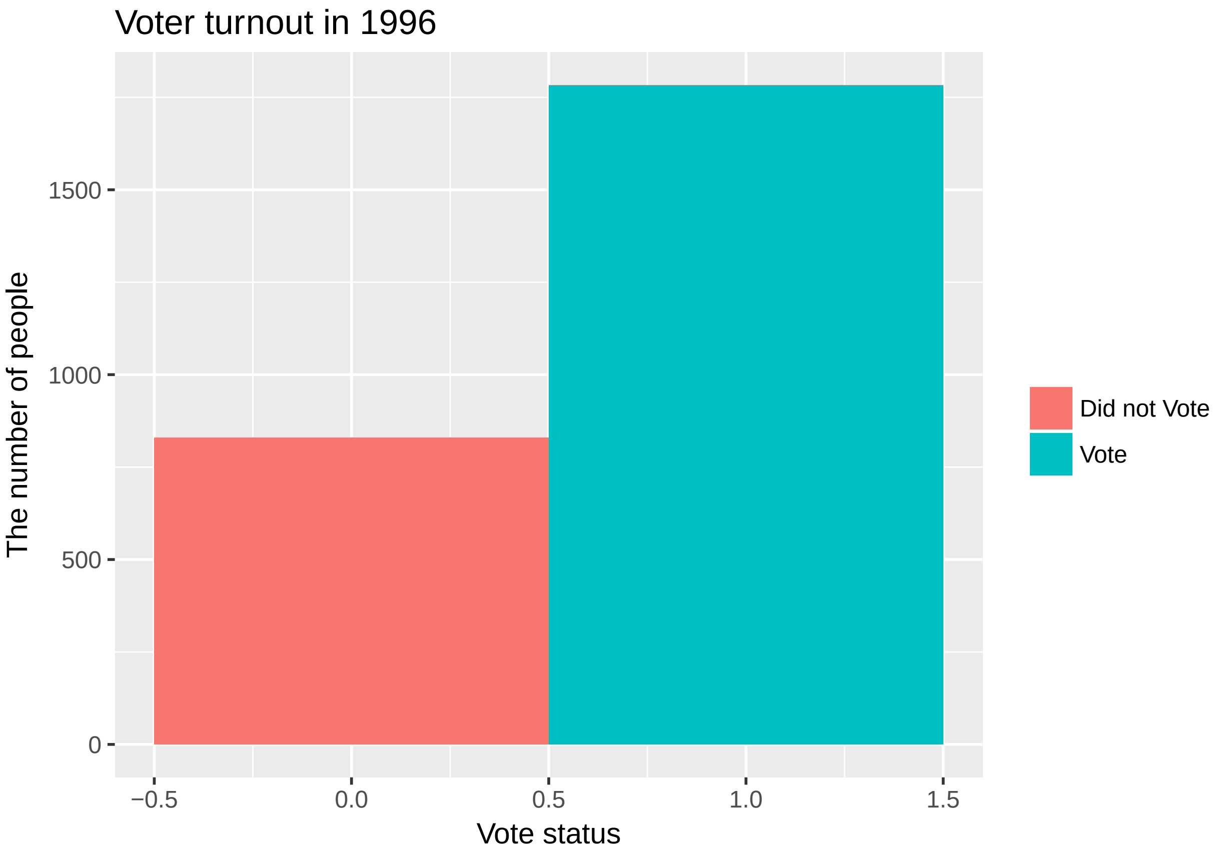
MACSS 30100 PS#6

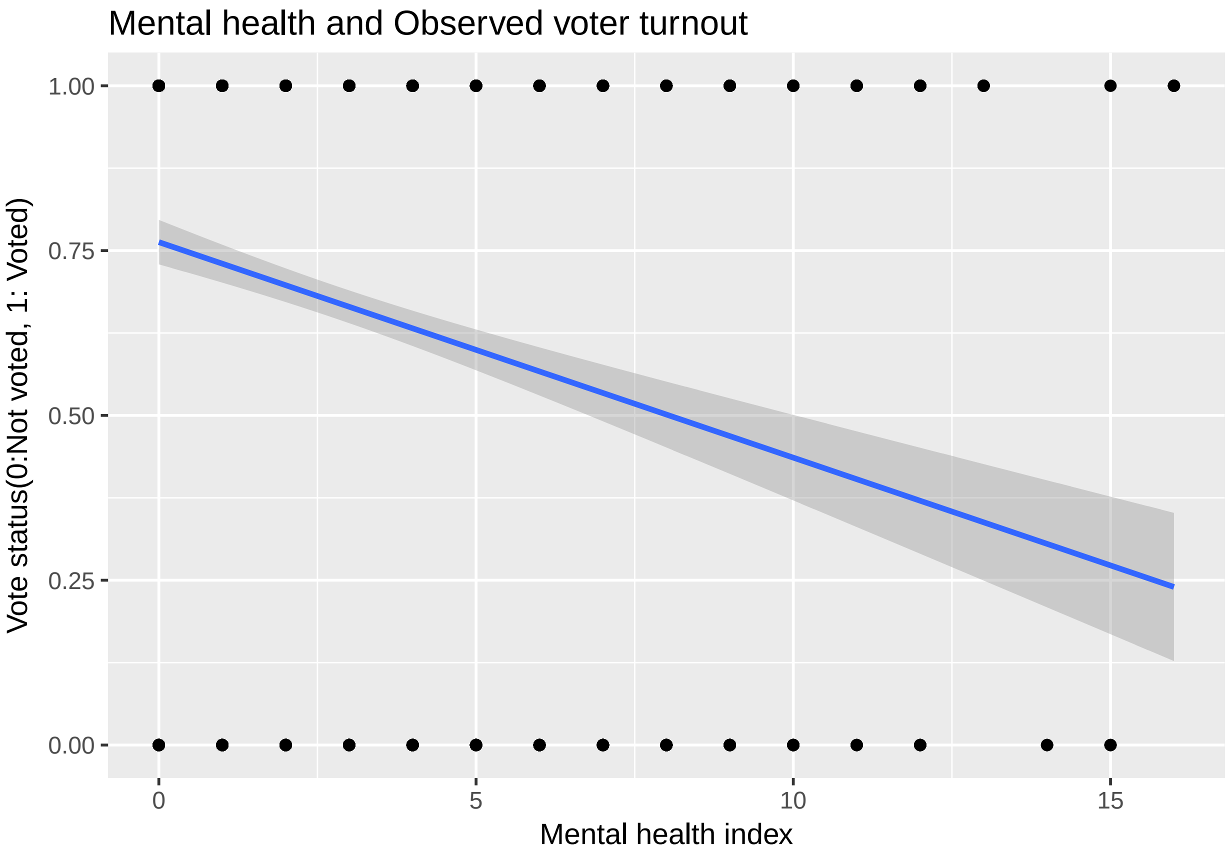
Alice Mee Seon Chung

2/18/2017

MACSS PS#6 Alice Mee Seon Chung

Describe the data





1. The unconditional probability of a given individual turning out to vote is 62.95904%.
2. The scatter plot and the linear smoothing line tell us that if a person had higher mental health index, which means in the severe depressed mood, it is likely that the person did not vote in 1996. The problems of the smooth linear line above is that it draws the response variables between the number 0 and 1, that is, it assumes the response variable is continous variables. But the response variables, vote96 ,are descrete variables (1: voted, 0: not voted). So when we interprete the smooth linear line, we can read the trend of the line, but the response variables can not be intepreted as discrete vote status.

