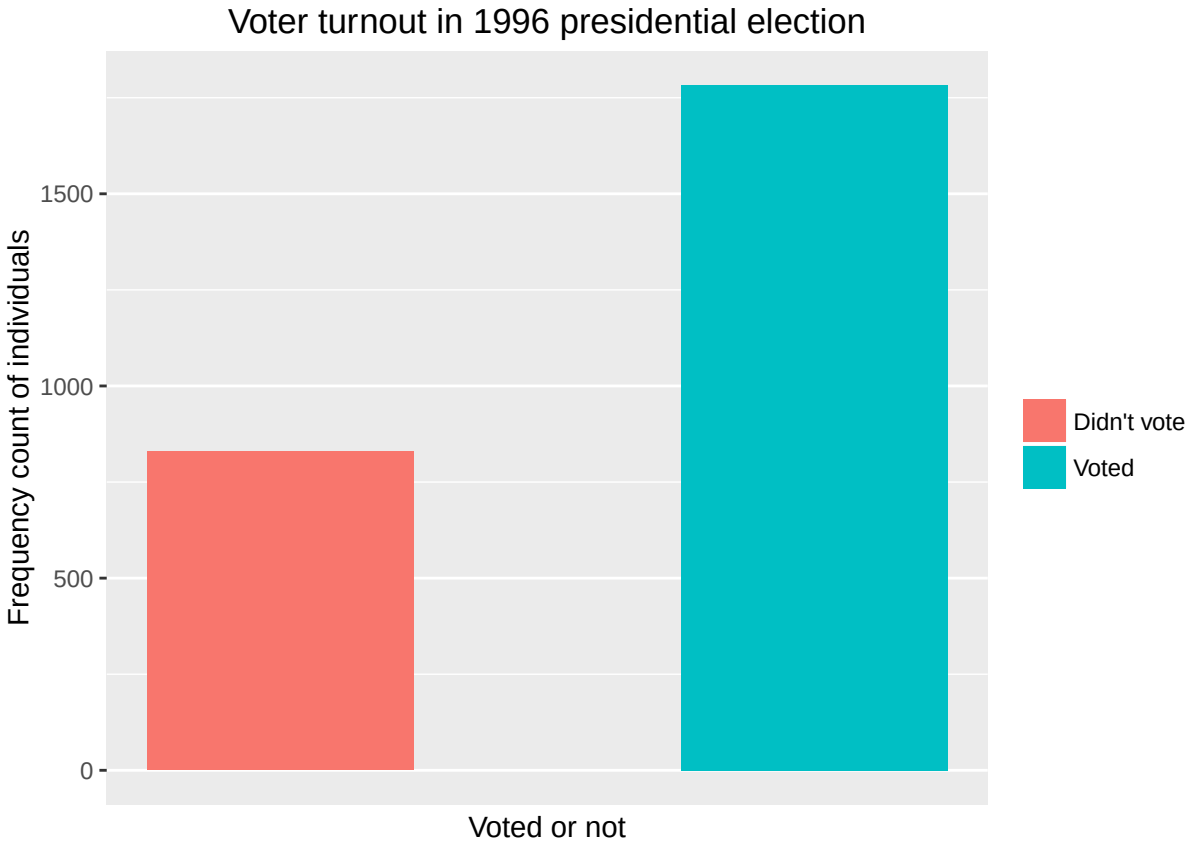
PS6: Generalized Linear Models

Ningyin Xu

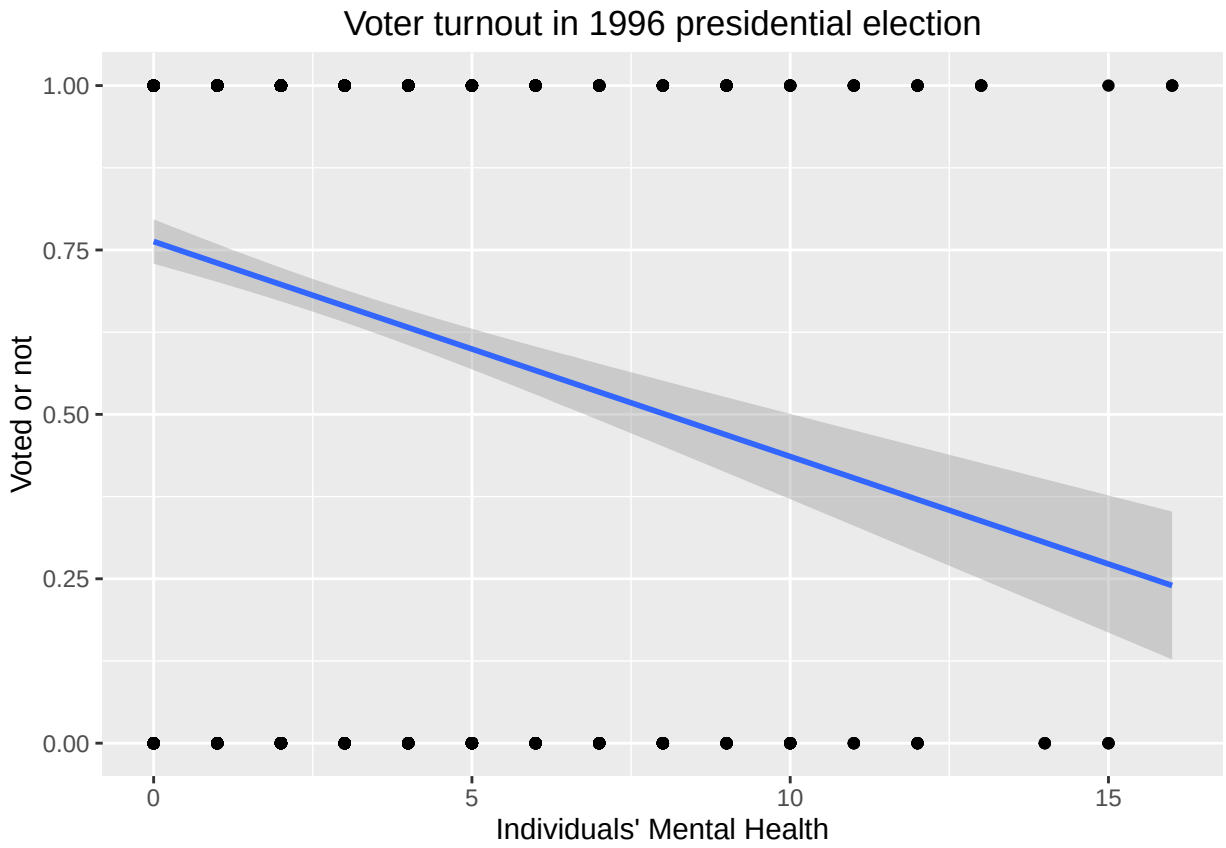
2/16/2017

Part 1: Modeling voter turnout

Problem 1. Describe the data



1). The unconditional probability of a given individual turning out to vote is 62.96%.



2). One can tell from the graph that there seem to be a negative correlation between individual's mental health and her/his voting decision, revealing depression decreases people's desire to participate in politics on some level.

However, this linear line has problems: the range of reponse is also problematic. The linear line shows the range of "voting decision" is any real number in `(0.25, 0.75)`, while response only have 2 values: 0, 1. This line doesn't explain the relationship between these two variables very well.

