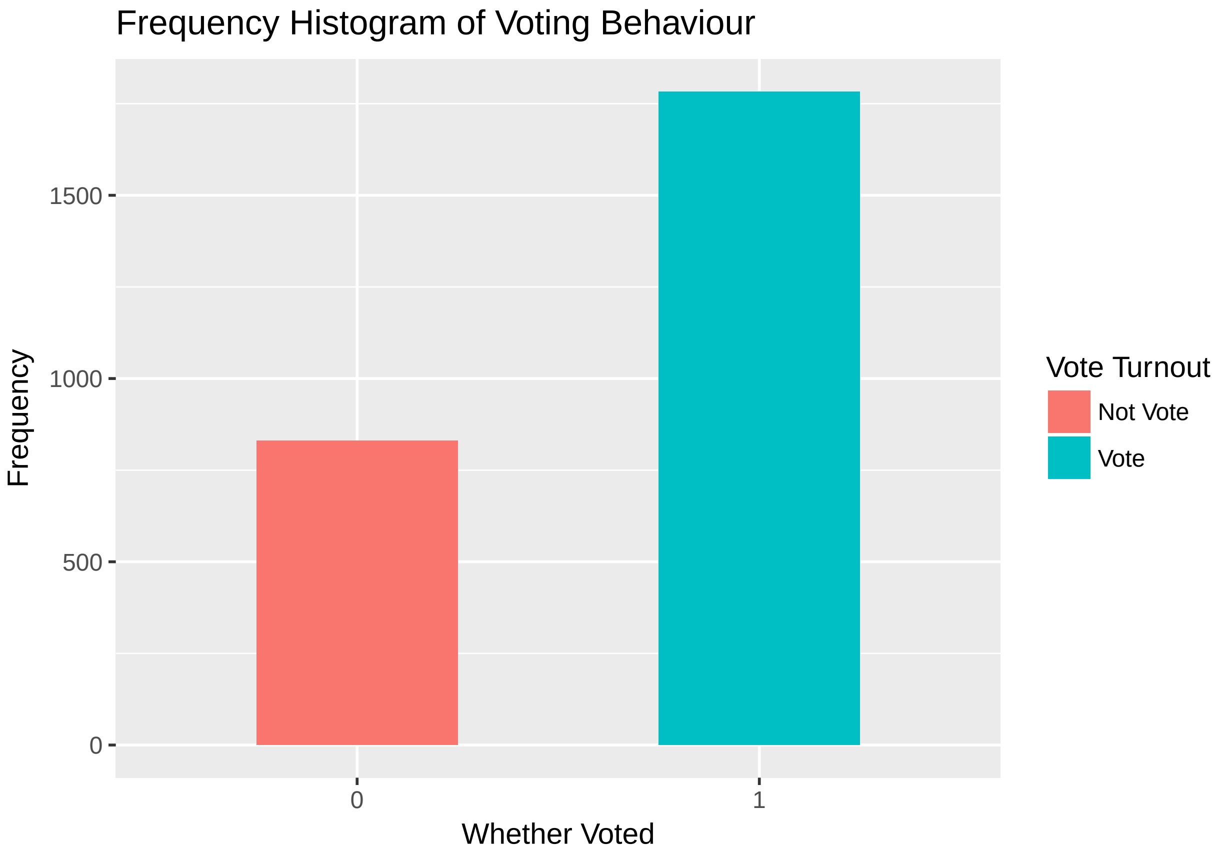
Problem Set 6

MACS 30100 - Perspectives on Computational Modeling Luxi Han 10449918

Problem 1

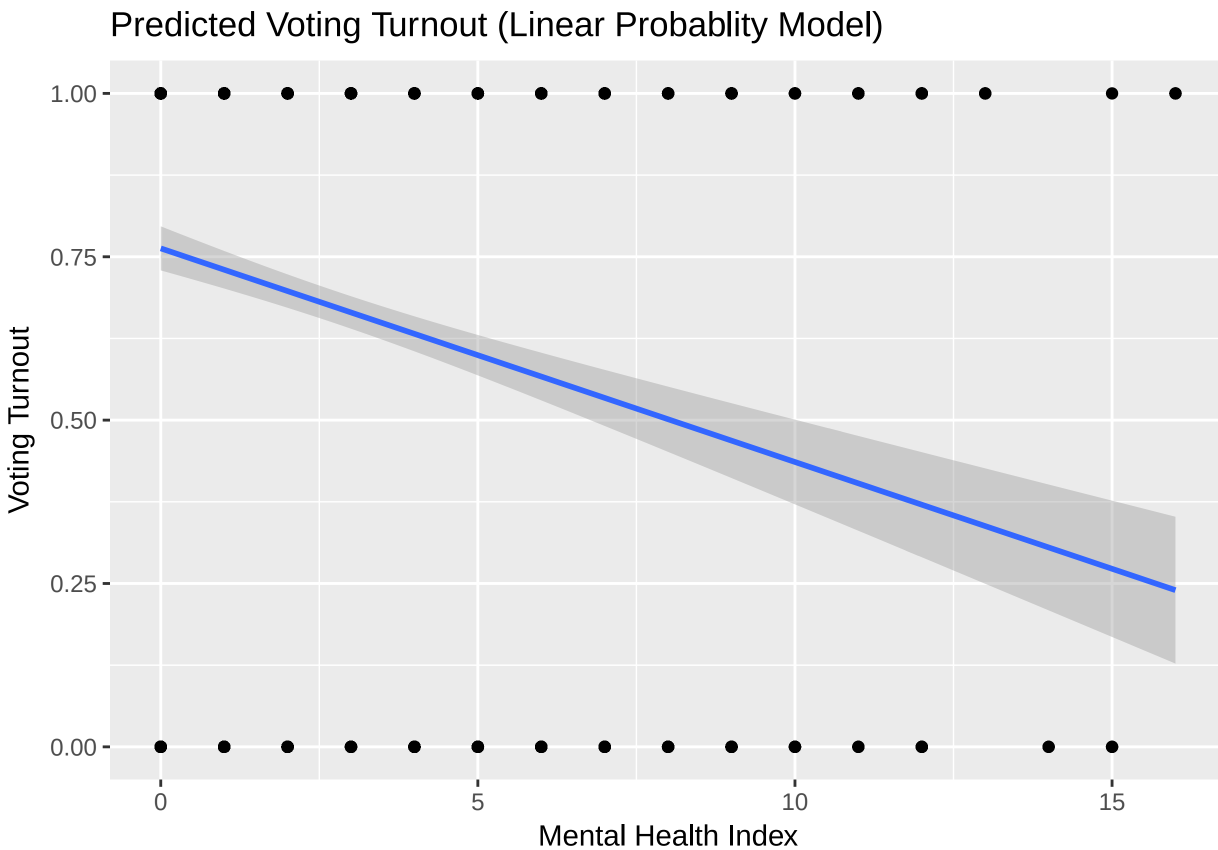
1. The following is the graph for the histogram of the variable




```
## [1] "The unconditional probablity a voter will vote is: 0.682357443551473"
```

The unconditional probablity of voter voting is about 68.24%.

2. The following is the scatter plot and the smoothed regression line:



This graph tells us that the relationship between voter turnout and mental health is negatively correlated. The reason why this graph is problematic is that: 1) voter turnout is a binary choice taking on values of either 0 or 1, while the predicted value is a strand of continous value between 0 and 1; 2) if we were to plot further down along the x axis, then we will get negative predicted voter turnout. This is unsenesible.