Basic model

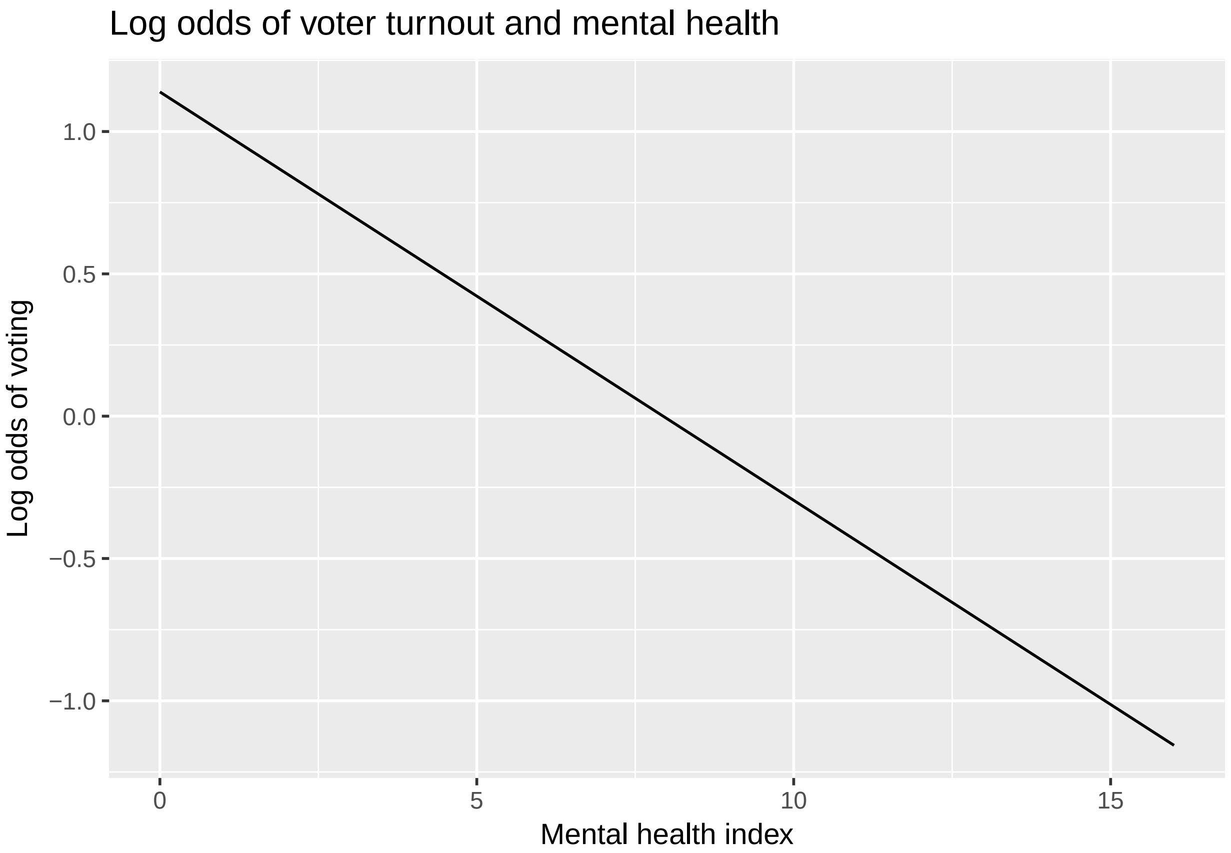
```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = "binomial", data = mental)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6834  -1.2977   0.7452   0.8428   1.6911
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.13921    0.08444  13.491  < 2e-16 ***
## mhealth_sum -0.14348    0.01969  -7.289 3.13e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.1  on 1321  degrees of freedom
## Residual deviance: 1616.7  on 1320  degrees of freedom
##   (1510 observations deleted due to missingness)
## AIC: 1620.7
##
## Number of Fisher Scoring iterations: 4
```

1. There is a statistically significant relationship between mental and voter turnout because its p-value is 3.13e-13 which is very close to zero, so it means that it is statisticaly significant.

