Problem set 6#Xuancheng Qian

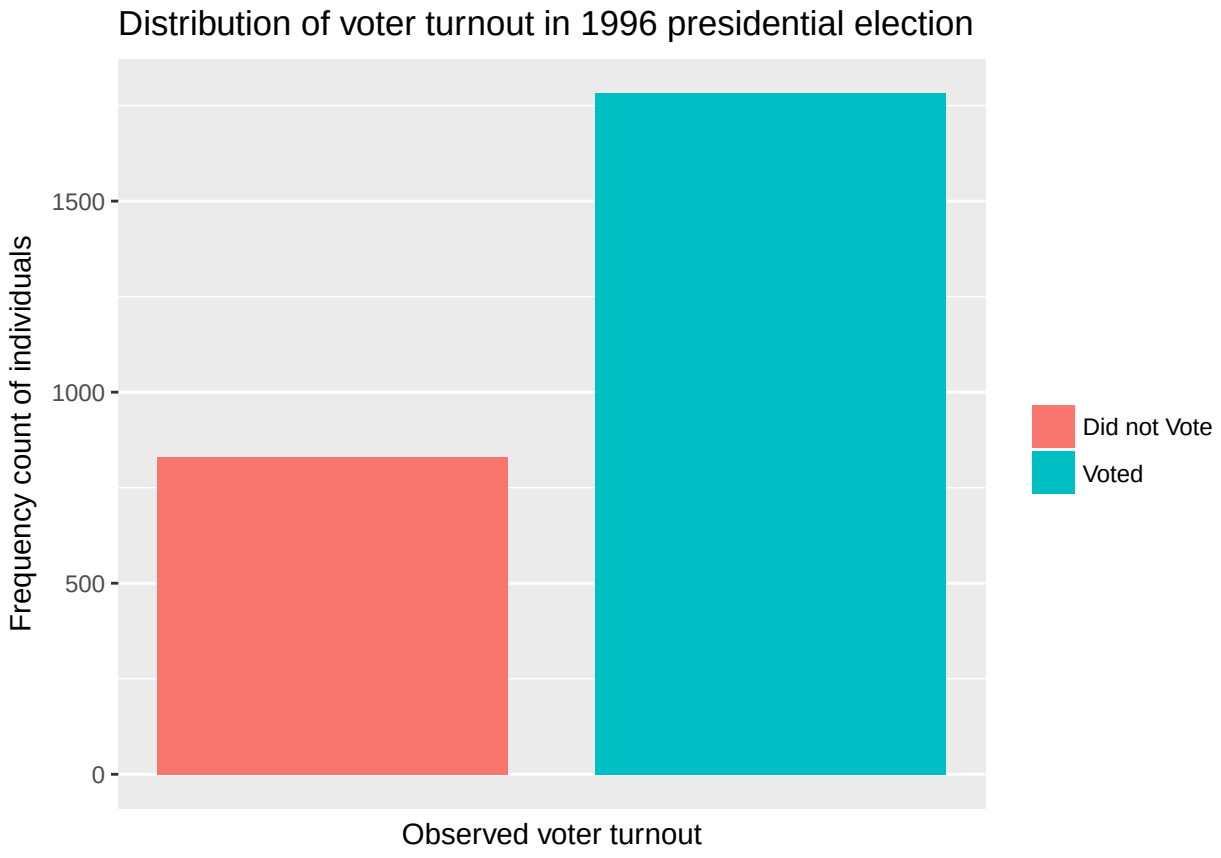
Xuancheng Qian

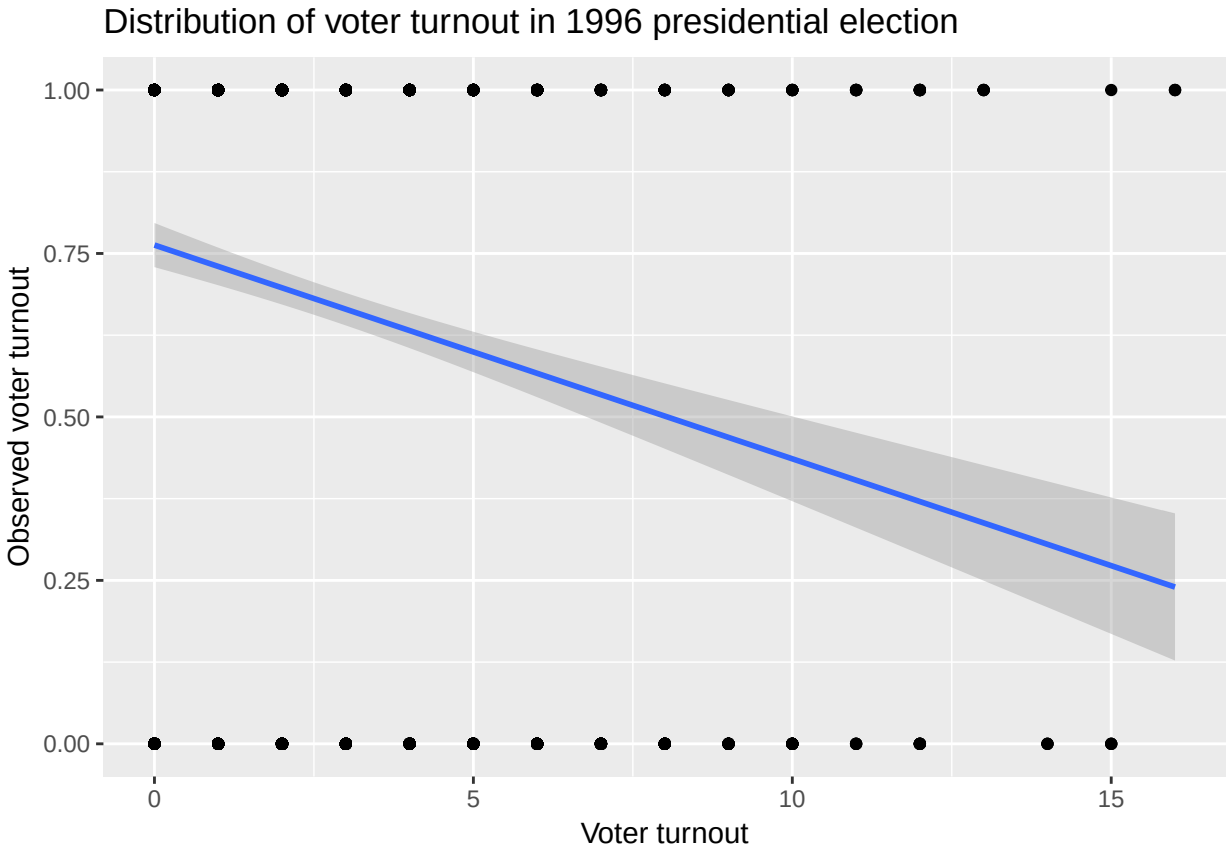
2/20/2017

Part 1: Modeling voter turnout

Describe the data (1 point)

1. Plot a histogram of voter turnout. Make sure to give the graph a title and proper x and y-axis labels. What is the unconditional probability of a given individual turning out to vote?
2. Generate a scatterplot of the relationship between mental health and observed voter turnout and overlay a linear smoothing line. What information does this tell us? What is problematic about this linear smoothing line?





- For the histogram plot, we remove 219 observations with missing values in voter turnout. And the unconditional probability of a given individual turning out to vote is 62.96%.
- From the scatterplot, we can see that in general, there exists a negative relationship between mental health and observed voter turnout. Higher mental health value (worse mental health) would decrease respondent's willingness to vote. Our voting turnout is binary (voted or not), however, this linear line indicates the response is continuous which did not explain the relationship between mental health and voter turnout well.

