Group8_Alcohol_use

Yichao Chen

12/3/2019

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.

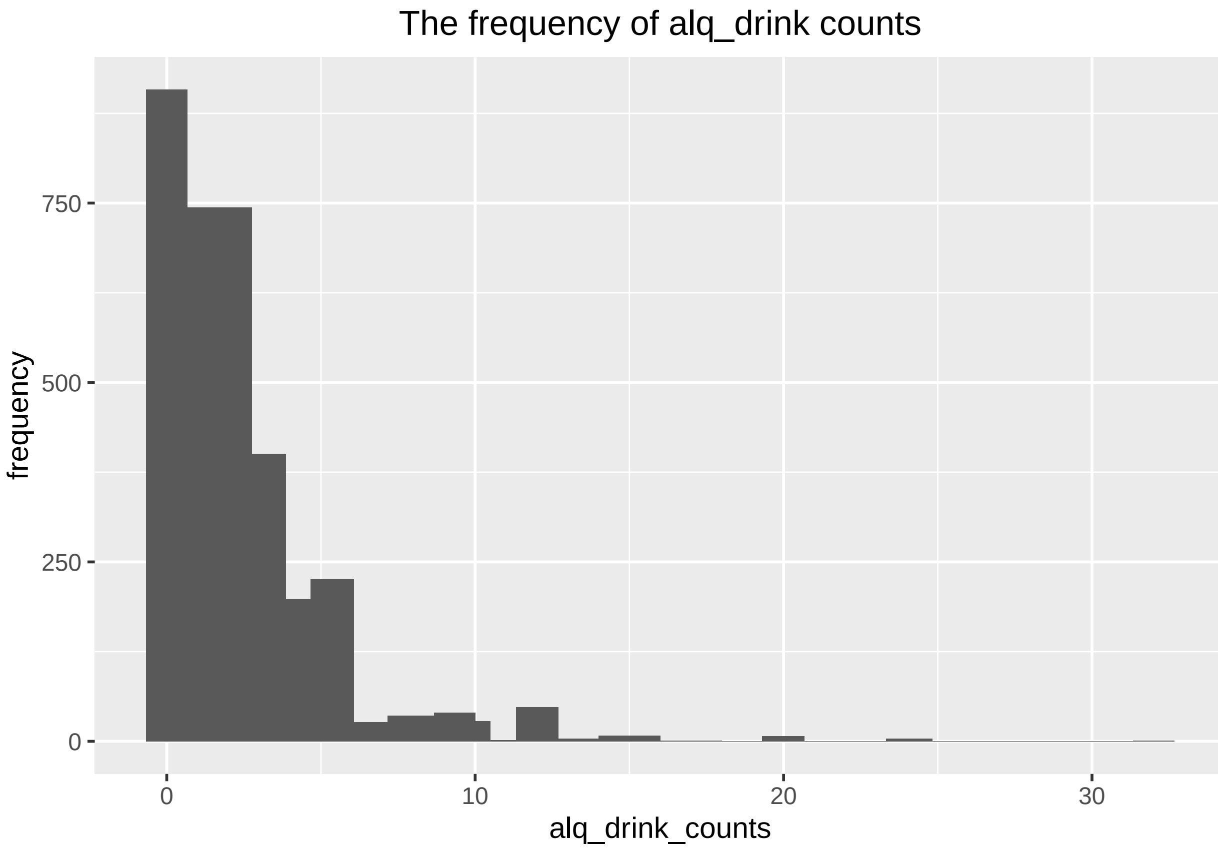
The mean and variance are calculated:

```
#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:

```r
# 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.

```r
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.

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
```

```r
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
```

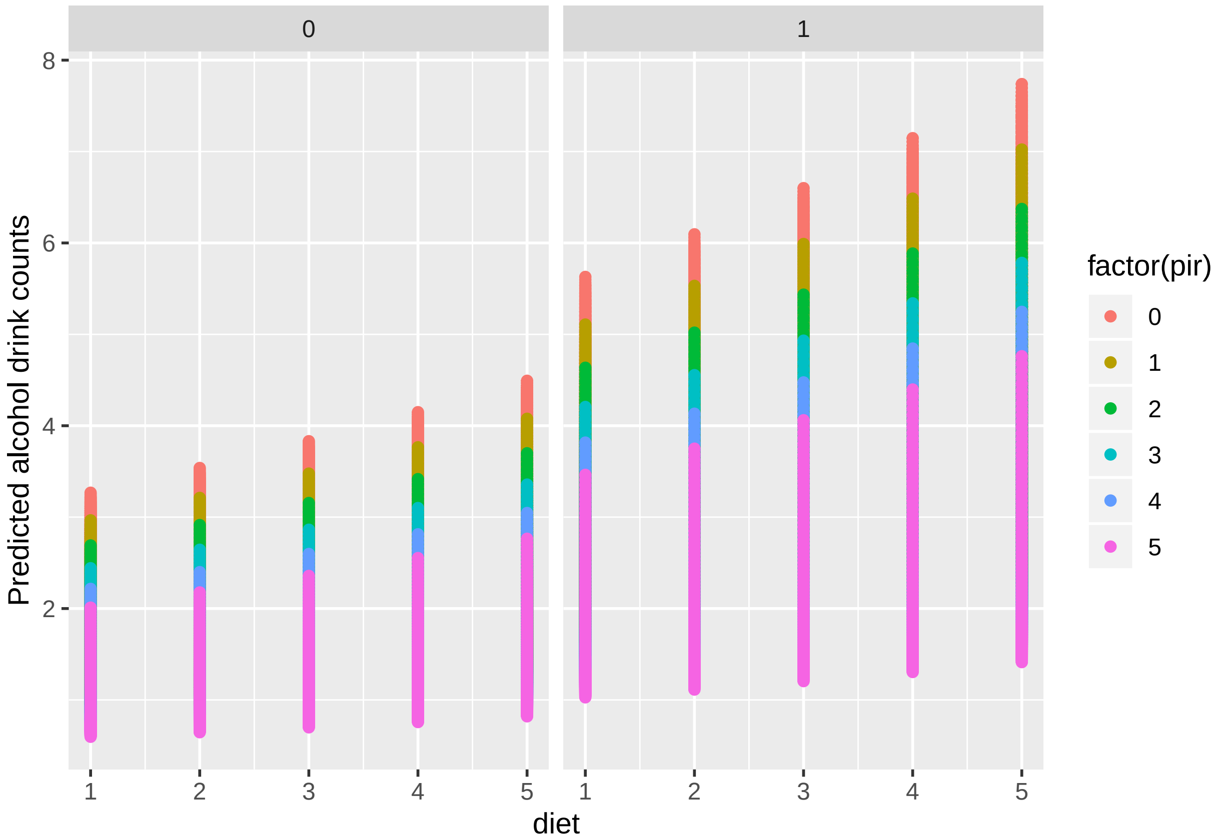
The summary of m2 shows that the variables of diet, gender, age and pir are all significant. The Vuong test suggests that zero-inflated negative binomial regression have better performance and its improvement is significant.

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.

```r
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.