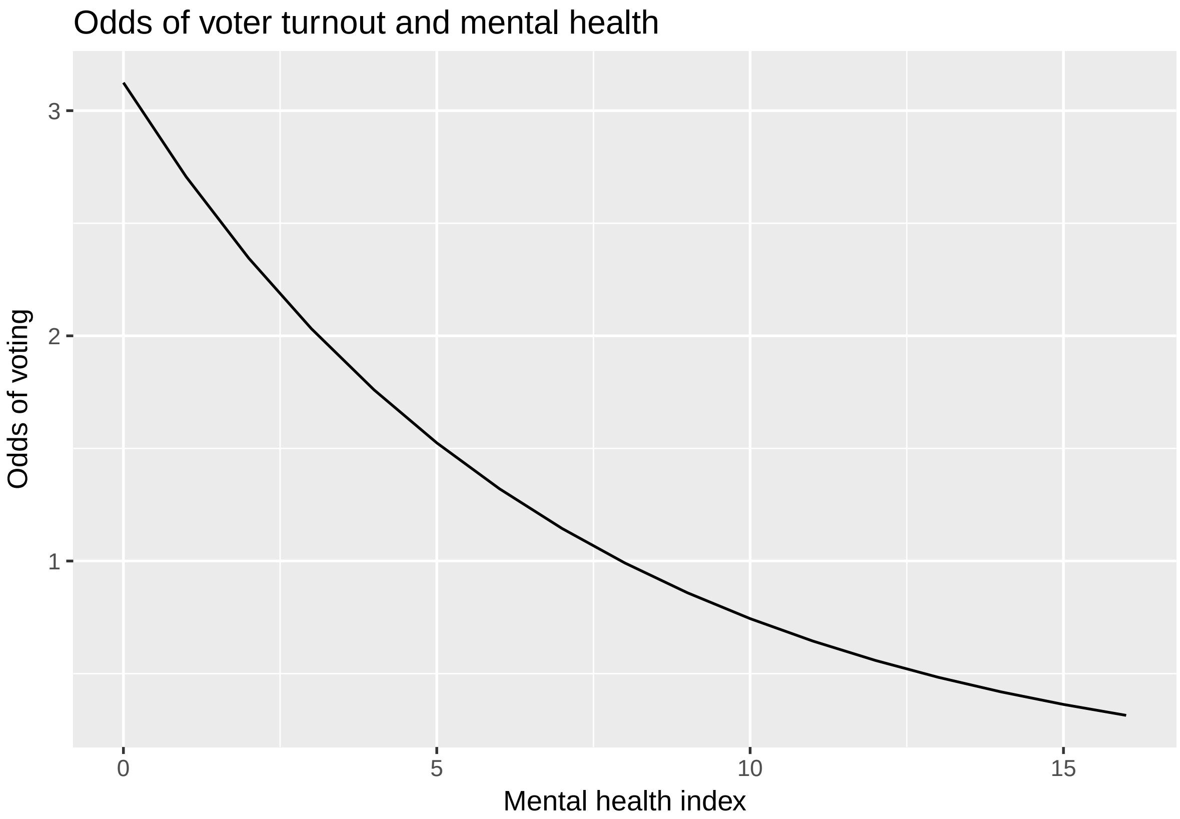
2.

```
##
## Call:
## glm(formula = logit ~ mhealth_sum, data = logit_mental)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.199e-14  -1.199e-14  -5.995e-15  -1.110e-16   3.553e-14
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.139e+00  3.614e-16  3.152e+15   <2e-16 ***
## mhealth_sum -1.435e-01  8.871e-17 -1.617e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.31325e-29)
##
##     Null deviance: 2.4360e+02  on 1321  degrees of freedom
## Residual deviance: 1.2293e-25  on 1320  degrees of freedom
## AIC: -81571
##
## Number of Fisher Scoring iterations: 1
```



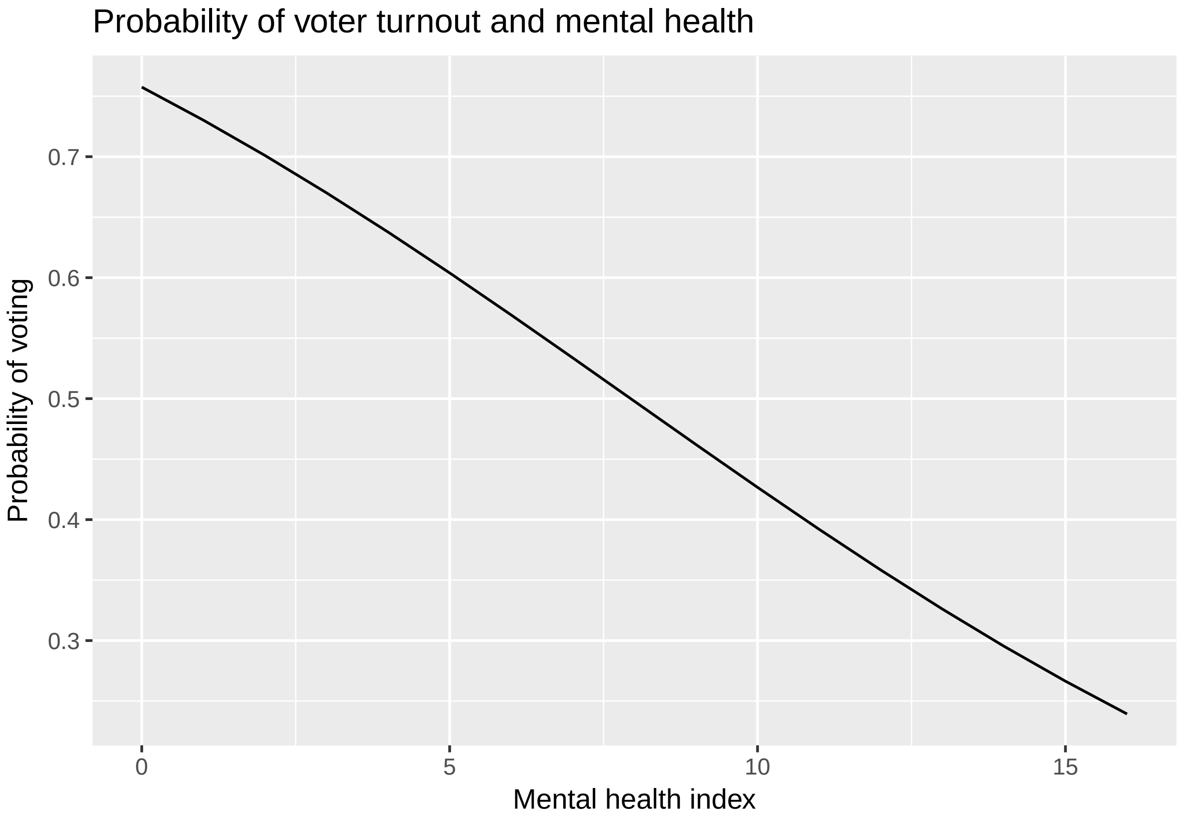
The coefficient of mental health index is -0.14348. It means that a one-unit increase in mental health index is associated with a 0.14348 decrease in the log-odds of voting.

