MA684 Midterm Project: Behavioral Risk Factor Analysis for Mental Health

Yoojin Jung
December 11, 2017

1. Research Background

Behavioral health refers to a person's overall wellbeing, which is affected by the person's behaviors. Behavioral health issues can be prevented, cured, or alleviated by behavioral choices that the individual person makes.

The Behavioral Risk Factor Surveillance System (BRFSS) is the system that conducts health-related telephone surveys and collects state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The data collected are used to target and build health promotion activities.

The goal of this study is to assess the relationship between behaviors and mental health, which includes stress, depression, and problems with emotions, using a dataset from BRFSS. Mental health in this survey is measured by the number of days during the past 30 days the mental health was not good. As the behavioral factors, I considered exercise and drinking alcohol occasionally and assessed whether these behavioral factors would make the mental health improve or worsen.

2. Data Description

The dataset is the survey data collected in 2010 and includes 451075 rows (observations). Mental health in this survey is measured by the number of days during the past 30 days the mental health was not good, and therefore the response ranges from 0 to 30. Among numerous question entries, exercise and drinking alcohol were considered as the behavioral predictors affecting the mental health. In addition to the main predictors and response, several demographic characteristics were included for the analysis.

The dataset includes an excess of zero counts (see Figure 1), indicating most of the survey participants felt mentally healthy. The general health for the zero counts varies from excellent to poor, implying a person with poor general health conditions could have been okay during the past 30 days. Another noticeable characteristic of the dataset is the large counts for the survey participants who felt mentally sick for the whole 30 days. These counts include even people who said their health in general is excellent. Such count peaks on both sides could be problematic when matching the data with statistical models. This issue will be examined in more detail below.

The dataset was cleaned up before any analysis. The responses of “don’t know/not sure” or “refused” were removed from the dataset. After the clean-up, the dataset includes 415420 rows.

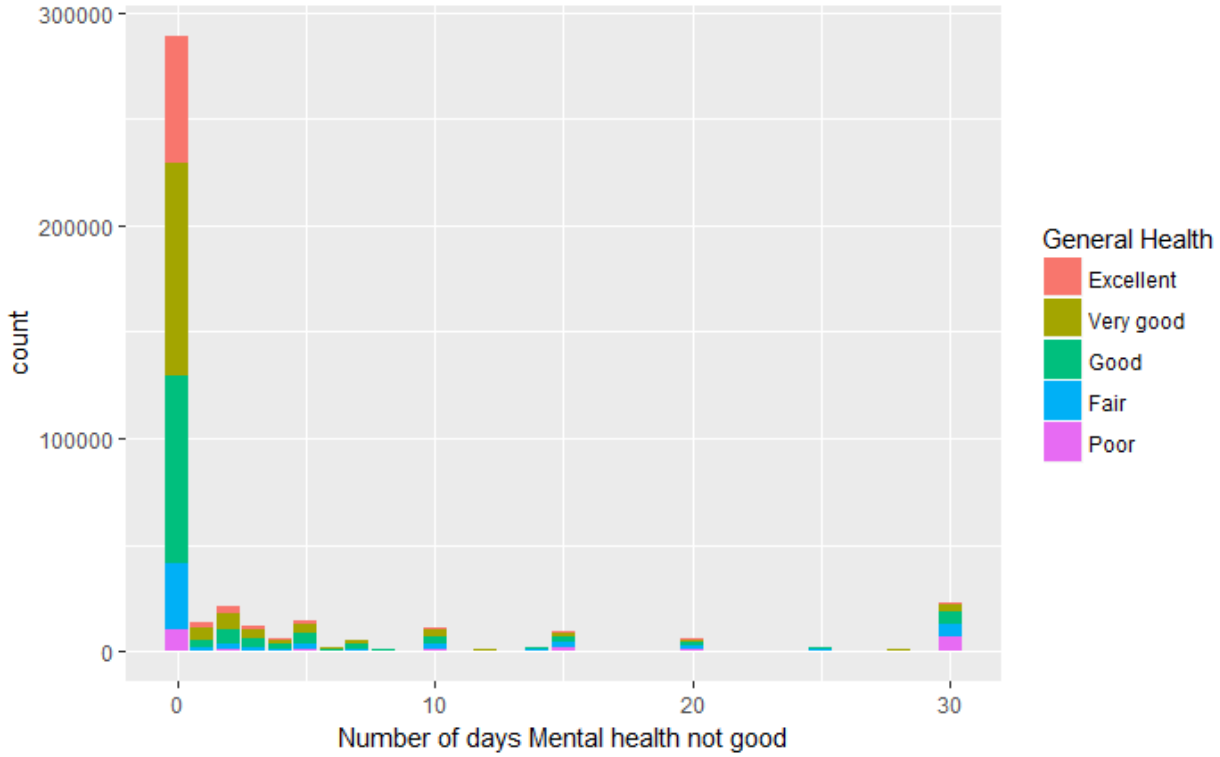

Figure 1. The distribution of number of days mental health not good.

3. Exploratory Data Analysis

Before conducting statistical analysis, I first performed the exploratory data analysis to visually understand the data and the potential relationship between the predictors and the response. Figure 2 shows the mental health level depending on whether they exercise, drink and their education level. The numbers on the right side represent the education level: 1) Never attended school or only Kindergarten, 2) Elementary, 3) Some high school, 4) High school graduate, 5) Some college, and 6) College graduate. According to Figure 2, people who exercise and drink occasionally tend to be mentally healthy. In addition, when people have higher education level, such as some college and college graduate, a larger portion of them seem to exercise and drink occasionally. The number of people who never attended school or only kindergarten was only 403, and therefore only some of days out of 30 days were only reported. Figure 3 shows the average mental health level summarized by the state. In general, the mentally sick days are fewer in the West North Central Division and greater in the East South Central Division.