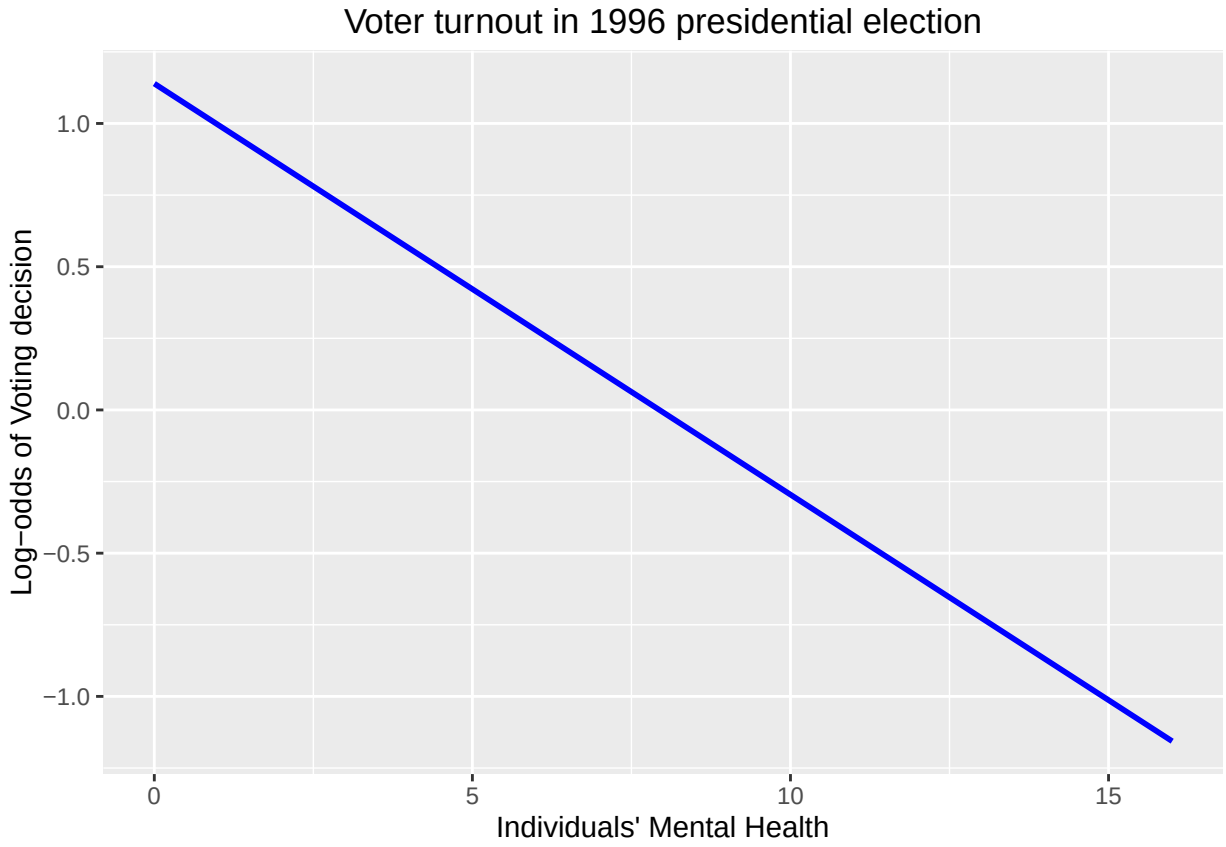
Problem 2. Basic Model

1). The relationship between mental health and voter turnour is statistically but not substantively significant.