3.



A one-unit increase in mental health index is associated with 1.154284 decrease in odds of voting on average.

4.



   A one-unit increase in mental health index is associated with 0.5358 decrease in probability of voting on average. The first difference for an increase in the mental health index from 1 to 2 is -0.0291782 and the difference for an increase in the mental health index from 5 to 6 is -0.0347782.

5.

With the threshold of .5, the accuracy rate is 67.78% and proportional reduction in error (PRE) is 1.62%, and the AUC for this model is 0.5401. I do not consider this model to be a good model. The proportional reduction in error is 1.62% and it means the statistical model reduced 1.62% of the prediction error. So it means it does not help in predicting. Moreover the AUC score is 0.6243 so it is 0.1243 higher than the 0.5 so it is hardly can tell as better classifier.

Multiple variable model

Among given 7 variables in the dataset, I choose 5 varaibles which are mental health index, age, education, married and income. I assume that mental health index affects on voting. Individuals with depression are less likely to participate in election than those without symptoms of depression. Also age affects on voting because older Americans, who typically have more time and higher incomes available to participate in politics, are more likely to participate in votings than younger Americans. In addition, education affects on voting because individuals with more years of education, who are generally more interested in politics, are more likely to participate in voting than individuals with fewer years of education. Also marriage affects on voting because marriage gives rise to a new and shared set of social and economic circumstances to individuals. Lastly, income level affects on voting and the assumption is based on the assumption of age.

1. The random component of the probability distribution for vote96 is binomial distribution. Each of $vote96_i$ is distributed as a binomial random variable.

$$Pr(\sum_{i=1}^{n} vote96_i = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

The linear predictor is

$$vote96_i = \beta_0 + \beta_1 mhealth\_sum + \beta_2 age + \beta_3 educ + \beta_4 married + \beta_5 inc10$$