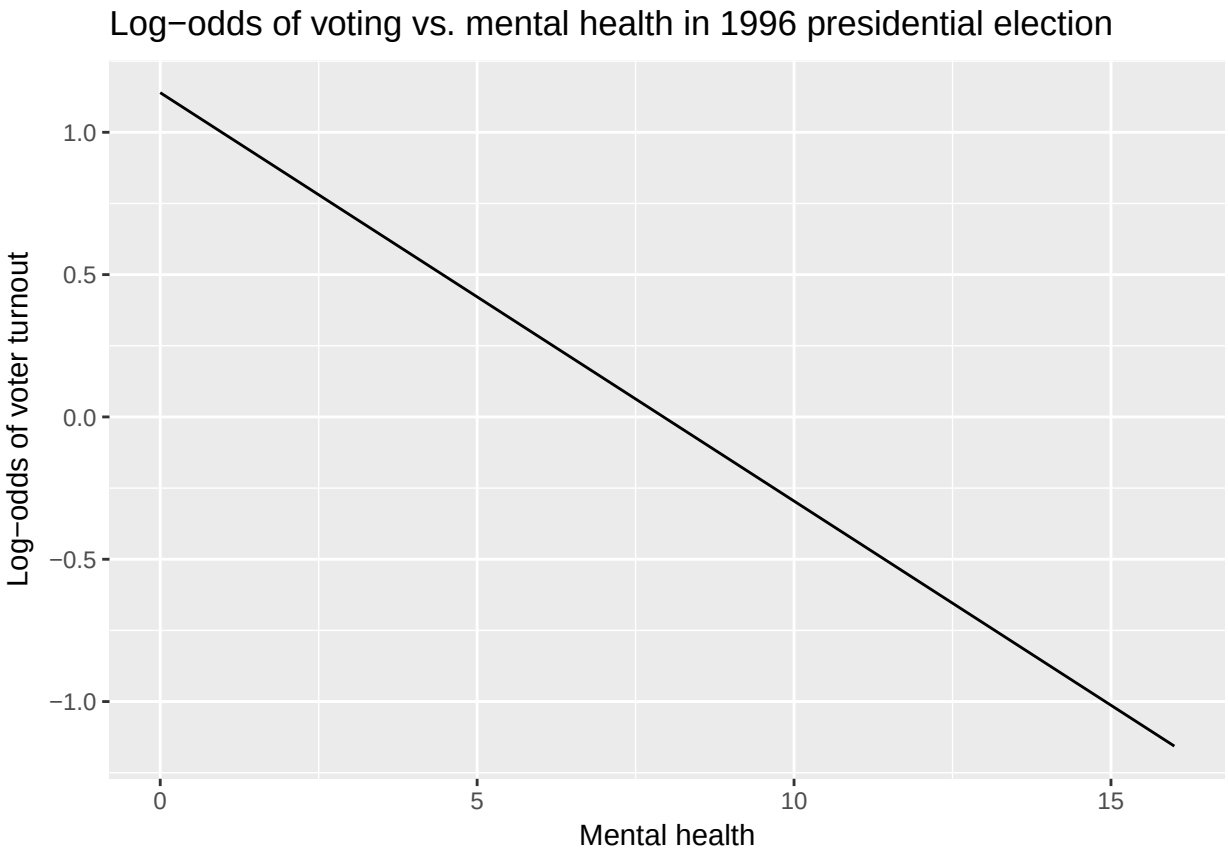
Basic model (3 points)

Estimate a logistic regression model of the relationship between mental health and voter turnout.

