Group8 Alcohol Use and Eating Habit

Xinrui Wu

12/5/2019

Introduction

Our group mainly focus on the relationship between eating habit and alcohol use of people while some demographic variables are also included. Zero inflation negative binomial regression model is used in our analysis process due to the structure of the response variable.

Data to Use

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) - "alcohol"
DBQ700 - How healthy is the diet (1-5:Excellent-poor) - "diet"
DBD091 - # of times/wk eat meals not from a home (1-21:range of values) - "meal_out"
RIAGENDR - Gender (1:Male 2:Female) - "gender"
RIDAGEYR - Age at Screening Adjudicated - Recode (0-84:range of values) - "age"
INDFMPIR - Family PIR (0-5:a ratio of family income to poverty threshold) - "pir"

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

Programming Method

R (using dplyr for basic data cleaning) and SAS

1. Data Cleaning

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

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.

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

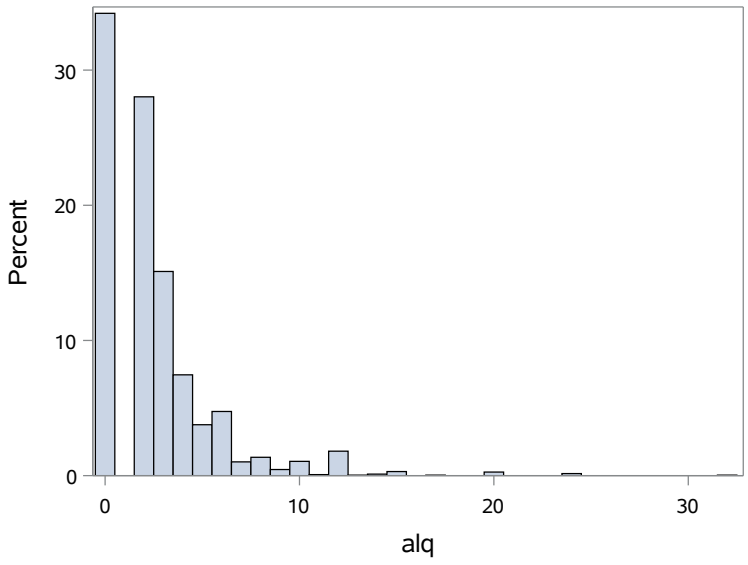
2.Check the data

Read the new dataset into SAS.
Check the distribution (mean and variance) of the variables.

The MEANS Procedure

Variable	Mean	Variance
alq	2.5092279	8.8920587
meal_out	3.5107345	13.9432721
diet	2.9457627	1.0196646
age	45.7781544	295.5977140
gender	0.5322034	0.2490567
pir	2.9519171	2.5836276

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



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

(1) negative binomial regression

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

The GENMOD Procedure

Model Information	
Data Set	WORK.ALCOHOL_DIET
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	alq

Number of Observations Read	2655
Number of Observations Used	2655

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	2650	3103.4288	1.1711
Scaled Deviance	2650	3103.4288	1.1711
Pearson Chi-Square	2650	2612.0316	0.9857
Scaled Pearson X2	2650	2612.0316	0.9857
Log Likelihood		1024.7471	
Full Log Likelihood		-5281.1922	
AIC (smaller is better)		10574.3843	
AICC (smaller is better)		10574.4161	
BIC (smaller is better)		10609.6895	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.4748	0.1031	1.2728	1.6768	204.78	<.0001
diet	1	0.0849	0.0207	0.0444	0.1255	16.87	<.0001
gender	1	0.6525	0.0418	0.5707	0.7344	244.18	<.0001
age	1	-0.0214	0.0013	-0.0239	-0.0189	275.35	<.0001
pir	1	-0.1043	0.0127	-0.1293	-0.0794	67.33	<.0001
Dispersion	1	0.6381	0.0351	0.5728	0.7107		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

(2) 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.

The GENMOD Procedure

Model Information	
Data Set	WORK.ALCOHOL_DIET
Distribution	Zero Inflated Negative Binomial
Link Function	Log
Dependent Variable	alq
Zero Model Link Function	Logit

Number of Observations Read	2655
Number of Observations Used	2655

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance		10344.0291	
Scaled Deviance		10344.0291	
Pearson Chi-Square	2648	2789.6923	1.0535
Scaled Pearson X2	2648	2789.6923	1.0535
Log Likelihood		-5172.0145	
Full Log Likelihood		-5172.0145	
AIC (smaller is better)		10360.0291	
AICC (smaller is better)		10360.0835	
BIC (smaller is better)		10407.1027	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.5471	0.0890	1.3726	1.7216	301.91	<.0001
diet	1	0.0796	0.0179	0.0445	0.1147	19.78	<.0001
gender	1	0.5439	0.0377	0.4700	0.6177	208.45	<.0001
age	1	-0.0161	0.0012	-0.0186	-0.0137	172.55	<.0001
pir	1	-0.0972	0.0110	-0.1187	-0.0756	78.12	<.0001
Dispersion	1	0.2236	0.0201	0.1874	0.2667		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

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.

4.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.