The link function is

$$g(vote96_i) = \frac{e^{vote96_i}}{1 + e^{vote96_i}}$$

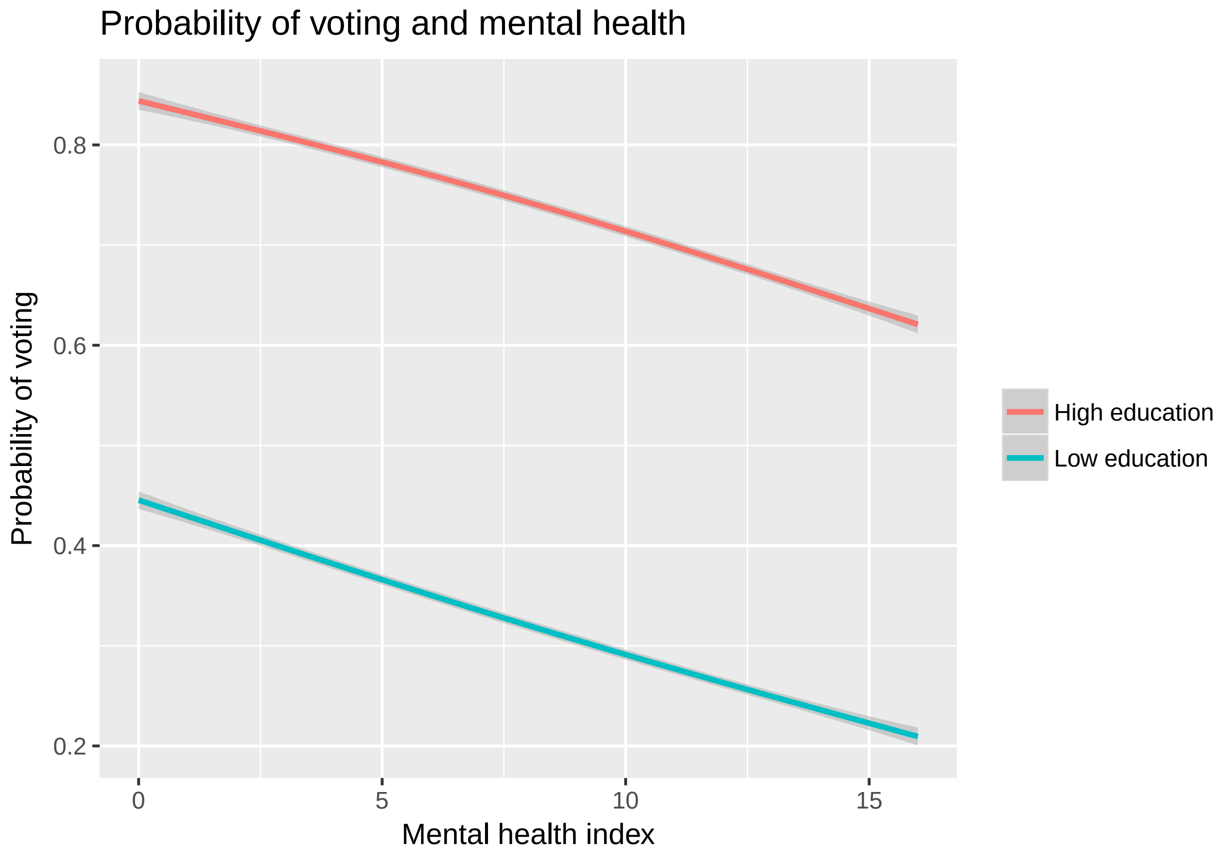
```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum + age + educ + married + inc10,
##     family = binomial, data = mental)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4907  -1.0297   0.5192   0.8418   2.1205
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.200133   0.497962  -8.435  < 2e-16 ***
## mhealth_sum -0.088332   0.023603  -3.742 0.000182 ***
## age          0.042113   0.004791   8.790  < 2e-16 ***
## educ         0.225271   0.029376   7.669 1.74e-14 ***
## married      0.293856   0.153009   1.921 0.054793 .
## inc10        0.066239   0.026207   2.528 0.011487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1468.3  on 1164  degrees of freedom
## Residual deviance: 1243.6  on 1159  degrees of freedom
##   (1667 observations deleted due to missingness)
## AIC: 1255.6
##
## Number of Fisher Scoring iterations: 4
```

β_0 of intercept in the multiple variable model is -4.200133, standard error is 0.497962 and p-value is less than 2e-16. β_1 for mental index sum is -0.088332,standard error is 0.023603 and p-value is 0.000182 β_2 for age is 0.042113, standard error is 0.004814 and p-value is less than 2e-16. β_3 for education is 0.225271, standard error is 0.029376 and p-value is 1.74e-14.β_4 for married is 0.293856, standard error is 0.153009 and p-value is 0.054793. β_5 for income is 0.066239, standard error is 0.026207 and and p-value is 0.011487.

3.

With the threshold of .5, the accuracy rate is 72.02% and proportional reduction in error (PRE) is 13.76%, and the AUC for this model is 0.7589. Comparing with above one-variable model, this multiple variable model fits the data well.

