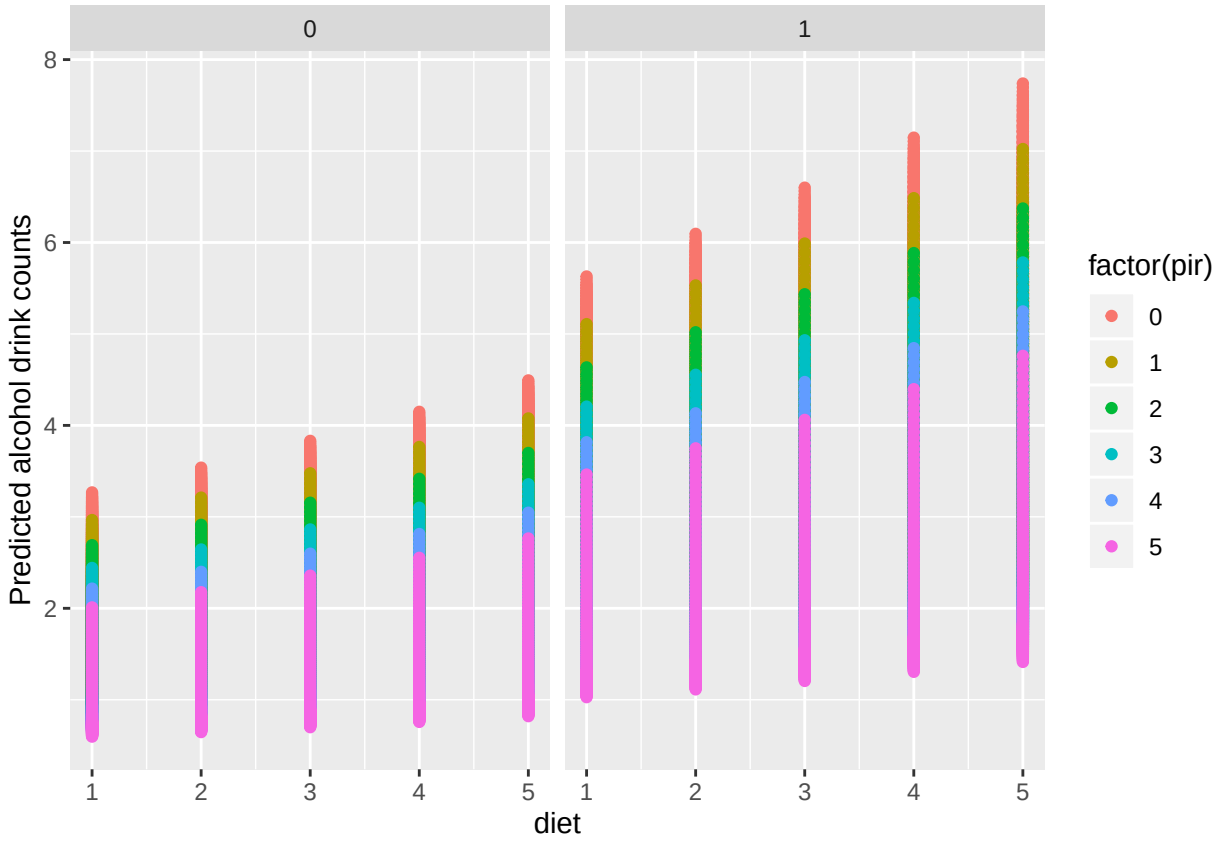
The summary of m2 shows that the variables of diet, gender, age and pir are all significant. The Vuong test suggests that zero-inflated negative binomial regression have better performance and its improvement is significant.

Predict alq_drink and Show results

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

The plot of predicted alq_drink counts and variables may help us see the potential relationship more directly.

```
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. This implies that the use of alchhol may have some correlation with gender, family PIR and diet habit.

Stata

Cleaning the data

```
cd "/Data/"

*Import data
import sasxport5 "ALQ_D.xpt", clear
save "Alcohol.dta", replace
import sasxport5 "DBQ_D.xpt", clear
save "Diet.dta", replace
import sasxport5 "Demo_D.xpt", clear
save "Demographic.dta", replace
```

Table 1: Non-Zero-Inflated Negative Binomial

X	Point_estimate	Standard_error	Z.value	P.value
Diet	0.085	0.021	4.119	0
Gender	0.652	0.042	15.630	0
Age	-0.021	0.001	-16.600	0
PIR	-0.104	0.013	-8.206	0
Constant	1.474	0.103	14.310	0

```
*Clean data and drop unwanted variables
use "Alcohol.dta", clear
keep seqn alq130
replace alq130=. if alq130==999
replace alq130=0 if alq130==1
rename alq130 alq_drink
save "Alcohol.dta", replace

use "Diet.dta", clear
keep seqn dbd091 dbq700
replace dbd091=0 if dbd091==6666
replace dbd091=21 if dbd091==5555
replace dbd091=. if dbd091==7777 | dbd091==9999
replace dbq700=. if dbq700==7 | dbq700==9
rename dbd091 meal_out
rename dbq700 diet
save "Diet.dta", replace

use "Demographic.dta", replace
keep seqn ridageyr riagendr indfmpir
replace riagendr=0 if riagendr==2
rename ridageyr age
rename riagendr gender
rename indfmpir pir
save "Demographic.dta", replace