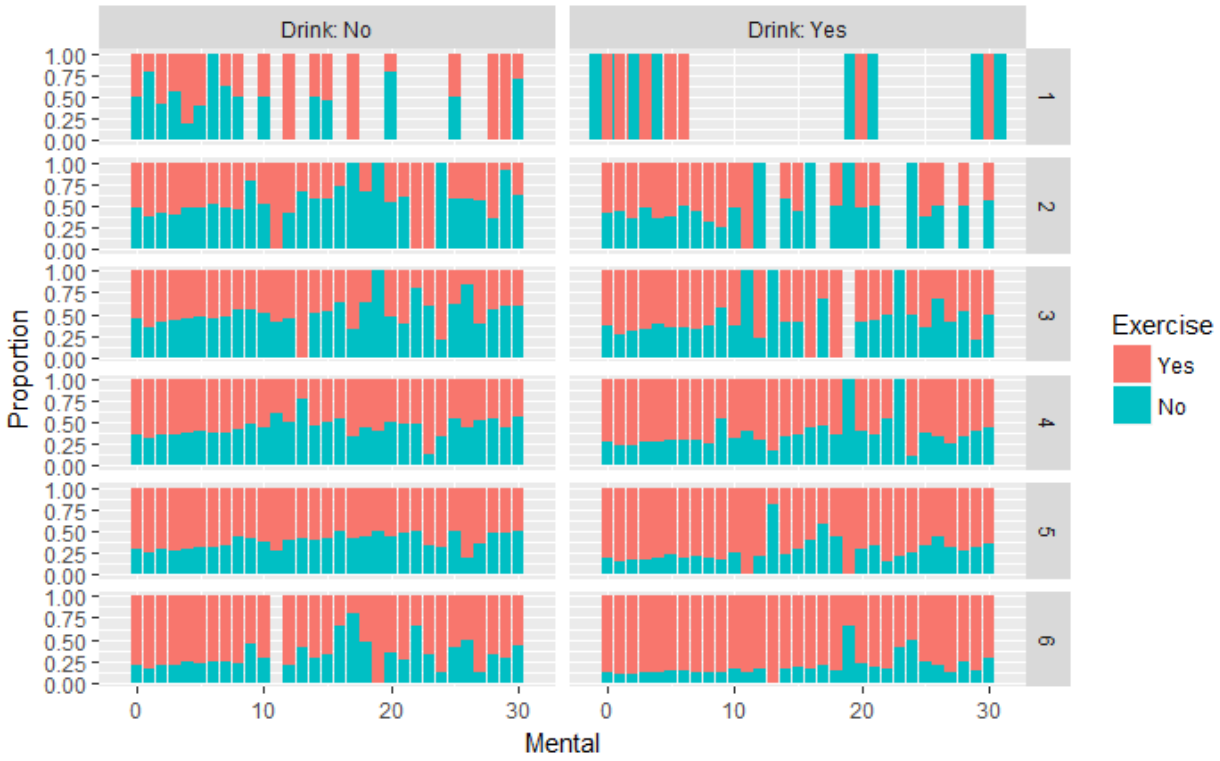

Figure 2. The distribution of number of days mental health not good by exercise, drink, and education.

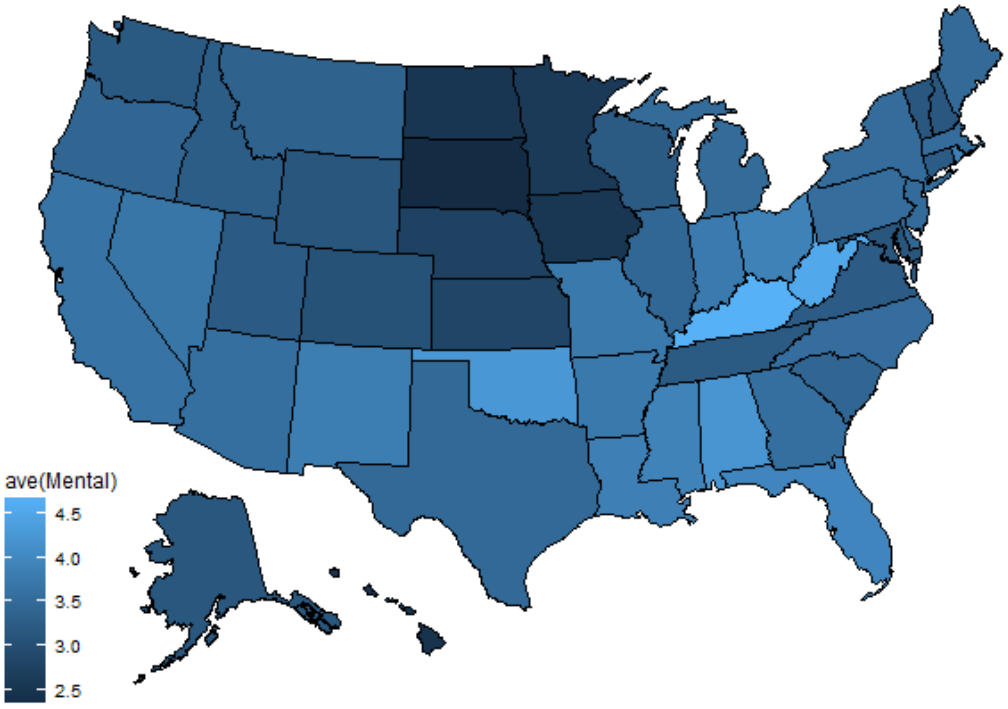

Figure 3. The average mental health level for each State.

4. Statistical Analysis

4.1. Logistic-binomial model

First, I started with a simple model, including only behavioral related covariates. Since each data point of the response (mentally sick days) can be interpreted as the number of "successes" out of the total trials (total 30 days), the logistic-binomial model was applied.