β_1 for mental index sum is -0.088332 and it means one-unit increase in mental health index is associated with 0.088332 decrease in log-odds of voting on average. β_2 for age is 0.042113, and it means one-unit increase in mental health index is associated with 0.042113 increase in log-odds of voting on average. β_3 for education is 0.225271 and it means one-unit increase in mental health index is associated with 0.225271 increase in log-odds of voting on average.β_4 for married is 0.293856 and it means one-unit increase in mental health index is associated with 0.293856 increase in log-odds of voting on average. β_5 for income is 0.066239 and it means one-unit increase in mental health index is associated with 0.066239 increase in log-odds of voting on average. Except married variables, each p-value of mental health index, age, education, and income are all less then 0.05 so it is statistically significant.



To see the relationshup between mental health index and probability of voting, I divide data into 2 groups, according to their education level and I set the criteria as 12 years because over 12 years means people get into college so I can assume that people had high education. From the above graph, we can tell as mental helth status gets severe, the probability of voting is decreasing. Morever, high education level people have higher probability of voting than the low education level people.

Part 2: Modeling tv consumption

Among given 11 variables in the dataset, I choose 4 varaibles which are education, grass, hrsrelax and black. I assume that education, grass, hrsrelax, race(black) affects on the number of hours of TV watched per day. I used step function to select best model.

1. The random component of the probability distribution for tvhours is poisson distribution. Each of tvhours is distributed as a poisson random variable.

$$P(tvhours_i = k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

The linear predictor is

$$tvhours_i = \beta_0 + \beta_1 educ + \beta_2 grass + \beta_3 hrsrelax + \beta_4 black$$

The link function for Poisson districution is log function,

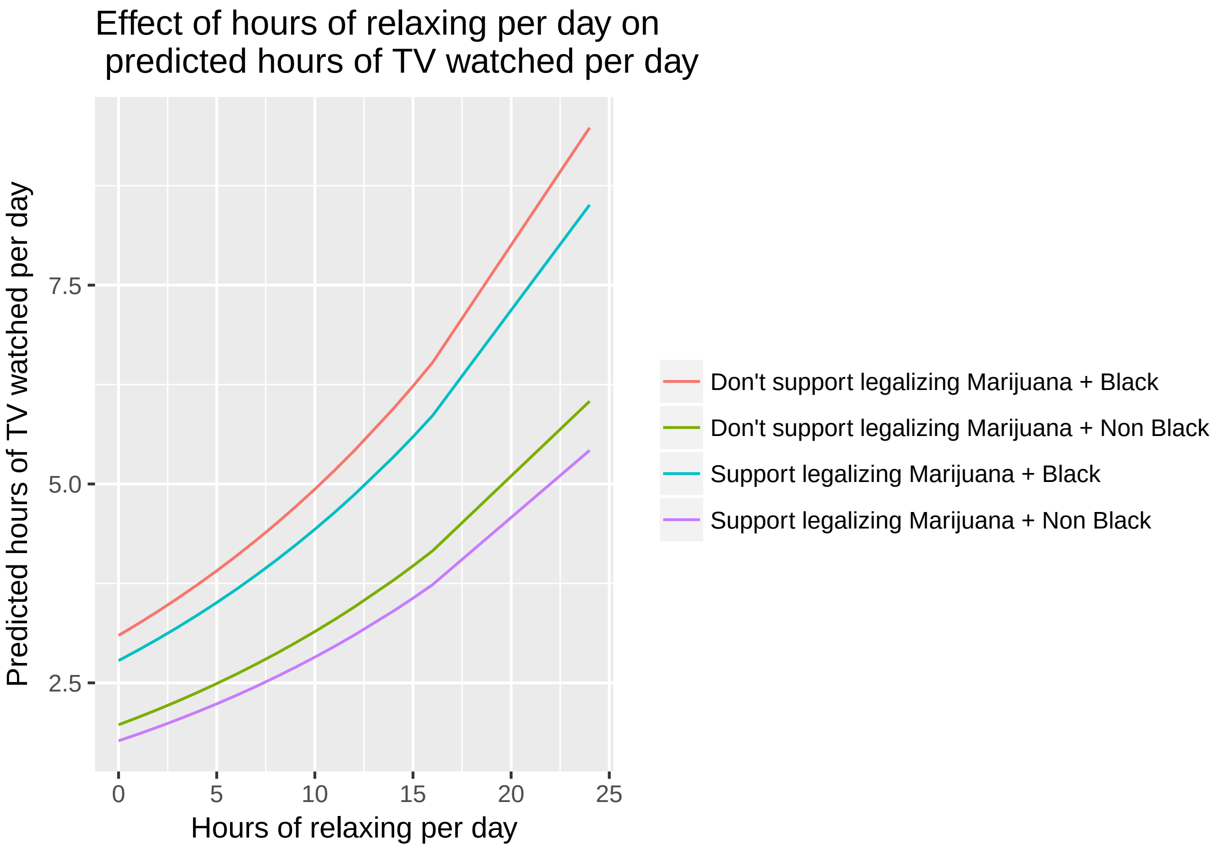
$$\mu_i = \ln(tvhours_i)$$

```
##
## Call:
```

```
## glm(formula = tvhours ~ educ + grass + hrsrelax + black, family = poisson,
##     data = gss2006)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0579  -0.7649  -0.0927   0.4634   5.4162
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.224882   0.169479   7.227 4.93e-13 ***
## educ        -0.038969   0.011063  -3.522 0.000428 ***
## grass       -0.107870   0.061855  -1.744 0.081174 .
## hrsrelax     0.046632   0.009462   4.928 8.29e-07 ***
## black        0.450638   0.072396   6.225 4.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 527.72  on 440  degrees of freedom
## Residual deviance: 442.48  on 436  degrees of freedom
## AIC: 1573.4
##
## Number of Fisher Scoring iterations: 5
```