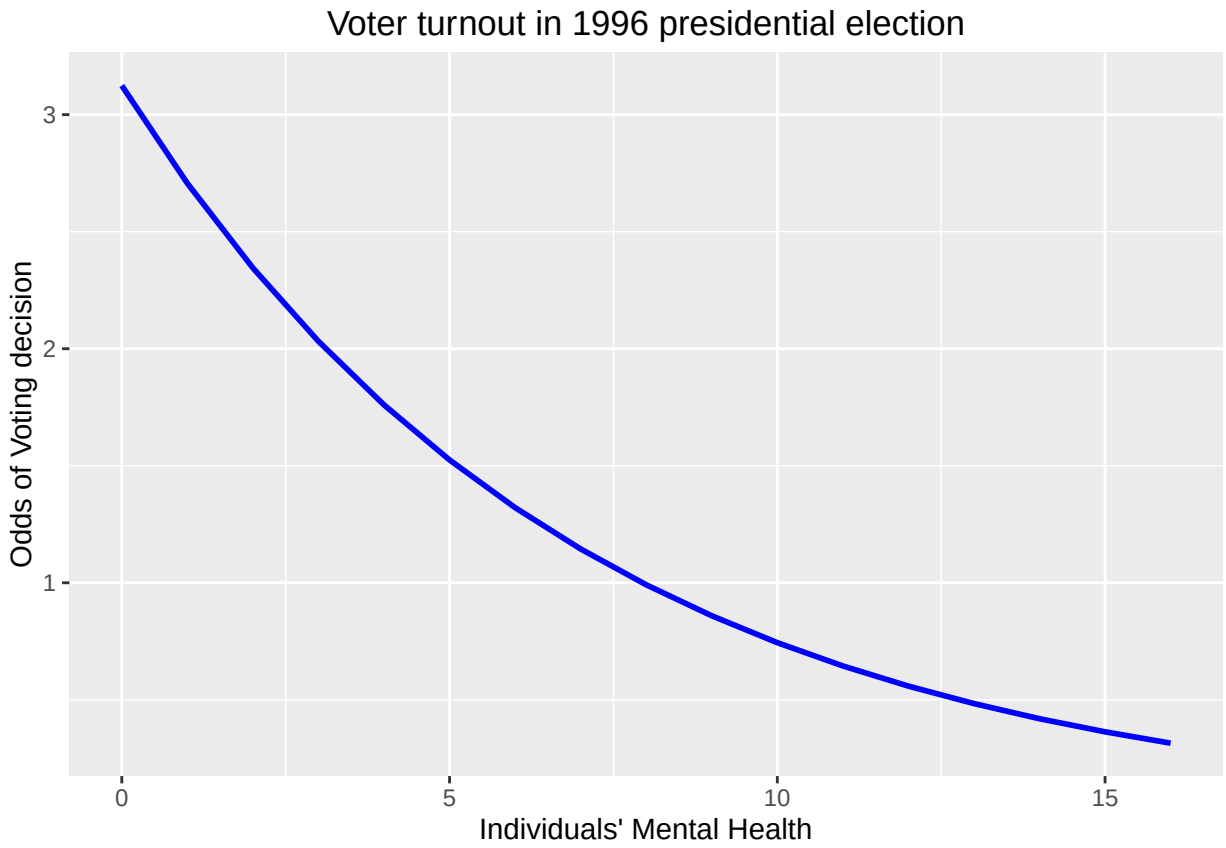
<1>. the p-value of mhealth_sum is very small, indicating the probablity to reject the null hypothesis is greather than 99%. This is statistical significance. <2>. The size of estimate parameter of mental health is -0.1435. It means with one unit increase in mental health, there would be 0.1435 decrease in the log-odds of voting decesion, so 1.1543 decrease in odds of voting. This effect size is relatively small, so the relationship is not substantively significant.