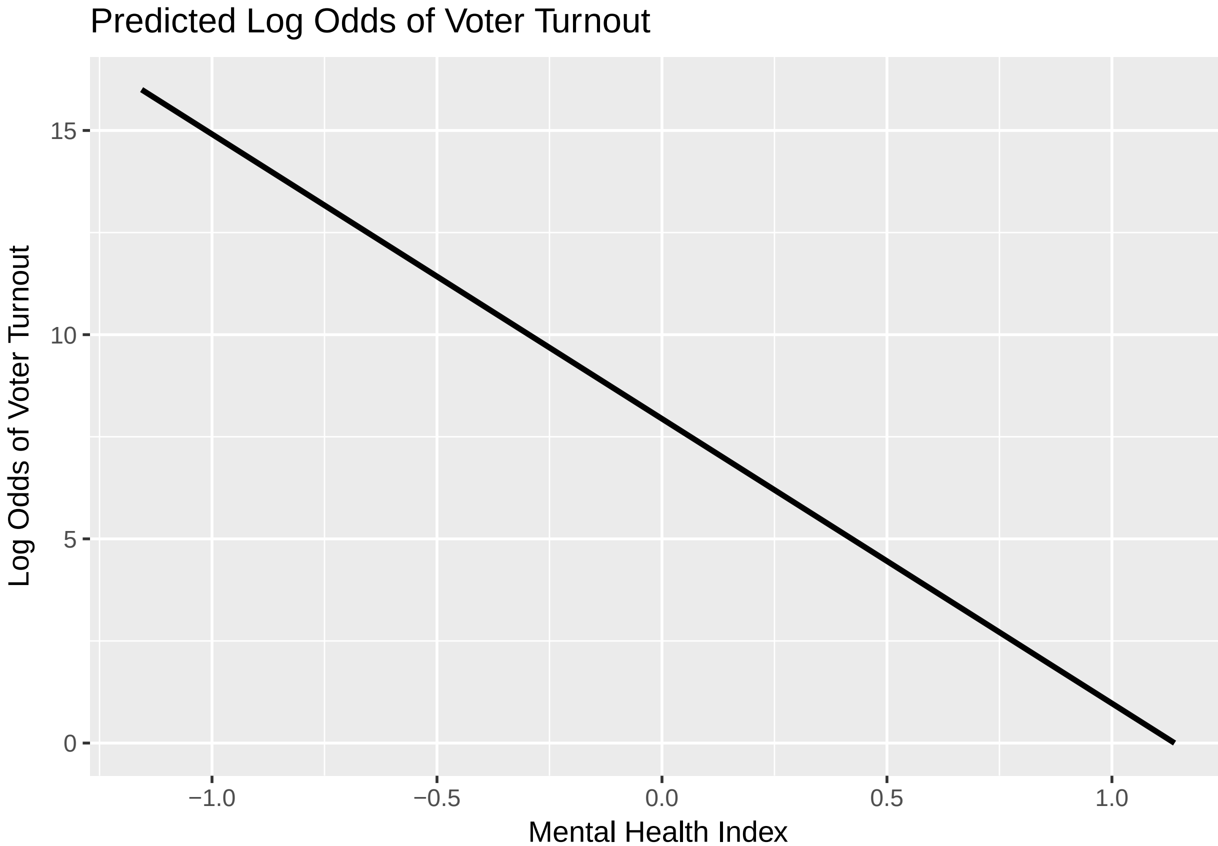
Problem 2

1. There is indeed a significantly negative relationship between voter turnout and meantal health. The esitmated parameter is about -0.14348 which is significant on 0.001 significance level.

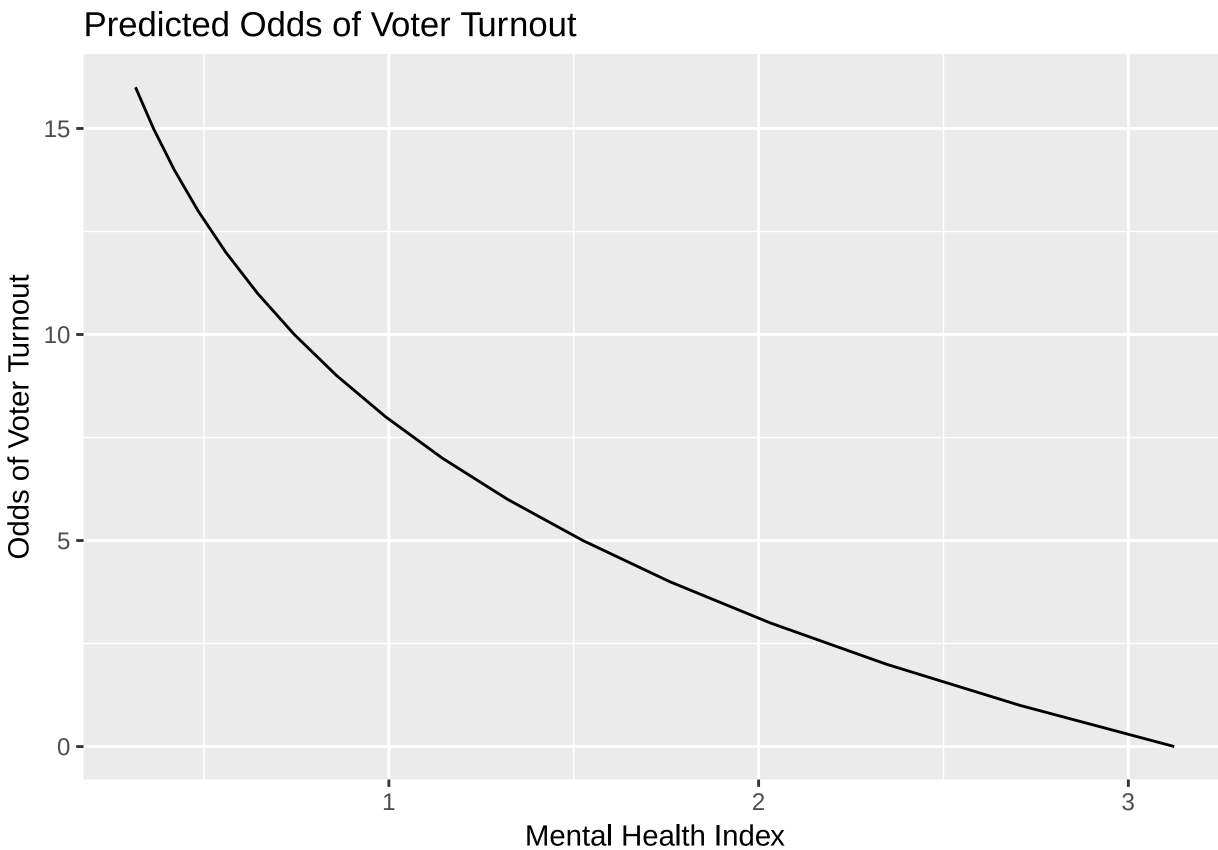
Table 1: Logit Regression With Single Predictor (Voter Turnout)

