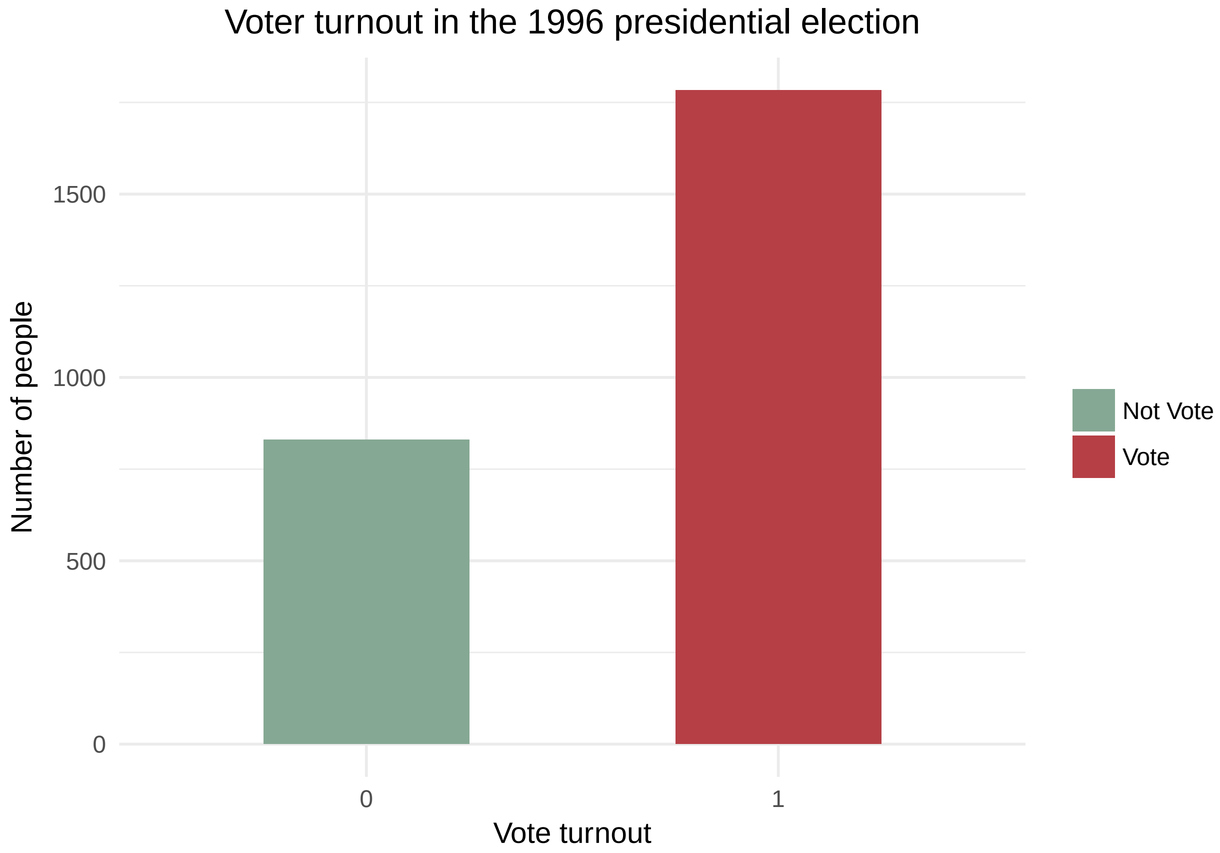
Problem set #6: Generalized linear models