Table 1: Regression of voting decision on mental health

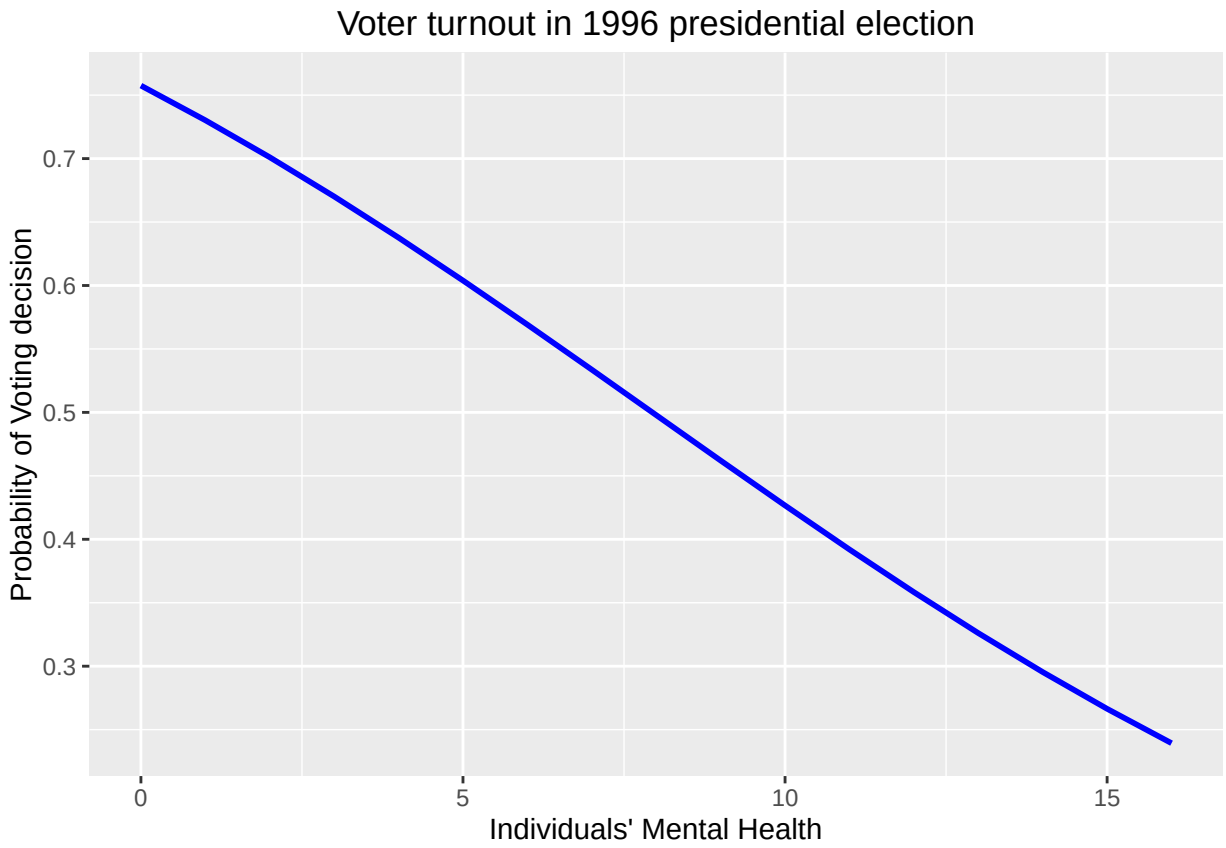
	Dependent variable:
	vote96
mhealth_sum	−0.143***
	(0.020)
Constant	1.139***
	(0.084)
Observations	1,322
Log Likelihood	−808.360
Akaike Inf. Crit.	1,620.720
Note:	*p<0.1; **p<0.05; ***p<0.01



2). With one unit increase in mental health (more depressed), there would be 0.1435 decrease in the log-odds of voting decesion.



3). With one unit increase in mental health (more depressed), there would be 1.1543 decrease in odds of voting.



4). With one unit increase in mental health (more depressed), there would be 0.5358 decrease in odds of voting. The first difference for an increase in the mental health index from 1 to 2 is -0.0292, for 5 to 6 is -0.0348. So the probability of voting would decrease by 2.92% when the individual's mental health increase from 1 to 2, the probability of voting would decrease by 3.48% when the individual's mental health increase from 5 to 6.

5). The accuracy rate, proportional reduction in error (PRE), and the AUC for this model are 67.78%, 1.62%, and 0.5401 respectively. This model is thus not so good. The accuracy rate seems okay but without baseline it doesn't say too much about the model's performance. The PRE says the statistical model reduces only little prediction error. The AUC is close to 0.5, which is the auc under random condition.

Problem 3. Multiple Variable Model

1). I chose the model based on several theoretcial assumptions. First, age might affected people's voting decision. Older Americans, who typically have more time and higher incomes available to participate in politics, should be more likely to participate in elections than younger Americans. Second, individuals with more years of education, who are generally more interested in politics and understand the value and benefits of participating in politics, are more likely to participate in elections than individuals with fewer years of education. Finally, Depression increases individuals' feelings of hopelessness and political efficacy, so depressed individuals will have less desire to participate in politics.