Table 1. Summary of the logistic-binomial model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-2.37	0.001	-1609.6	<2e-16
Exercise1	0.67	0.002	357.2	<2e-16
Drink1	0.20	0.002	108.0	<2e-16

The estimated coefficients are all significant based on the standard error and z value. The intercept term is the prediction of mentally sick days if the participant engaged in any physical activities or exercise during the past 30 days and occasionally consumed alcohol, which is 2.6 days. The coefficient of Exercise1 is the expected difference in the mentally sick days if the respondent did not participated in any physical activities or exercise during the past 30 days, which corresponds to 9% increase in the probability. The coefficient of Drink1 is the expected difference in the mentally sick days if the respondent did not drink alcohol occasionally, which corresponds to 2% increase in the probability. However, the overdispersion rate was 14.5, indicating the data have more variation that is explained by the model. To account for the overdispersion, the quasibinomial family was used for the bionomial regression. Then, the standard errors were increased by a factor of 4. The interaction term was considered as well, but the estimated coefficients were almost identical to those from the regression model without the interaction term. And as shown in Figure 4, the confidence interval for the interaction term crosses zero, indicating that this term is not statistically significant.

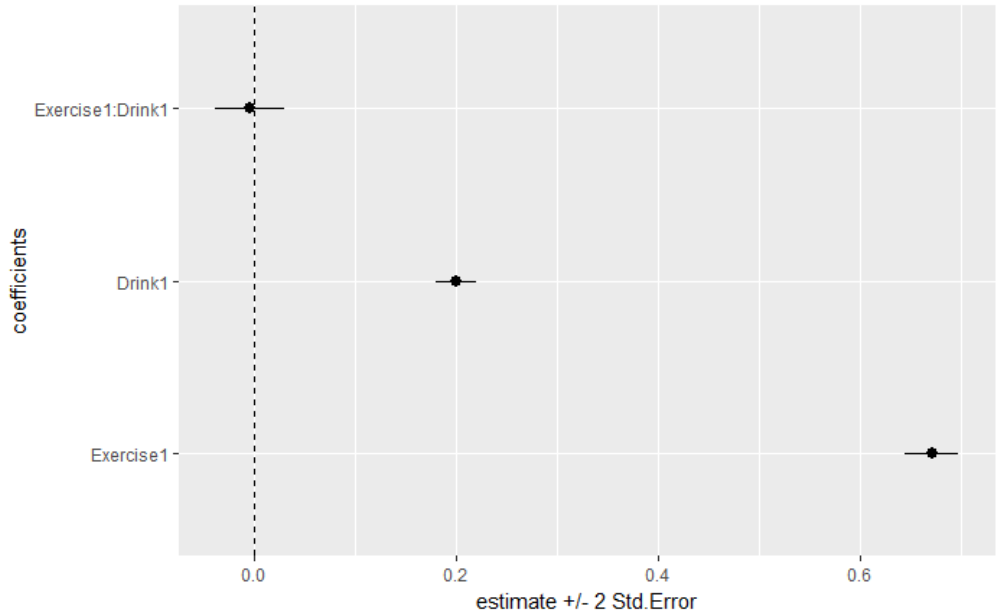

Figure 4. The coefficient plot with the confidence interval for the quasibinomial-logistic model with the interaction term.

While the model estimates seem good, the probability plot in Figure 5 reveals that the fitted model only captures part of the observed data. Figure 6 clearly shows the difference in the distribution of number of days mental health not good between the observed data (red) and the simulated data (blue), indicating the logistic-binomial model might not be appropriate to explain the data.

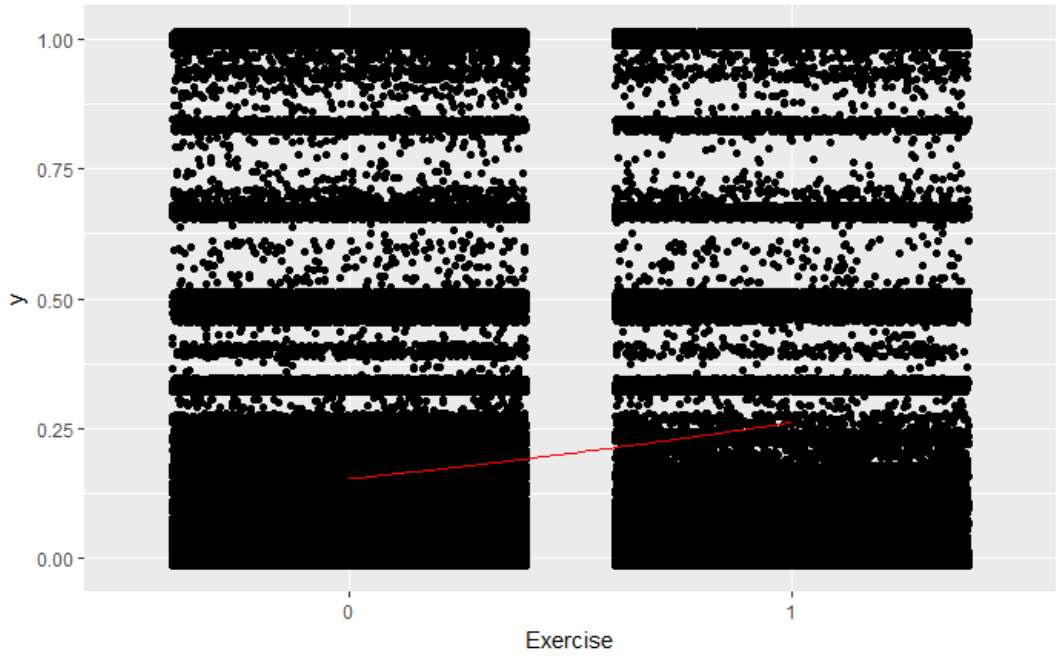

Figure 5. The probability of being mentally sick as a function of exercise with the fitted model.