*Merge data together and drop minors
use "Alcohol.dta", clear
merge 1:1 seqn using "Diet.dta", nogen
merge 1:1 seqn using "Demographic.dta", nogen
drop if Age<21
save "Final.dta", replace
```

Negative binomial model (without zero-inflation)

```
nbreg alq_drink diet i.gender age pir
outreg2 using NegBin.xls, stats(coef se tstat pval) noaster replace
```

Zero-Inflated Negative binomial model

```
zinb alq_drink diet i.gender age pir, inflate(meal_out) forcevuong
outreg2 using ZeroNegBin.xls, stats(coef se tstat pval) noaster replace
```

Table 2: Zero-Inflated Negative Binomial

X	Point_estimate	Standard_error	Z.value	P.value
Diet	0.080	0.018	4.447	0
Gender	0.544	0.038	14.440	0
Age	-0.016	0.001	-13.140	0
PIR	-0.097	0.011	-8.839	0
Meal_out	-0.055	0.019	-2.924	0
Constant	1.547	0.089	17.380	0

Consistent with the results using the other methods, the Vuong test shows that the zero-inflated negative binomial model is preferred ($p<0.000$).

SAS

Data Cleaning

Read the data the join variables to use in a new dataset.

```
# read data :
alcohol = sasxport.get("ALQ_D.XPT")
diet = sasxport.get("DBQ_D.XPT")
demographic = sasxport.get("DEMO_D.XPT")

# select variables and merge :
alcohol_need = select(alcohol, id = seqn, alq = alq130)
diet_need = select(diet, id = seqn, meal_out = dbd091, diet = dbq700)
demo_need = select(demographic, id = seqn, age = ridageyr, gender = riagendr, pir = indfmpir)
alq_diet = merge(alcohol_need, diet_need, by = "id")
alq_diet = merge(alq_diet, demo_need, by = "id")
```

Since in the raw dataset, people who don't drink report "1" in the variable "ALQ130", transfer 1's in the related variable to 0's. Delete all missing values.

```
alq_diet[which(alq_diet$alq %in% c(777,999)),"alq"] = NA
alq_diet[which(alq_diet$alq == 1),"alq"] = 0
alq_diet[which(alq_diet$meal_out == 5555), "meal_out"] = 21
alq_diet[which(alq_diet$meal_out == 6666), "meal_out"] = 0
alq_diet[which(alq_diet$meal_out %in% c(7777,9999)),"meal_out"] = NA
alq_diet[which(alq_diet$gender == 2), "gender"] = 0
alq_diet = alq_diet%>%
  filter(age>=21) %>%
  filter(!is.na(alq) & !is.na(meal_out) & !is.na(pir) & !is.na(diet))
```

Write the new dataset to a csv file for later using in SAS.

```
write.csv(alq_diet, file = "alq_diet.csv")
```

Check the data

Read the new dataset into SAS.

Check the distribution (mean and variance) of the variables.

```
knitr::include_graphics("dist_check.pdf")
```

Draw a histgram to see the distribution of the response variable ("alcohol use").

```
knitr::include_graphics("hist_alq.pdf")
```

The response variable (alcohol use) is a count variable, so generally we can use Poisson regression. However, the mean and variance of the response shown above as "alq" are not the same, which does not fit the assumption of Poisson regression. So we choose to use negative binomial regression instead.

From the histgram we see that many people do not drink thus there are many 0's in the response. We can imagine that people's not drinking is influnced by another process compared to how much people drink. Thus we use zero-inflated negative binomial regression so that the 0's can be independently modeled.

3.Build Model

negative binomial regression

Firstly, use negative binomial regression to fit model alcohol ~ diet + gender + age + pir . The result is as below.

zero-inflated negative binomial regression

Then, use zero-inflated negative binomial regression to fit model alcohol ~ diet + gender + age + pir where alcohol is not 0 and fit model alcohol ~ meal_out where alcohol is 0. The result is as below.

We can see from the results that in both models, all the variables are significant. By the AIC value, we can see that the zero-inflated negative binomial model is better.

Comparing and Analysis

In this part, I'll use Vuong test to compare the performance of the two models. And then do some further analysis.
This part is to be done.

Results

Overall, we found a statistically significant and positive relationship between the self-reported healthfulness of a person's diet and their alcoholic consumption, indicating that people with less healthy diets tend to drink more alcohol on average. Specifically, 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.

When gender change from 0 to 1, the change in log(alq_drink) is 0.544, indicating that men tend to drink more than women do.

The expected change in log(alq_drink) for one-unit increase in age is -0.016 holding other variable constant. This means older people tend to drink less alcohol.

The expected change in log(alq_drink) for one-unit increase in age is -0.097 holding other variable constant, which means that wealtheir families consume use less alcohol on average.

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.

Discussion

References

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Negative_Binomial_Regression.pdf

https://www.stata.com/manuals13/rzinb.pdf