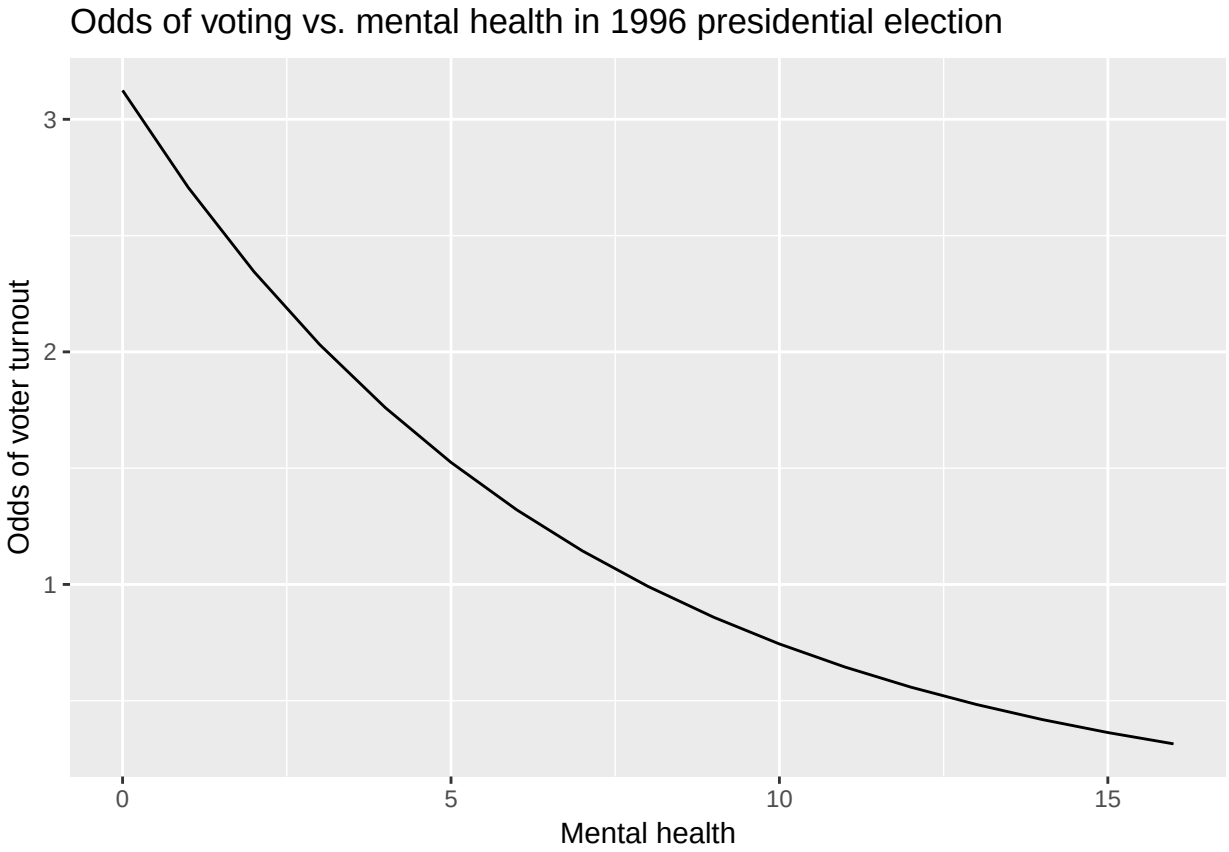
- The relationship between mental health and voter turnout is statistically significant as p-value is very small, which is 3.13e-13. The coefficient of mental health is -0.1434752 , which indicates that in general, the odds ratio associated with one unit increase in mental health is 0.8663423, which is 14.347% decrease in the odds of voting and this indicates that relationship is substantively significant in negative directions.

Table 1: Logistic Regression Results

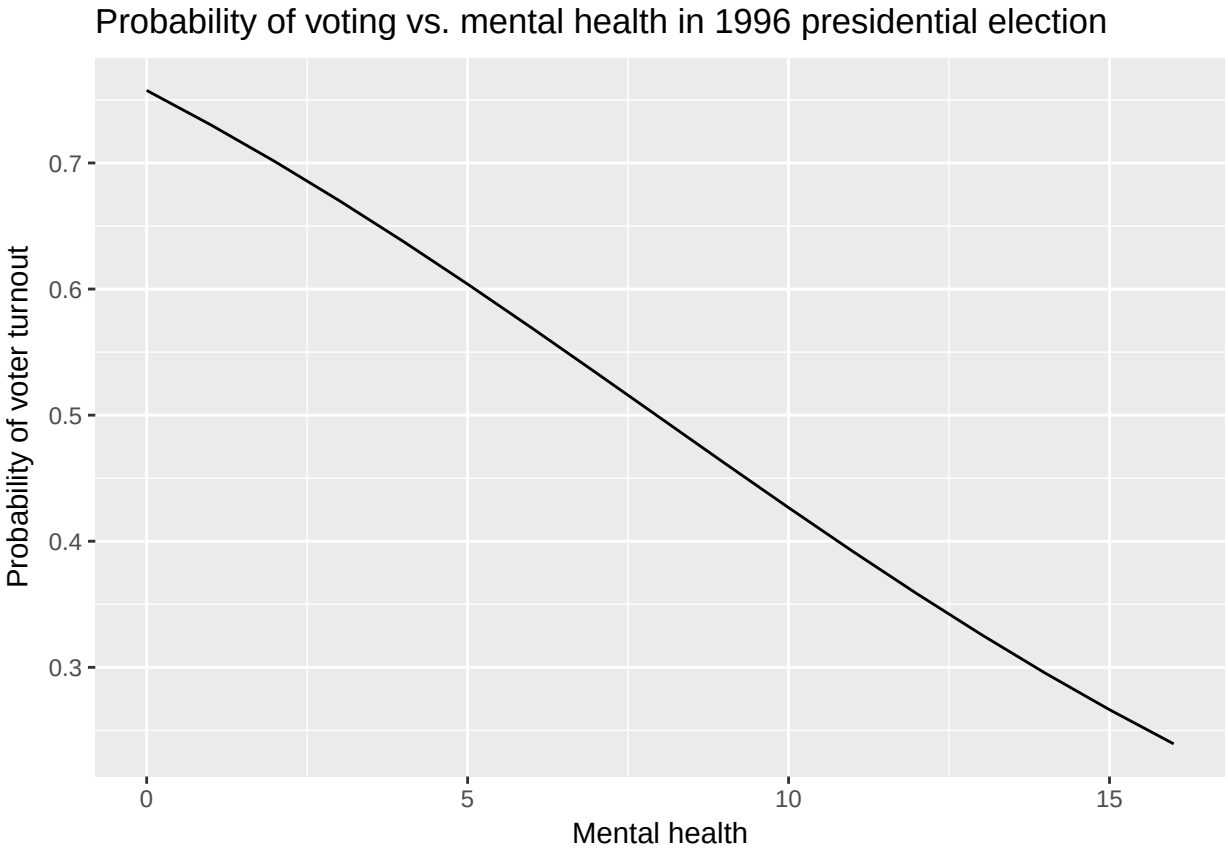
	Dependent variable:
	vote96
mhealth_sum	−0.143***
	(0.020)
Constant	1.139***
	(0.084)
Observations	1,322
Log Likelihood	−808.360
Akaike Inf. Crit.	1,620.720
Note:	*p<0.1; **p<0.05; ***p<0.01