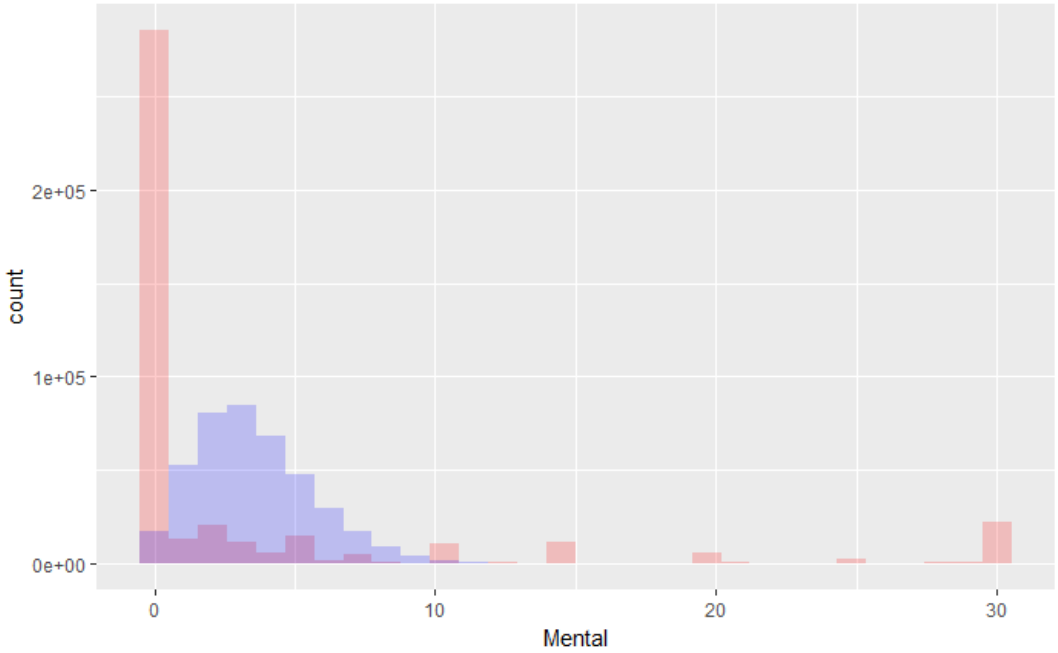

Figure 6. The distribution of number of days mental health not good. The red color represents the observed data, and the blue color the simulated data by the logistic-binomial model.

4.2. Negative binomial model

Knowing the dataset is overdispersed and the response is the count data, I considered the negative binomial model. The issue with using this model is that the dataset has a limit, ranging only up to 30, whereas the negative binomial model assumes that the response does not have a natural limit. This limitation will be discussed further in the discussion section.

Table 2. Summary of the negative binomial model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.94	0.007	134.0	<2e-16
Exercise1	0.58	0.010	54.8	<2e-16
Drink1	0.18	0.009	18.7	<2e-16
theta	0.12	0.0004		

The estimated coefficients are again all significant based on the standard error and z value. The intercept term is the log of the expected counts when people exercise and drink occasionally (that is, the expected counts in this case is 2.6 days, which is actually identical to that in the logistic-binomial model). The coefficient of Exercise1 is the difference in the logs of expected mentally sick days if the respondent did not participate in any physical activities or exercise during the past 30 days. The coefficient of Drink1 can be similarly interpreted. As before, the effect on the response was bigger when behavior changes regarding exercise. Figure 7 shows the rootogram for number of days mental health not good. The rootogram compares the observed and expected values graphically. The hanging plot shows the observed counts (bars) hanging from the curve representing the expected counts. The suspended plot mainly emphasizes the deviations between the expected and observed counts. This negative binomial model was able to account for the excess zeros; the zero counts were 69% of the total survey dataset, and the zero counts by the negative binomial model were 68%. However, the model was not able to capture the large count at 30 days and included some counts that were more than 30 days, which does not make sense. This was an expected issue of applying the negative binomial model for the dataset with the limit. There are other regression models that might match the observed dataset better, which could handle censored data on both sides, and will be considered later after having better understanding on those models. For now, I will use the negative binomial model for multilevel models since the negative binomial model in general seems to capture the dataset better as shown in Figure 8. The Q-Q plot for the logistic-binomial model indicates this model barely captures the data. On the other hand, the Q-Q plot for the negative binomial model indicates the positively skewed distribution with a fat positive tail (and a slightly thin negative tail), but mostly captures the data except the counts that represent the cases where mentally sick days are more than 30 days.