	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

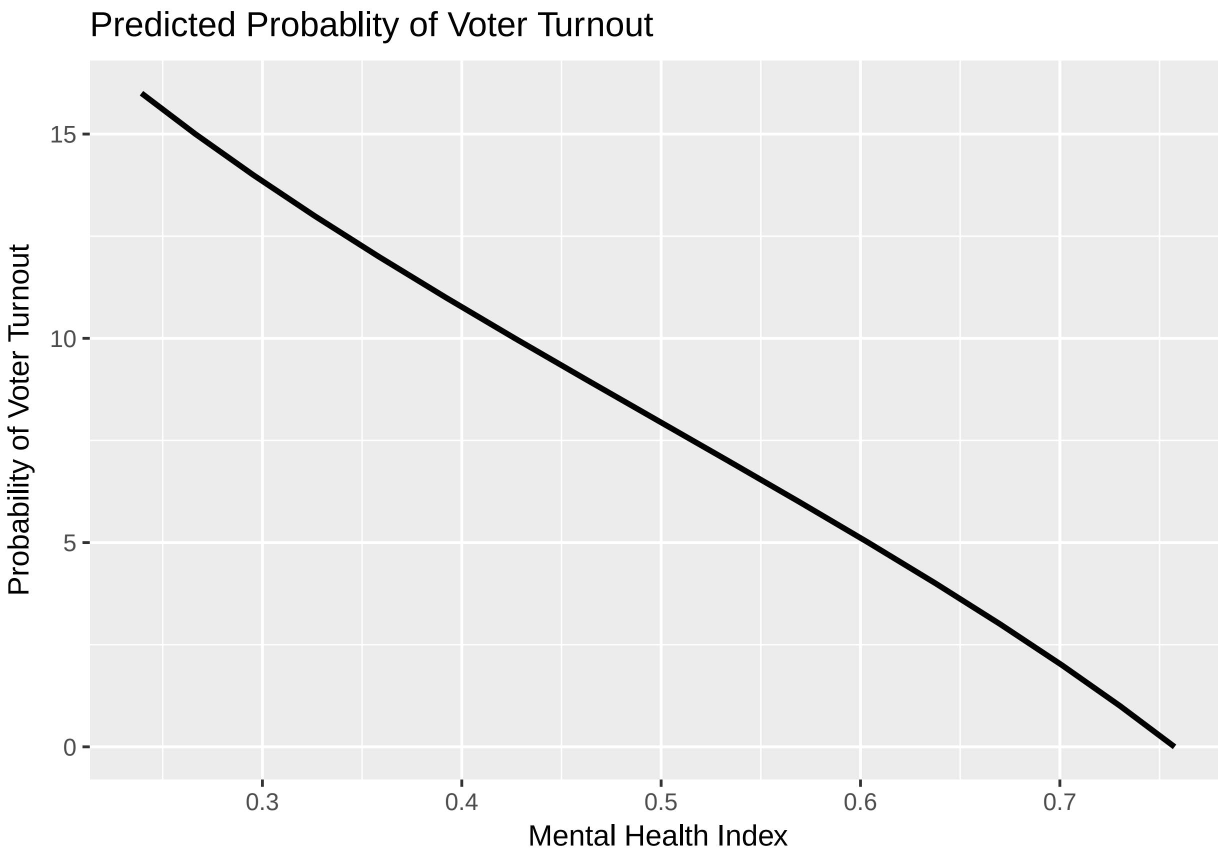
2. When the evaluation on mental health index increases by one unit, the log odds of voter voting against not voting decreases by -0.14348. The following is the graph:



3. The estimator on odds can be interpreted as percent change. When the evaluation on mental health increases by one unit, the odds of voter voting against not voting decreases by -14.348 percent(%).



4. The interpretation of the estimator from the perspective of probablity is not certain. Since the first difference typically depend on the initial age.




```
## [1] "first difference going from 1 to 2 is  -0.0291782383716035"
## [1] "first difference going from 5 to 6 is  -0.0347782137951934"
```

The first difference for an increase in the mental health index from 1 to 2 is -0.0292; from 5 to 6 is -0.0348.