- One unit increase in mental health (worse mental health) would lead to 0.1434752 decrease by average in the log-odds of voting decision.



- One unit increase in mental health (worse mental health) would lead to 14.347% decrease by average in the odds of voting decision.



- One unit increase in mental health (worse mental health) would lead to 0.464 change by average in the probability of voting decision. The first difference for an increase in the mental health index from 1 to 2 is -0.0291782, so the probability of voting would decrease by 2.92% when the mental health increase from 1 to 2. The first difference for an increase in the mental health index from 5 to 6 is -0.0347782, so the probability of voting would decrease by 3.48% when the mental health increase from 5 to 6.

The accuracy rate is 67.78%, the proportional reduction in error (PRE) is 1.62%, and the AUC for this model is 62.43%. The model is not a good model. The proportional reduction in error says that this model only reduces 1.62%, which is very small, and the AUC is very close to the performance of random condition.

Multiple variable model (3 points)

- The probability distribution is Bernoulli distribution.

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

- The linear predictor:

$$vote96_i = \beta_0 + \beta_1 mhealth_sum + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 female + \beta_6 married + \beta_7 inc10$$

- The link function:

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

- In this multiple variable logistic regression model, the response variable is observed voter turnout, which is binary (voted or did not vote). The predictors include mental health index, age, education, race, gender, marital status and family income.The regression results indicate that four of the coefficients