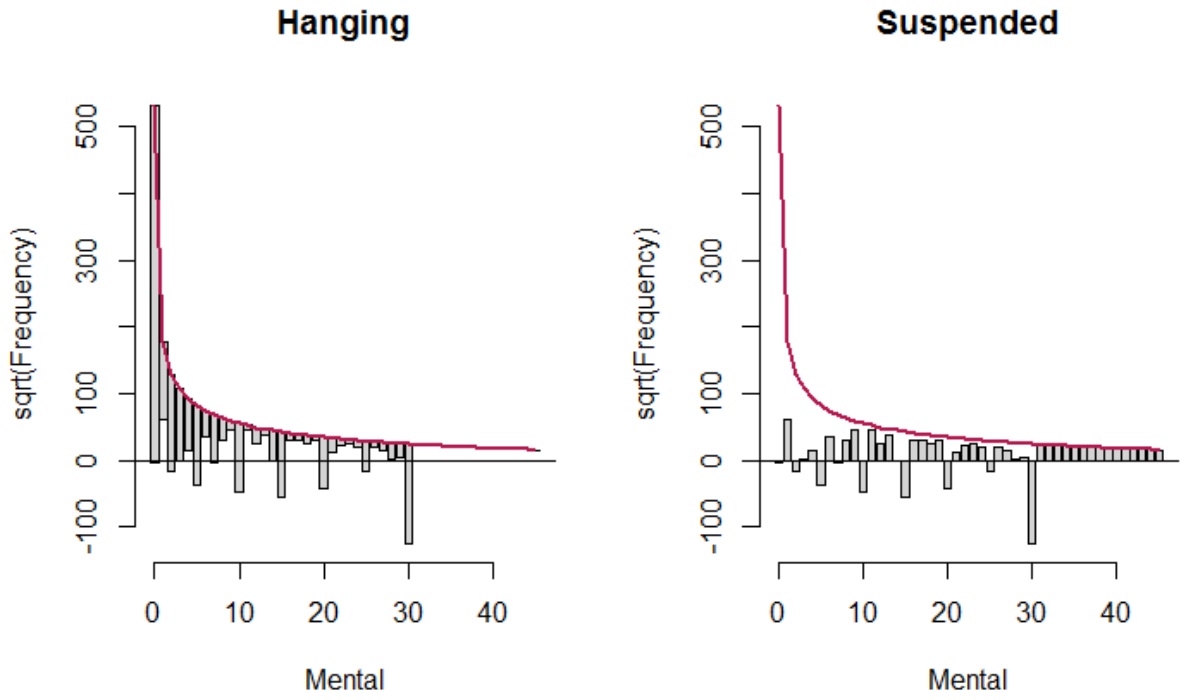

Figure 7. The hanging and suspended rootograms for number of days mental health not good.

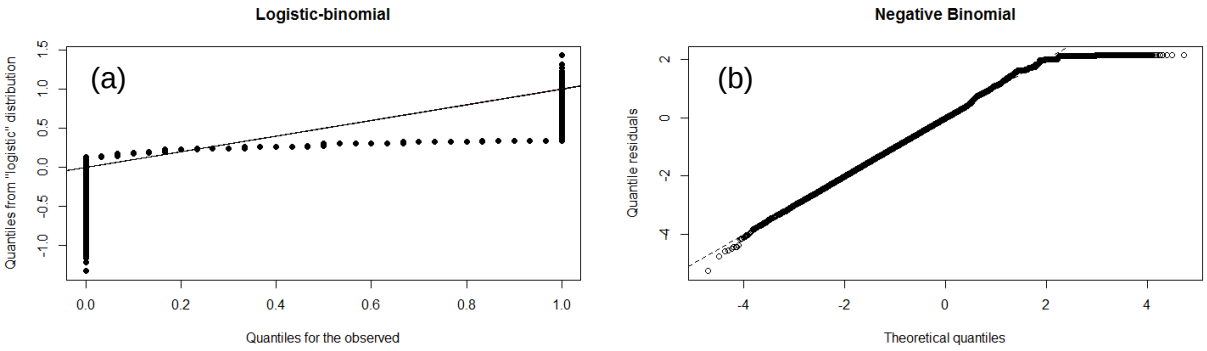

Figure 8. The Q-Q plots for (a) the logistic-binomial model and (b) the negative binomial model.

4.3. Multilevel models

In this study, I considered two factors as random effect: the states and the education levels. The dataset includes the survey data from all the states in the U.S., and here the differences between the states were considered as random effect. While education is often considered as the attribute for a regression model, the education level was considered as random effect as well. For multilevel models, I used a subset of the dataset to facilitate fitting the multilevel generalized linear models. Among the education levels, the first level of Never attended school or only Kindergarten is rather incomplete in terms of the distribution of number of days mental health not good, which resulted in quite different estimates for that level depending on the sampling size. Therefore, after excluding this education level, 10,000 sample data were randomly selected. The estimated coefficients for the sampled dataset were similar to those for the original dataset (see Tables 2 and 3), indicating the sample dataset well represents the original dataset. Due to the decrease in the sample size, the standard errors were increased by a factor of 7, but the coefficients were still significant based on the standard error and z value.

Table 3. Summary of the negative binomial model for the sampled dataset

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.93	0.05	20.5	<2e-16
Exercise1	0.57	0.07	8.4	<2e-16
Drink1	0.16	0.06	2.7	0.007
Theta	0.11	0.003		

The estimates of mental health from a multilevel model with the states as random effect revealed that the difference between the states is negligible (the variance for the random effect was zero). The coefficients for the fixed effects were identical to those from the single-level model (complete-pooling estimates). For comparison, the states were included as the attribute for the single-level model. Figure 9 shows the incident rate ratios for this case. While some variance between the states were observed, the coefficients in terms of incident rate ratios were not significant (the confidence intervals cross the vertical line of one), except District of Columbia, supporting the result of the multilevel estimates.

The estimates of mental health from a multilevel model with the education levels as random effect shows a minor variance of 0.02, but the standard error of the random effect was 0.15, indicating the variance is not significant. Figure 10 shows the best linear unbiased prediction (BLUP) for the estimates of random effects, which is basically the estimated intercept for each education level group. The only education level group that has a significantly different intercept is college graduate (group 6).