For the model I chose, the Probability distribution is the Bernoulli distribution:

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

The linear predictor is:

$$vote96_i = \beta_0 + \beta_1 MentalHealth_i + \beta_2 Age_i + \beta_3 Education_i$$

The link function is:

$$\pi_i = \frac{e^{vote96_i}}{1 + e^{vote96_i}}$$

The estimation result of this model is:

Table 2:

	Dependent variable:
	vote96
mhealth_sum	-0.099^{***}
	(0.021)
age	0.045^{***}
	(0.004)
educ	0.260^{***}
	(0.027)
Constant	-4.375^{***}
	(0.463)
Observations	1,317
Log Likelihood	−710.201
Akaike Inf. Crit.	1,428.403
Note:	*p<0.1; **p<0.05; ***p<0.01

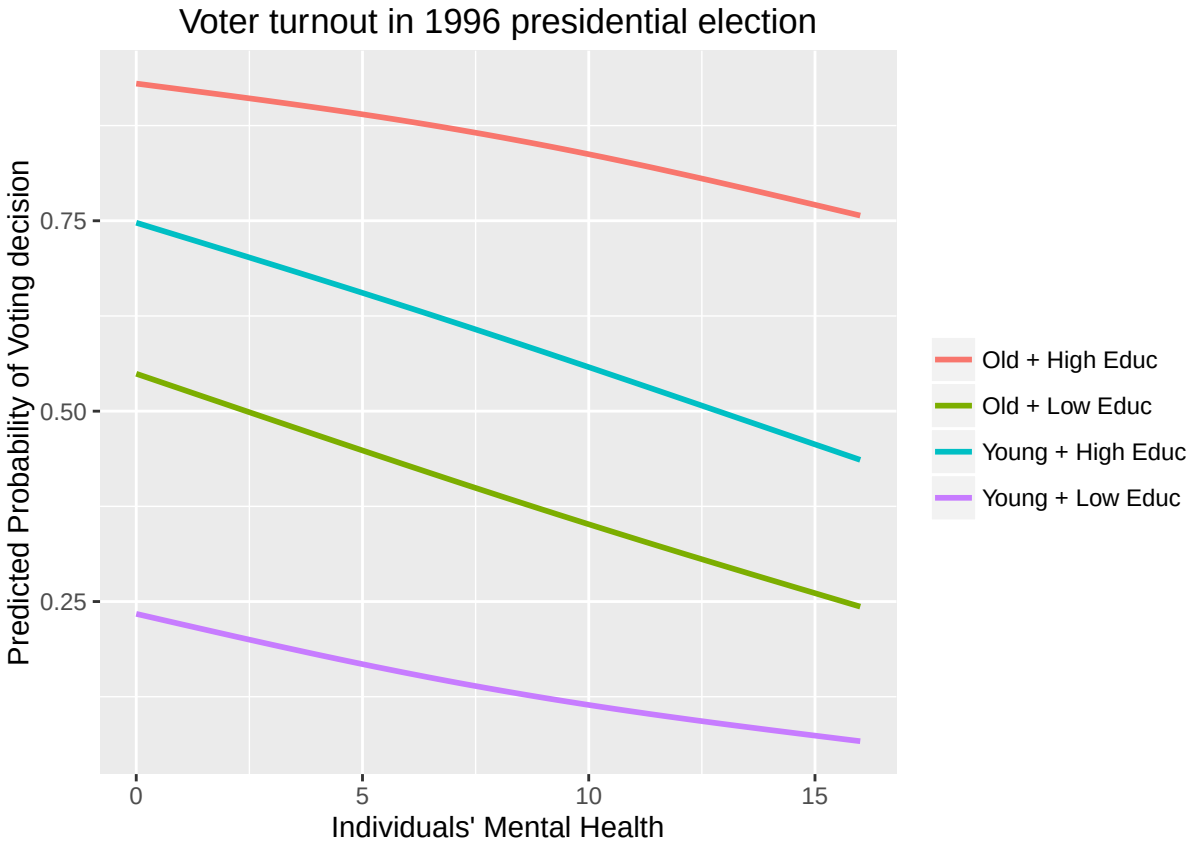
3). From the results of the model, one can see the model performs well: <1>. The chosen variables all have significant relationship with the response. The three predictors we chose are all statistically significant. <2>. The model's fits the reality relatively well. The accuracy rate, proportional reduction in error (PRE), and the AUC for this model are 72.29%, 15.12%, and 0.6379 respectively. Comparing to the signle-variable model, the accuracy rate has been improved, there's more reduction in error, and the AUC is enlarged. One can also use p-value 0, which is way less than 0.05, showing that this model as a whole fits significantly better than an empty model.