Table 2: Multiple Logistic Regression Results

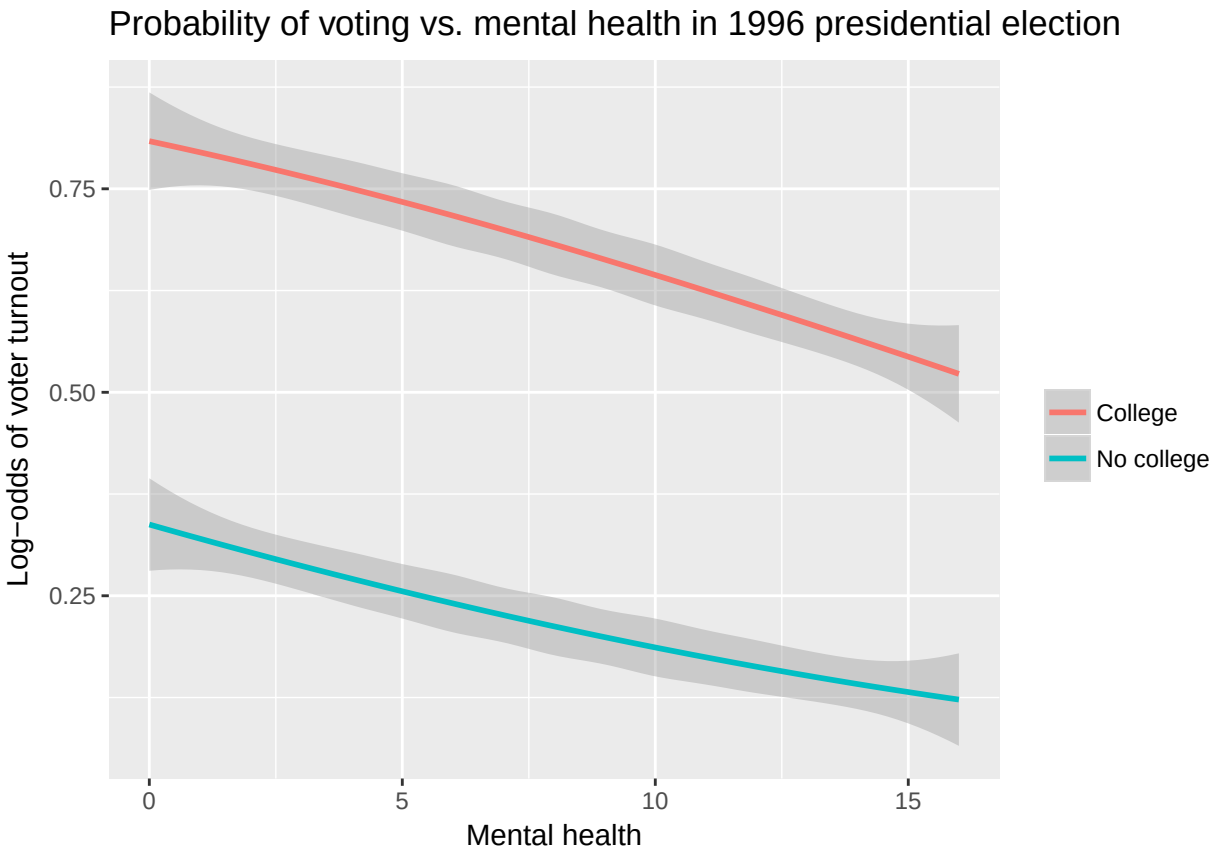
	Dependent variable:
	vote96
mhealth_sum	-0.089^{***}
	(0.024)
age	0.043^{***}
	(0.005)
educ	0.229^{***}
	(0.030)
black	0.273
	(0.203)
female	-0.017
	(0.140)
married	0.297^{*}
	(0.153)
inc10	0.070^{***}
	(0.027)
Constant	-4.304^{***}
	(0.508)
Observations	1,165
Log Likelihood	-620.883
Akaike Inf. Crit.	1,257.767
Note:	*p<0.1; **p<0.05; ***p<0.01

are statistically significant; these coefficients are, respectively, -0.089102 for the mental health index, 0.042534 for age, 0.228686 for education and 0.069614 for income.

The coefficient of mental health index is -0.089102, which means one unit increase in mental health (worse mental health), there would be `r 0.089102` decrease by average in the log-odds of voting, and 0.9147523 decrease by average in odds of voting. The coefficient of age is 0.042534, which means one unit increase in age, there wouold be 1.0434515 increase by average in odds of voting. The coefficient of education is 0.228686, which means one unit increase in education level, there would be 1.2569473 increase by average in odds of voting. The coefficient of income is 0.069614, which means one unit increase in income there would be 0.9147523 increase by average in odds of voting.

The accuracy rate is 72.36%, the proportional reduction in error (PRE) is 14.81%, and the AUC for this model is 75.96%. The model is not a good model. The proportional reduction in error says that this model only reduces 14.8%, which is very small, and the AUC is very close to the performance of random condition.

We can also compare the first difference in mental health index with previous model. To fix other predictors, I will use 25-year-old black female with 12 years of education, single, and income of $50,000. The first difference for an increase in the mental health index from 1 to 2 is -0.022268, so the probability of voting would decrease by 2.23% when the mental health increase from 1 to 2. The first difference for an increase in the mental health index from 5 to 6 is -0.021477, so the probability of voting would decrease by 2.15% when the mental health increase from 5 to 6.



Part 2: Modeling TV consumption

- The probability distribution is Poisson distribution (y_i indicate the TV hours)

$$Pr(Y_i = y_i|\mu) = \frac{\mu^{y_i} e^{-\mu}}{y_i!}$$

- The linear predictor:

$$\eta_i = \beta_0 + \beta_1 age + \beta_2 childs + \beta_3 educ + \beta_4 female + \beta_5 grass + \beta_6 hrsrelax + \beta_7 black$$