Table 4. Summary of the multilevel negative binomial model with the education level as the attribute

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.02	0.09	11.3	<2e-16
Exercise1	0.53	0.07	7.7	1.74e-14
Drink1	0.13	0.06	2.0	0.04
Random Effects	Variance	Std. Error		
Education	0.02	0.15		

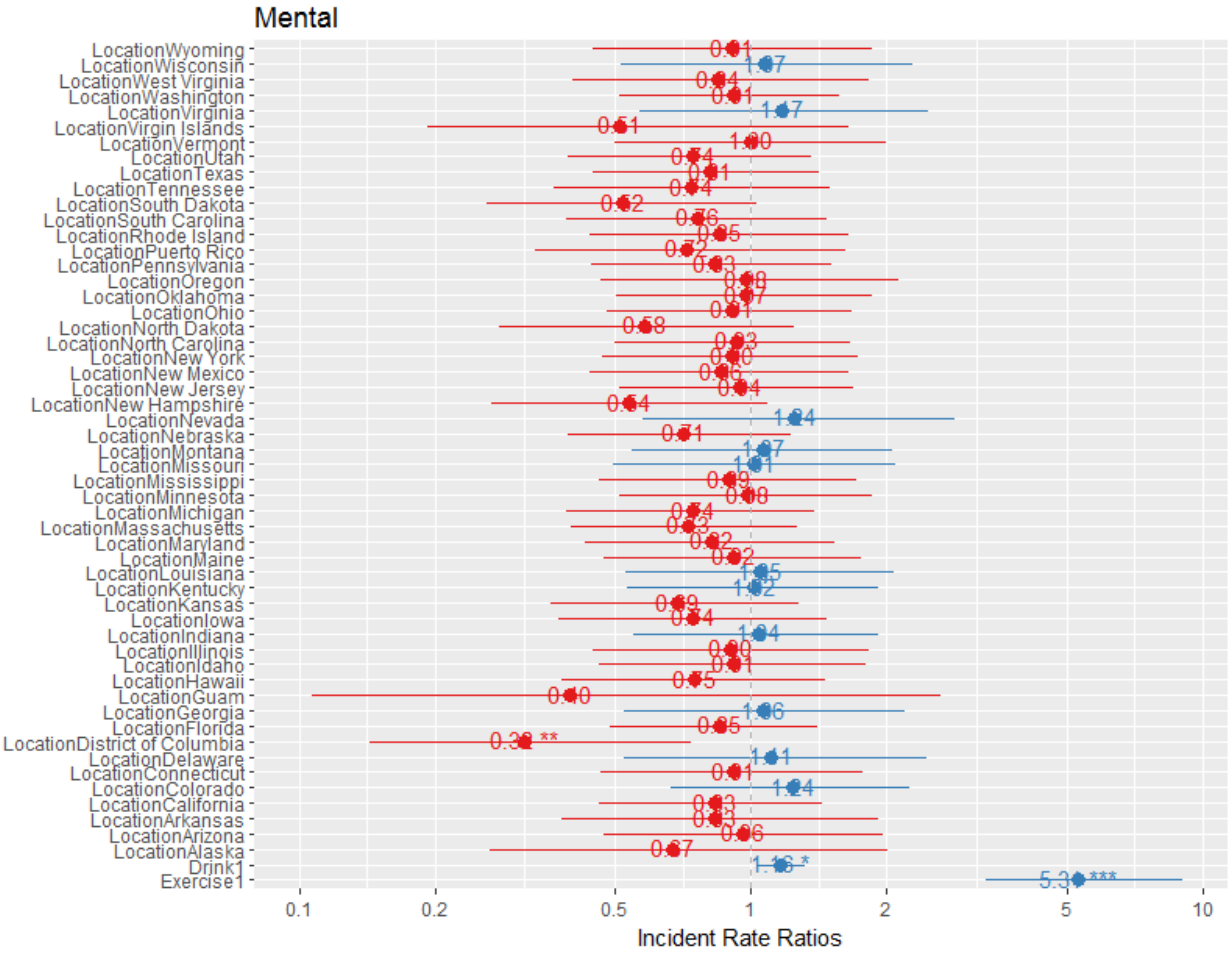

Figure 9. The incident rate ratios for the negative binomial model with the states as the attribute.

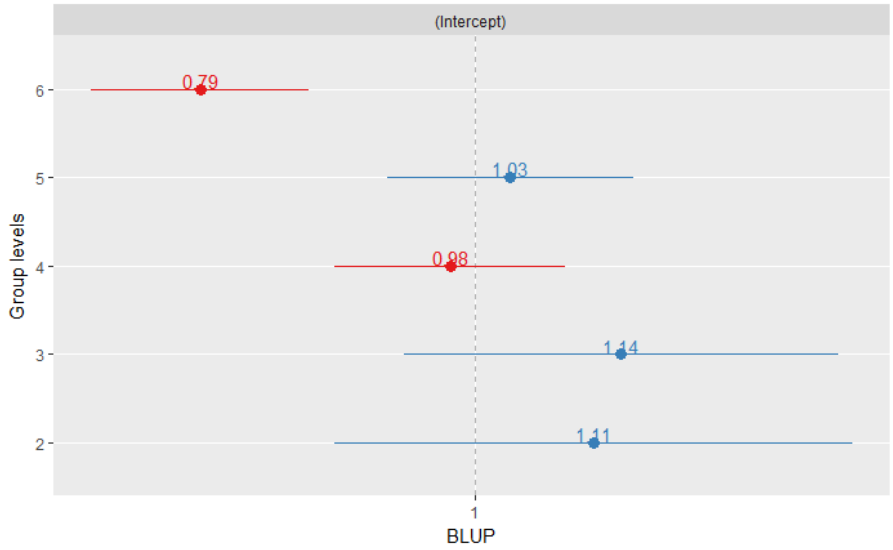

Figure 10. The best linear unbiased prediction (BLUP) for the estimates of random effects (vary intercepts by education levels).

The results from the multilevel model were also compared with those from no-pooling. Figure 11 shows the estimated coefficients from no-pooling analysis, but it should be noted that the education level 2 is not included because the negative binomial model failed to converge due to the relatively smaller sample size for the education level 2 (289 observations). The most variation by education is observed with the intercept term. The variation among the estimated intercepts from no-pooling analysis was slightly bigger than that in the multilevel model (see Figure 10).