5.

```
## [1] 0.677761
## [1] 0.01616628
## Area under the curve: 0.6243
```

The accuracy rate is 0.6778. The prediction error reduction is 1.62%. The AUC is 0.6243. The model doesn't really explain the binary choice of voting very well. We can see that the prediction error reduction is only around 1.6%, which is a small magnitude with a 0-100% scale.

Problem 3

1. We have the following: random component is bernouli distribution:

$$Pr(Y_i = y_i|\pi_i) = (\pi_i)^{y_i}(1-\pi_i)^{1-y_i}$$

Then we know π_i is the population 'mean' we want to model;

linear predictor is:

$$\begin{aligned}\eta_i = &\beta_0 + \beta_1 mhealth_sum_i + \beta_2 age_i + \beta_3 educ_i + \\ &\beta_4 black_i + \beta_5 black_i + \beta_6 female_i + \beta_7 married_i + \beta_8 inc10_i\end{aligned}$$

the link function is:

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

2. The following is the regression result:

Table 2: Logit Regression With Multiple Variables (Voter Turnout)

	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

3.

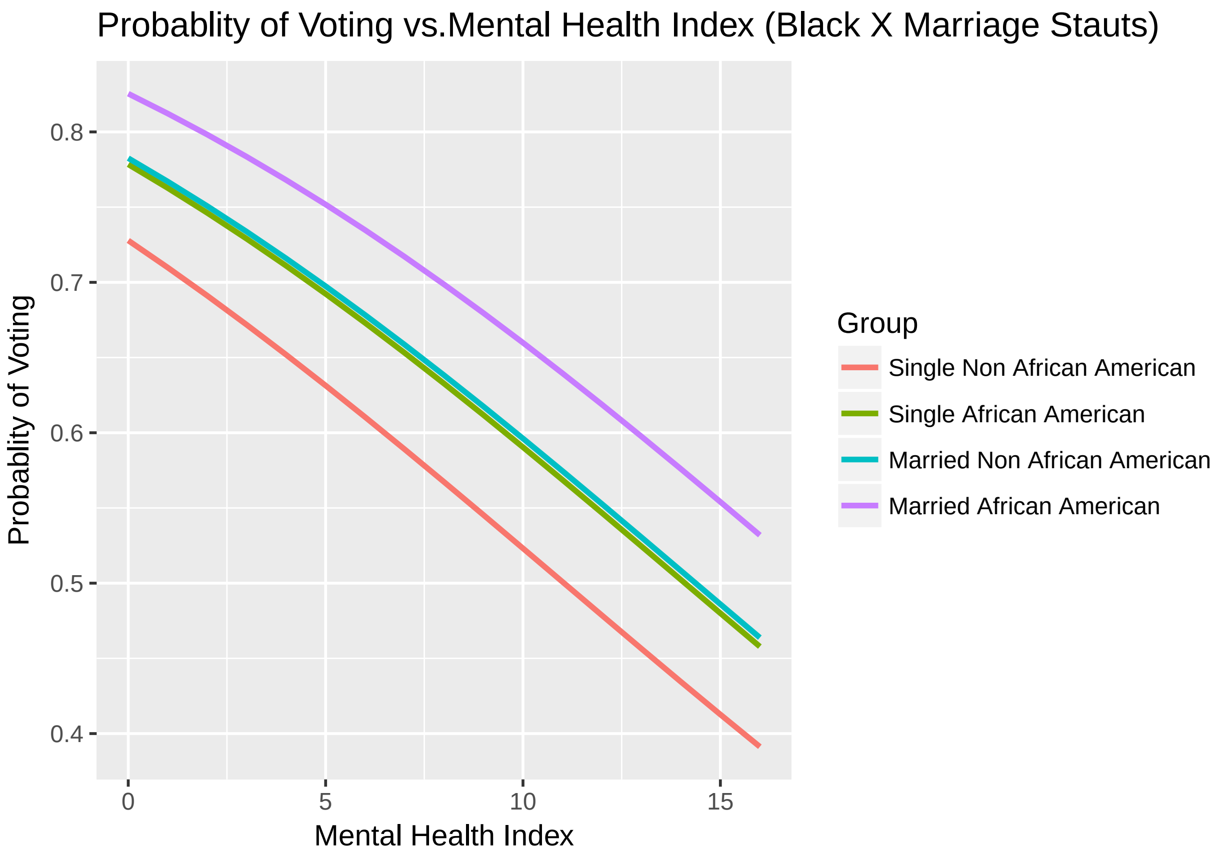
```
## [1] 0.1481481
```

Overall, the preformance or prediciton power of this model improves significantly relative to the model in last question. Using prediction error reduction as a criterion, we get the result that the prediction error reduces by 14.8% compared to the baseline model where we predict one individual will always vote.