Based on the fitness of this model, one could interpret the relationship between response and the three predictors using this model. The size of estimate parameter of mental health is -0.0985. It means with one unit increase in mental health, there would be 0.0985 decrease in the log-odds of voting decesion, so 1.1036 decrease in odds of voting.

The size of estimate parameter of age is 0.0449. It means with one unit increase in age, there would be 0.0449 increase in the log-odds of voting decesion, so 1.046 increase in odds of voting.

The size of estimate parameter of education level is 0.2605. It means with one unit increase in education level, there would be 0.2605 increase in the log-odds of voting decesion, so 1.2975 increase in odds of voting.

First difference could be demonstrated from mental health, age, and education level. To compare with the basic model before, I'm only gonna use mental health here, controlling age and education level as 22 and 15. The first difference for an increase in the mental health index from 1 to 2 is -0.0238, for 5 to 6 is -0.0246. So the probability of voting would decrease by 2.38% when the individual's mental health increase from 1 to 2, the probability of voting would decrease by 2.46% when the individual's mental health increase from 5 to 6.



To better illustrate the relationship between the response and predictors, I use the above graph, in which I divide age and educ (education level) to two groups based on their median value. And from the graph one could tell, individual's mental health has negative influence on voting decision in general, age has a positive influence since both curves for older people are higher comparing to younger people from this plot, and people who receive higher education seems to be more willing to vote. This proves the theorectical assumptions.

Part 2: Modeling TV Consumption

Estimate a regression model

1). I used the backward AIC selection to choose the model. For the model I chose, the Probability distribution is:

$$Pr(tvhours = tvhours_i|\mu) = \frac{\mu^k e^{tvhours_i}}{tvhours_i!}$$

The linear predictor is:

$$tvhours_i = \beta_0 + \beta_1 educ_i + \beta_2 grass_i + \beta_3 hrsrelax_i + \beta_4 black_i$$

The link function is:

$$\mu_i = ln(tvhours_i)$$

The estimation result of this model is:

3). From the results of the model, one can see the model performs well: <1>. The chosen variables all have significant relationship with the response (all the predictors except grass is significant at the 95% level, grass is significant at 90% level). The four predictors we chose are all statistically significant.