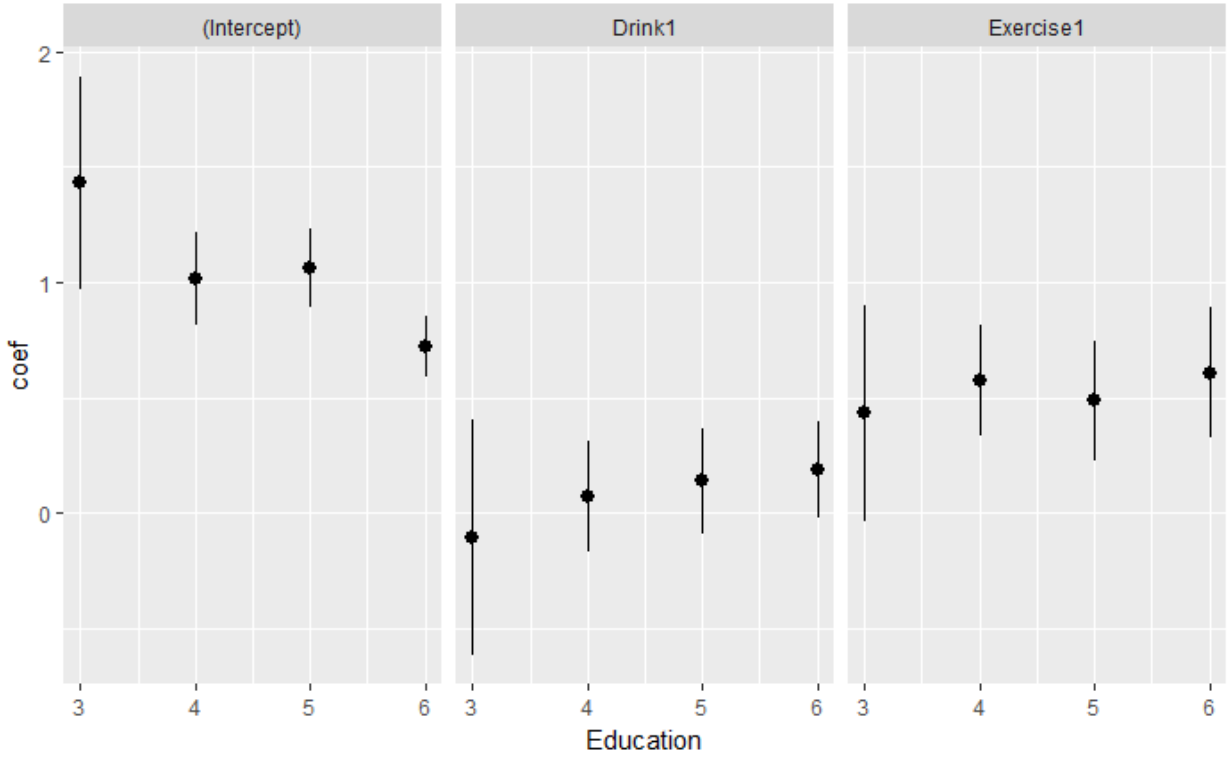

Figure 11. The estimated coefficients from no-pooling analysis.

Another comparison is to consider the education as an attribute. The summary of the estimates are shown in Table 5. The exponent of the intercept term is the expected counts when people exercise and drink occasionally with the education level of element, which is 3.5 days. The difference between the education levels is only significant for the education level 6, which corresponds with the result of the multilevel model.

In general, the difference between the models tested is not significant. However, the AIC value was the smallest for the single-level model with the education level as an attribute (33193.0), and the biggest for the single-level without the education level as an attribute (33210.8). The AIC of the multilevel model was 33200.0. Figure 12 shows the rootograms for number of days mental health not good when the education level is included as an attribute. They look very similar to the plots in Figure 7, supporting that the difference between the models is minor. The Q-Q plot shown in Figure 13 shows the distribution obtained by this updated single-

level negative binomial model is improved slightly (still positively skewed distribution with a fat positive tail but no thin negative tail).

Table 5. Summary of the negative binomial model including the education as an attribute

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	1.26	0.18	6.8	7.6e-12
Education3	-0.02	0.22	-0.1	0.92
Education4	-0.26	0.18	-1.4	0.16
Education5	-0.20	0.19	-1.1	0.28
Education6	-0.50	0.19	-2.7	0.007
Exercise1	0.53	0.07	7.5	5.3e-14
Drink1	0.11	0.06	1.8	0.07
Theta	0.12	0.003		