Among all of the independent variables, the mental health index, age, education and income turn out to be significant variables. Mental health, age and education are significant on the significance level of 0.001, while income is significant on the level of 0.05. Mental health index has a negative relationship with the voter turnout. On average, one level increase in the mental health index will reduce the odds, defined by the probablity of a voter voting versus not voting, by 1 percent. On the other hand, age, education and income all have positive effect on voter turnout. Specifically, one year increase in age will increase the odds of voting by 4.2%; one year increase in years of educatoin will on average increase the odds of voting by 22.86%; and every ten thousand dollar increase in income will on average increase the odds of voting by 7.0%.

Marriage status is significant on the 0.1 significance level. While whether a person is african american, and gender is statistically insignificant. We plot the predicted probablity of voting against mental health index, which we divide into four groups: married african american people, married non-african american people, unmarried african american people and unmarried non-african american people. We take the mean of all of the continous variables and the median of all the discrete(categorical) varaibles to fix all other predictors fixed.

We can see that, in this case since we don't have any interactive terms, we have both varaibles serving as a shifter of probablity. Married people and African people both have higher probablity of voting. Though we see an incrase that is almost 0.1 in probablity, we still can't jump into conclusion on whether this variable has a large effect of not. On the one hand thses variables are statistically insignificant on 0.05 level, on the other hand a 0.1 incrase in voting probablity is not a negligible effect. Serveral factors can casue this problem. The most probable one is multicolinearity with other variables. For example, marriage status can be closely correlated with age and income.



Problem 4

1. We have the following: random component is poisson distribution:

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

Then we know π_i is the population 'mean' we want to model;

linear predictor is:

$$\begin{aligned}\eta_i =&\beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 education_i + \beta_4 female_i + \beta_5 grass_i +\\ &\beta_6 hrsrelax_i + \beta_7 black_i + \beta_8 social\_connect_i + \beta_9 voted04_i +\\ &\beta_{10} xmovie_i + \beta_{11} zodiac_i\end{aligned}$$

the link function is:

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

This should be the right form. In class, we wrote the opposite which is wrong (in class we said μ equals to log of η_i, this is wrong). Instead the mean function(the inverse of link function) is:

$$\mu_i = g(\eta_i) = e^{\eta_i}$$

2. The following is the regression result:

Table 3: Poisson Regression of Number of Hours Wathcing TV per Day

	Dependent variable:
	tvhours
age	0.001 (0.003)
childs	−0.001 (0.024)
educ	−0.029** (0.012)
female	0.042 (0.065)
grass	−0.098 (0.067)
hrsrelax	0.047*** (0.010)
black	0.462*** (0.076)
social_connect	0.043 (0.040)
voted04	−0.096 (0.077)
xmovie	0.086 (0.076)
zodiacAries	−0.119 (0.149)
zodiacCancer	0.008 (0.143)
zodiacCapricorn	−0.233 (0.164)
zodiacGemini	0.007 (0.145)
zodiacLeo	−0.178 (0.153)
zodiacLibra	−0.057 (0.135)
zodiacNaN	−0.314 (0.211)
zodiacPisces	−0.163 (0.163)
zodiacSagittarius	−0.236 (0.156)
zodiacScorpio	0.033 (0.149)
zodiacTaurus	−0.147 (0.163)
zodiacVirgo	−0.144 (0.155)
Constant	1.112*** (0.236)
Observations	446
Log Likelihood	−783.571
Akaike Inf. Crit.	1,613.143
Note:	*p<0.1; **p<0.05; ***p<0.01

3.

We first perform the goodness of fit test. Essentially, the goodness of fit test is to take the 'difference' of the predicted value using the model specified and the true value of the counts and perform a test on the difference. The difference conforms to a chisqure distribution.

```
## [1] 0.3679743
```

As we can see the p-value is approximately 0.368. Remeber the null hypothesis for the goodness of fit test is: the model fits the data (deviance equals to zero). Since this is a chisqure distribution and indeed if the statistics approaches zero, the deviance is approaching zero indicating small difference between the predicted and real value. Thus we cannot reject this null hypothesis. Judging by the p-value, this model provides a good fit of the data.