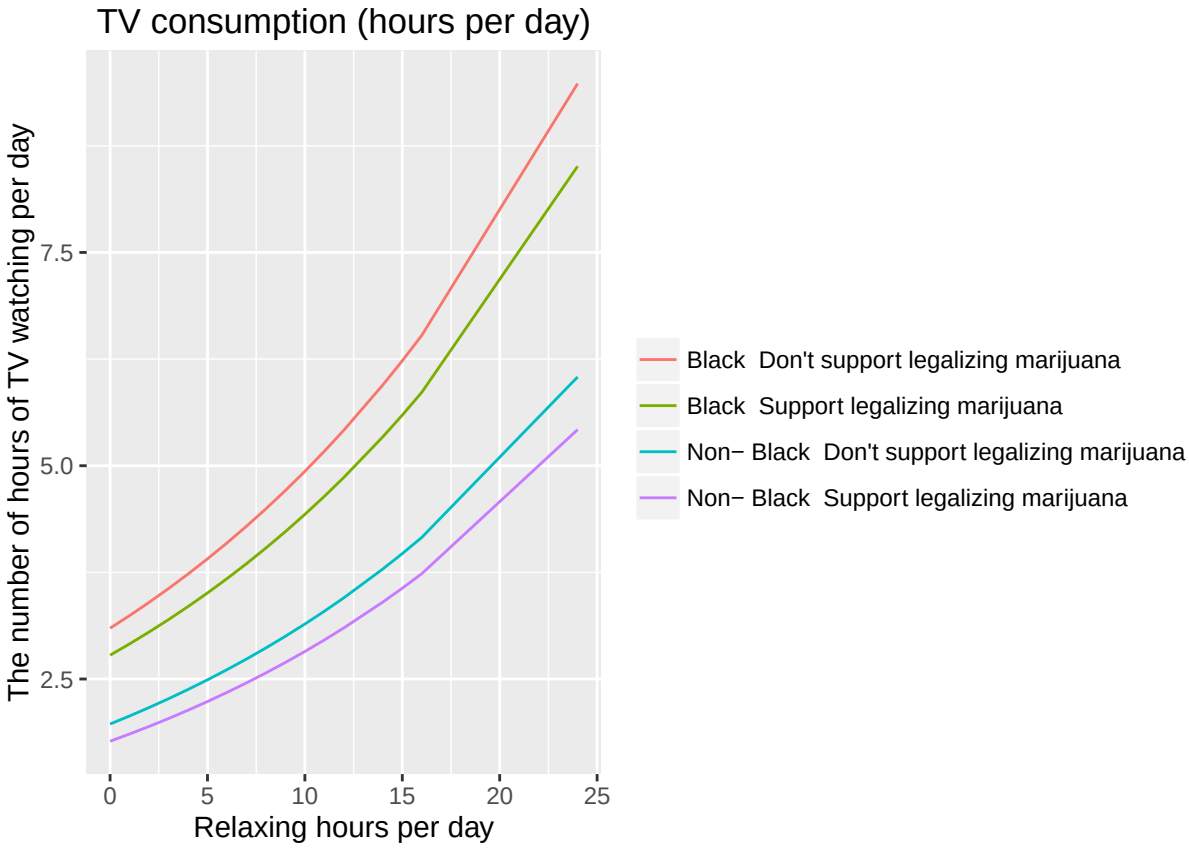
Table 3:

	Dependent variable:
	tvhours
educ	-0.039^{***}
	(0.011)
grass	-0.108^{*}
	(0.062)
hrsrelax	0.047^{***}
	(0.009)
black	0.451^{***}
	(0.072)
Constant	1.225^{***}
	(0.169)
Observations	441
Log Likelihood	−781.721
Akaike Inf. Crit.	1,573.442
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##      res.deviance  df         p
## [1,]     442.4819 436 0.4047867
```

<2>. One can use the residual deviance to perform a goodness of fit test for the overall model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. Therefore, if the residual difference is small enough, the goodness of fit test will not be significant, indicating that the model fits the data. We conclude that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant (p-value is 0.4048).

Based on the fitness of this model, one could interpret the relationship between response and the four predictors using this model. The size of estimate parameter of education is -0.039. It means with one unit increase in education level, there would be 0.039 expected decrease in the log count of tv consumption hours. The size of estimate parameter of grass is -0.1079. It means if an individual thinks marijuana should be legalized, there would be 0.1079 expected decrease in the log count of tv consumption hours. The size of estimate parameter of relaxing hours is 0.0466. It means with one unit increase in education level, there would be -0.0466 expected increase in the log count of tv consumption hours. The size of estimate parameter of relaxing hours is 0.4506. It means if an individual is black, there would be -0.4506 expected increase in the log count of tv consumption hours.



To better illustrate the relationship between the response and predictors, I use the above graph, in which I fixed education level on its median value. And from the graph one could tell, individual's relaxing hours per day has positive influence on tv consumption in general, being black has a positive influence since both curves for black people are higher comparing to non-black people from this plot, and people who doesn't support legalizing marijuana seems to spend more time on tv.