- The link function

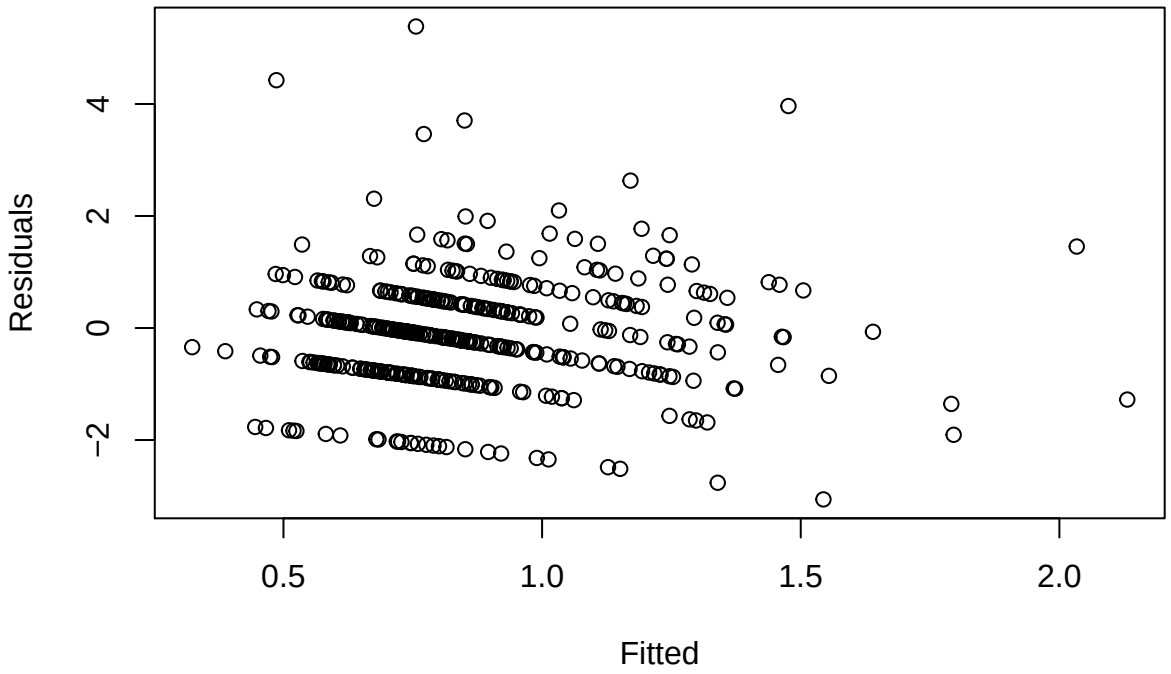
$$\eta_i = \log(\mu_i)$$

In this Poisson regression model, the response variable is the number of TV hours per day. The predictors include age, number of children, education, gender, opinion on legalizing marijuana, hours to relex, and race. The regression results shows that there are three predictors education, hours to relax and race are statistically significant under $\alpha = 0.001$.

The coefficient of education is -0.0393047, which means that one unit increase in education would lead to 0.9614577-fold change in the number of hours of watching TV by average. The coefficient of hours to relax is 0.0471133, which means one unit increase in hours to relax would lead to 1.048241-fold change in the number of hours of watching TV by average. The coefficient of race is 0.4495892, which being black would lead to 1.567668-fold change in the number of hours of watching TV by average.

Table 3: Poisson Regression Results

	Dependent variable:
	tvhours
age	0.0002
	(0.003)
childs	−0.004
	(0.024)
educ	−0.039***
	(0.011)
female	0.018
	(0.064)
grass	−0.109*
	(0.063)
hrsrelax	0.047***
	(0.010)
black	0.450***
	(0.074)
Constant	1.218***
	(0.197)
Observations	441
Log Likelihood	−781.665
Akaike Inf. Crit.	1,579.330
Note:	*p<0.1; **p<0.05; ***p<0.01



Then we also look at the residual vs. fitted plot and we see clear evidence of nonconstant variance, we could estimate a dispersion parameter, which is 1.113, indicating that original model is overdispersion and we could adjust our model summary by using quasi-poisson.

Table 4: Quasipoisson Regression Results

	Dependent variable:
	tvhours
age	0.0002
	(0.003)
childs	−0.004
	(0.025)
educ	−0.039***
	(0.012)
female	0.018
	(0.067)
grass	−0.109
	(0.067)
hrsrelax	0.047***
	(0.010)
black	0.450***
	(0.078)
Constant	1.218***
	(0.208)
Observations	441
Note:	*p<0.1; **p<0.05; ***p<0.01