β_0 of intercept in the model is 1.224882, standard error is 0.169479 and p-value is 4.93e-13. β_1 for education is -0.038969, standard error is 0.011063 and p-value is 0.000428. β_2 for grass is -0.107870, standard error is 0.061855 and p-value is 0.081174. β_3 for hrsrelax is 0.046632, standard error is 0.009462 and p-value is 8.29e-07.β_4 for black is 0.450638, standard error is 0.072396 and p-value is 4.83e-10.

With the threshold of .5, the accuracy rate is 22.68% and proportional reduction in error (PRE) is 0%, and the AUC for this model is 0.5488. β_1 for education is -0.038969 and it means one-unit increase in mental health index is associated with 0.038969 decrease in log-count of hours of TV watched per day on average. β_2 for grass is -0.107870, and it means one-unit increase in mental health index is associated with 0.107870 decrease in log-count of hours of TV watched per day on average. β_3 for hrsrelax is 0.046632 and it means one-unit increase in mental health index is associated with 0.046632 increase in log-count of hours of TV watched per day on average.β_4 for black is 0.450638 and it means one-unit increase in mental health index is associated with 0.450638 increase in log-count of hours of TV watched per day on average.



I will focus on one of 4 variables which is not a binary and has small p-value, hrsrelax. To see the relationshup between hours of relaxing per day and predicted hours of TV watched per day, I divide data into 4 groups, according to their race is black or non black and their opinions about legalizing marijuana and draw 4 seperate lines. From the above graph, we can tell hours of relaxing per day has positive effect on the hours of TV watched per day. To be more specific, race(black) has higher positive effect on the hours of TV watched per day since two graphs of two race(black) groups are above than two non-black groups.That is black people tend to spend more time on watching TV per day. Also when we categorized by opinions on legalizing marijuana, the graphs of two opposition on marijuana groups are both higher than two supporting groups when the race is same.

```
##
## Call:
## glm(formula = tvhours ~ educ + grass + hrsrelax + black, family = "quasipoisson",
##     data = gss2006)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0579  -0.7649  -0.0927   0.4634   5.4162
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.224882   0.178263   6.871 2.21e-11 ***
## educ        -0.038969   0.011637  -3.349 0.000882 ***
## grass       -0.107870   0.065061  -1.658 0.098040 .
## hrsrelax     0.046632   0.009952   4.686 3.74e-06 ***
## black        0.450638   0.076148   5.918 6.61e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for quasipoisson family taken to be 1.106342)
##
##     Null deviance: 527.72  on 440  degrees of freedom
## Residual deviance: 442.48  on 436  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

To test for under- or over-dispersion, I used quasipoisson model. From the summary, dispersion parameter for quasipoisson family is 1.106342. Since the parameter is greater than 1, the conditional variance of tvhours increases more rapidly than its mean, so we can say it is overdispersion.