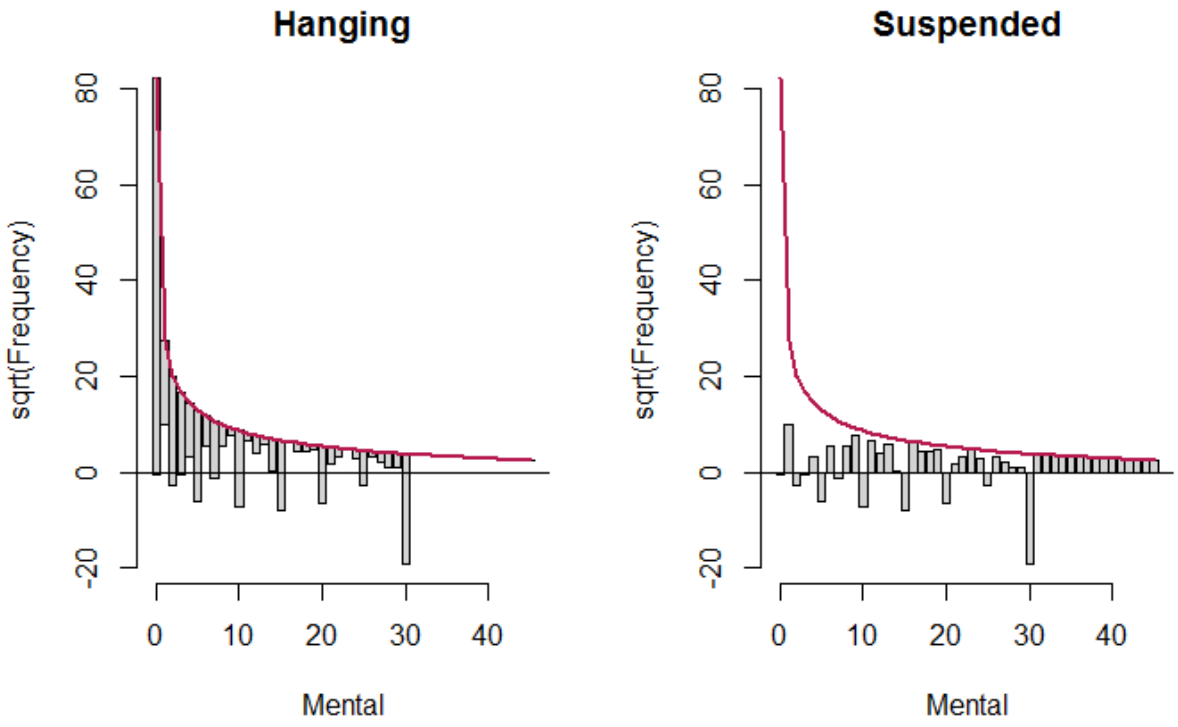

Figure 12. The hanging and suspended rootograms for number of days mental health not good when the education level is included as an attribute.

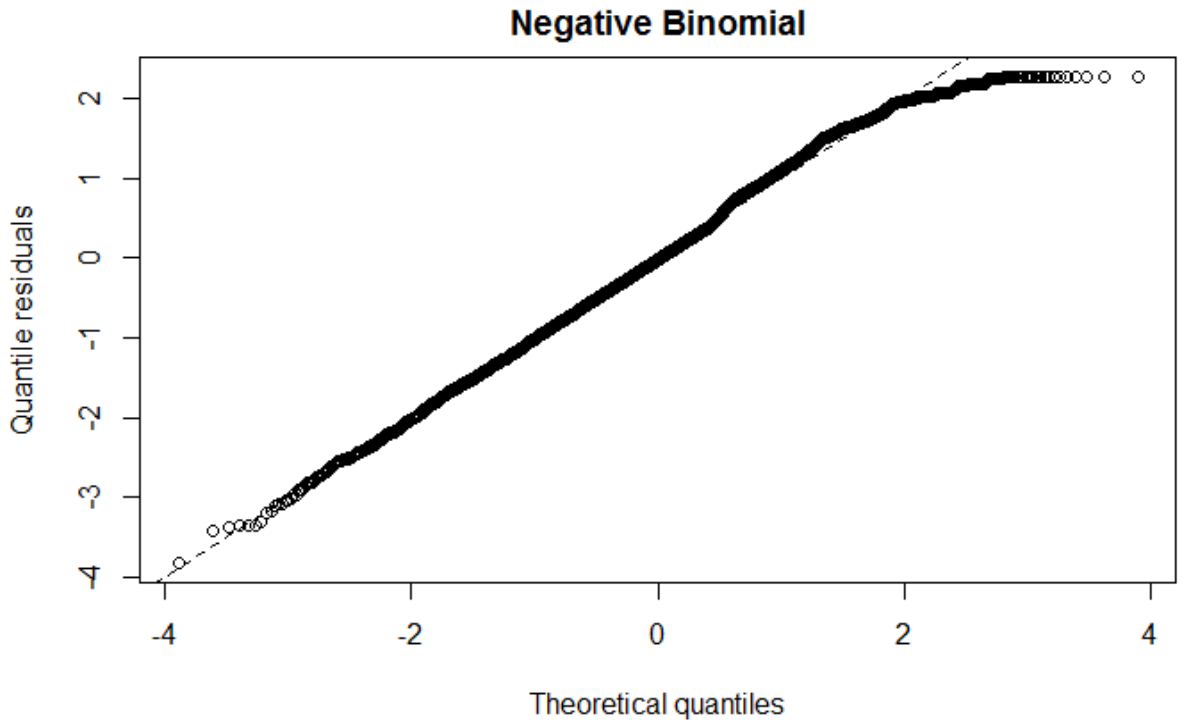

Figure 13. The Q-Q plots for the negative binomial model with the education level as an attribute.

5. Conclusion and discussion

In this study, I analyzed the behavioral survey data from the BRFSS to compare the effect of behaviors on mental health. Whether people exercise or drink occasionally are considered as the behavioral factors. The key challenges in formulating a regression model to fit the dataset were twofold. First, the multilevel responses, ranging from 0 to 30, need to be explained by two yes/no predictors on exercise and drink. Even after including the education level as an attribute, which has 6 levels, the response levels are greater than the combinations of the predictors. Second, the response data are censored on both sides and include the excess zeros and a large count for the maximum mentally sick days. Another issue with this large count for the maximum mentally sick days is that even people who reported their general health is excellent answered they were mentally sick for 30 days, indicating the accuracy of some of the dataset might not be satisfactory.

Based on the understanding on the dataset and limitations, I attempted to match the data using two regression model types: logistic-binomial model and negative binomial model. Even if the negative binomial model has a few apparent limitations such as generating overestimates that do not make sense and being unable to capture the right censored data points, it seems to describe the dataset better in general. The random effects of the states or the education levels were negligible, implying the dataset does not include any group-level traits. In terms of the attributes, exercise has the stronger effect on the mental health than drink. The effect of the education level only stands out for the highest education group.

A few model improvements should be considered in a future study. The models tested are overly simplified. Although the exploratory data analysis showed that the predictors (exercise,

drink, and education) have some correlations with the mental health measure, none of them might not be the major factor determining the mental health and key predictors might have been left out. Combining a different dataset with more descriptive predictors should be considered. In addition, a different regression model that can handle the censored data on both ends should be considered at least for a single-level regression.