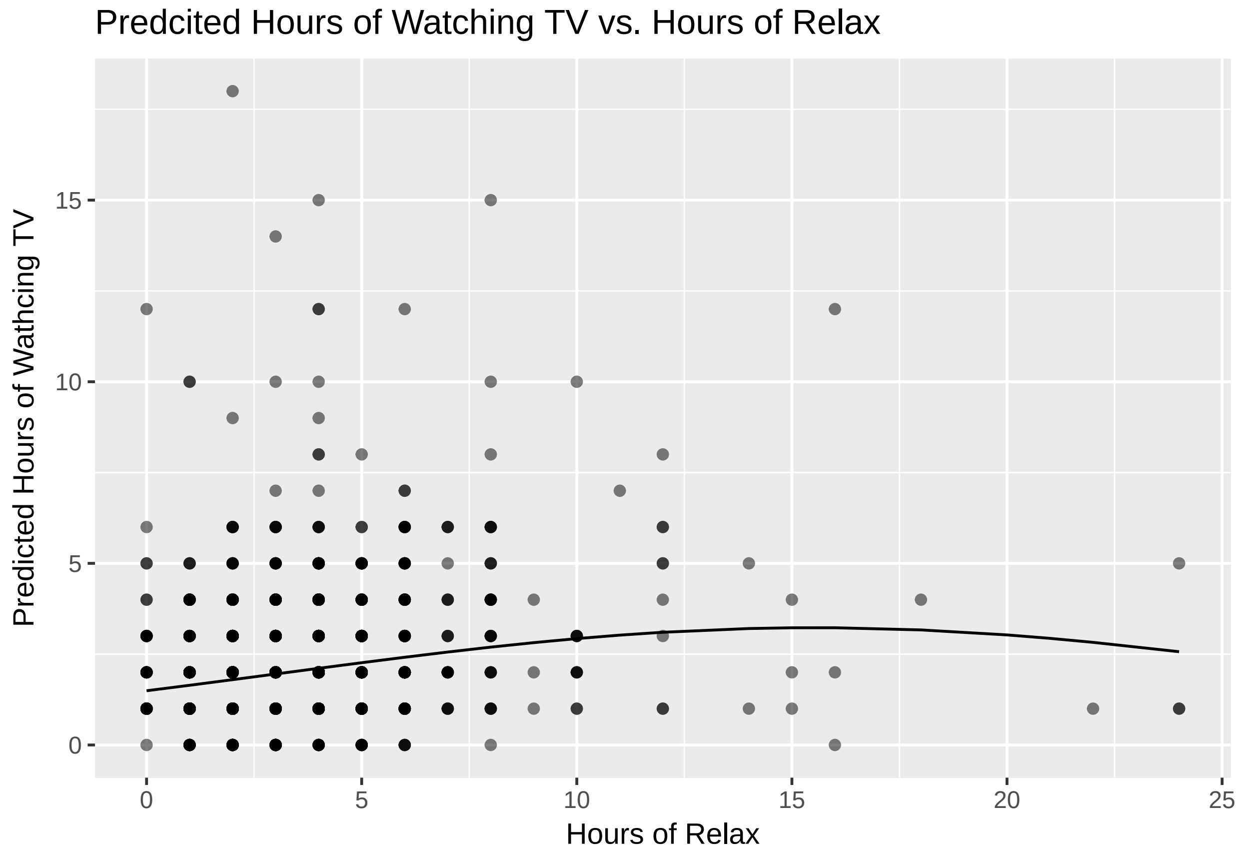
But we can take a close look at the regression table. Years of edcuation, hours of relax per day and whether people are areican american are significant in 0.005 level. Specifically, 1 year incrase in education on average cause 0.033 unit decrease in log of hours of watching TV (or 3.3% percent decrease); one hour incrase in hours of relax per day on average causes 0.046 units incrase in log of hours of watching TV (or 4.6%); holding all other constant, being an African American people on average incrase the log of hours of watching TV by 0.44. This is a farily large incrase, considering $\frac{\partial log(mu_i)}{\partial black}$ represents the precent increase.

The regression result speaks for itself. More educated one person is, less hours they watch TV. This may come as a result of having more work to do for their job or they have other ways of entertainment. Black people tend to watch more TV. This result on its surface doesn't make sense. There could be other socio-economic factors that we don't take into account, for example, income, whether in a food stamp program, or whether on a social welfare program, etc. Another interesting varaible is hours of relax. As hours of relax increases, number of hours of watching television also increases. But as hours of relaxation increases, watching TV is not the only way of entertainment. My hypothesis is that after hours of relax increase to a certain amount, people invest more time to other types of entertainment. There should be a non-linear relationship between hours of wathcing TV and hours of relax. We run the following regression adding the square of hours of relax:

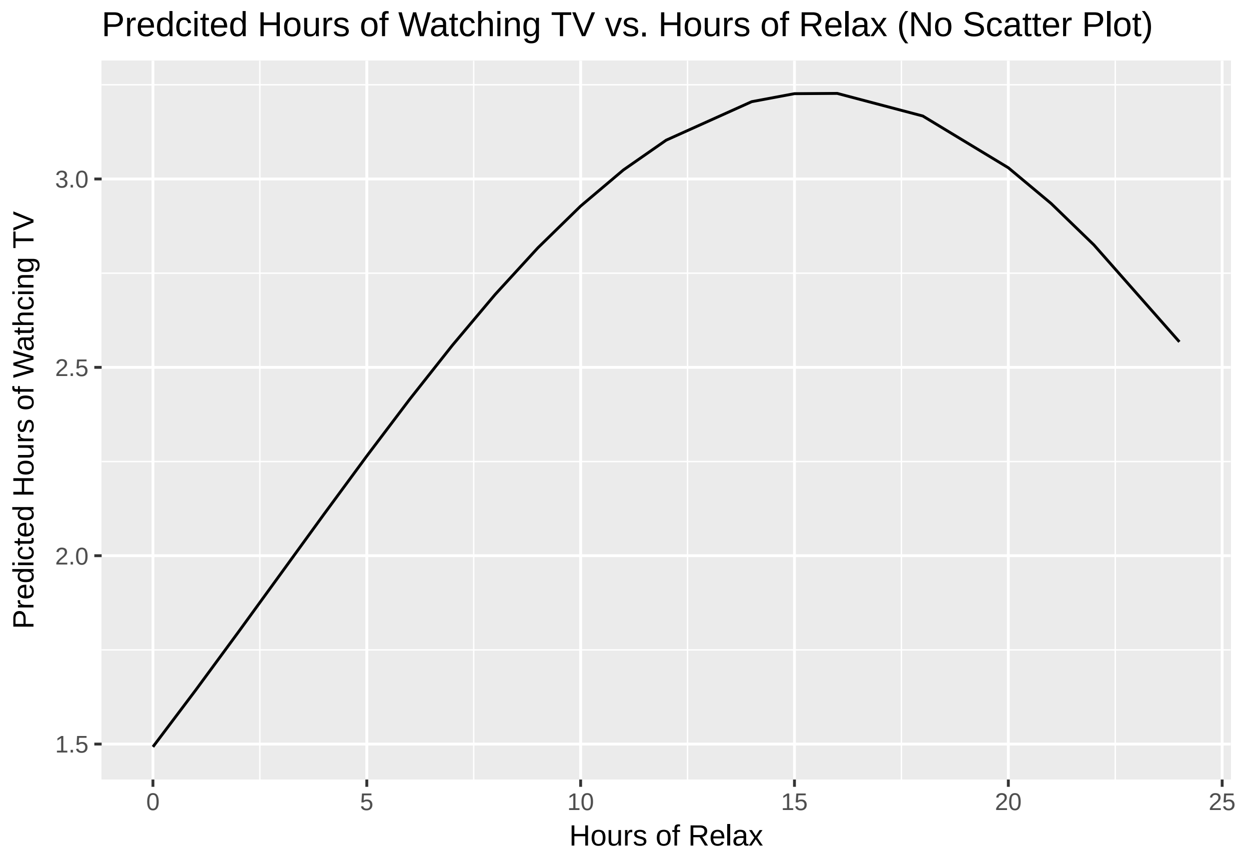
Table 4: Poisson Regression of Number of Hours Wathcing TV per Day

	Dependent variable:
	tvhours
age	0.001 (0.003)
childs	0.007 (0.024)
educ	-0.027^{**} (0.012)
female	0.043 (0.064)
grass	-0.102 (0.067)
hrsrelax	0.099^{***} (0.026)
I(hrsrelaxˆ2)	-0.003^{**} (0.001)
black	0.478^{***} (0.076)
social_connect	0.049 (0.040)
voted04	-0.108 (0.077)
xmovie	0.073 (0.076)
zodiacAries	-0.118 (0.149)
zodiacCancer	0.018 (0.143)
zodiacCapricorn	-0.208 (0.164)
zodiacGemini	0.019 (0.145)
zodiacLeo	-0.174 (0.153)
zodiacLibra	-0.046 (0.135)
zodiacNaN	-0.304 (0.211)
zodiacPisces	-0.165 (0.163)
zodiacSagittarius	-0.176 (0.157)
zodiacScorpio	0.050 (0.149)
zodiacTaurus	-0.138 (0.163)
zodiacVirgo	-0.128 (0.155)
Constant	0.942^{***} (0.248)
Observations	446
Log Likelihood	-780.926
Akaike Inf. Crit.	1,609.852
Note:	*p<0.1; **p<0.05; ***p<0.01

Now we can see that the square of hours of relax is negatively significant. This suggests there is a firstly incrasing and then decreasing relationship between hours of relax and hours of watching TV. We then plot the predicted counts against hours of relax. In this case, we take all other predictors as their median value (except for zodiac, I took Aries as the predictor value)



This indeed is a hump shaped curve. Furthermore, if we get rid of the scatter plot, we have:



As for other predictors, they are not significant in 0.1 level. But notice, zodiac is not related to hours of watching TV. This may be a supporting evidence that zodiac is just a relfect of random month to be born in.