Yiqing Zhu

Part 1: Modeling voter turnout

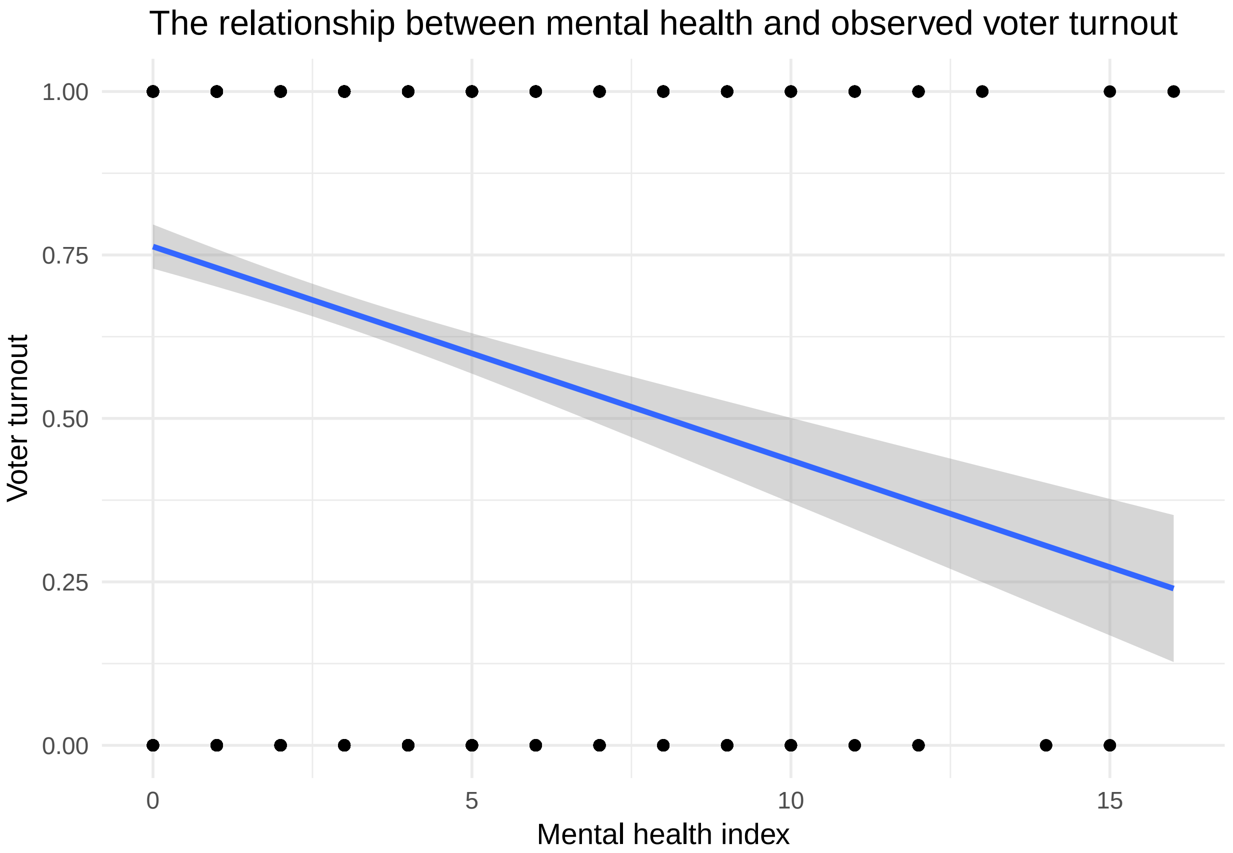
Problem 1: Describe the data

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?



The unconditional probability of a given individual turning out to vote is 68.2357444%.

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?



The plot tells us that the more depressed an individual is, the less desire he or she has to participate in politics.

The problematic about this linear smoothing line is first, the only possible values for `vote96` are 0 and 1, yet the linear regression model gives us predicted values such as .4 and .25. Second, even if these values are predicted probabilities, that is, the estimated probability an individual will vote given their mental health index, the line is linear and continuous, so it extends infinitely in both directions of mental health index, but we cannot have a probability outside of the [0, 1] interval.

Problem 2: Basic model

The summary of the logistic regression model of the relationship between voter turnout and mental health estimated is shown below:

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = binomial, data = mhealth_nona)
##
## 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
##   (1291 observations deleted due to missingness)
## AIC: 1620.7
##
## Number of Fisher Scoring iterations: 4
```

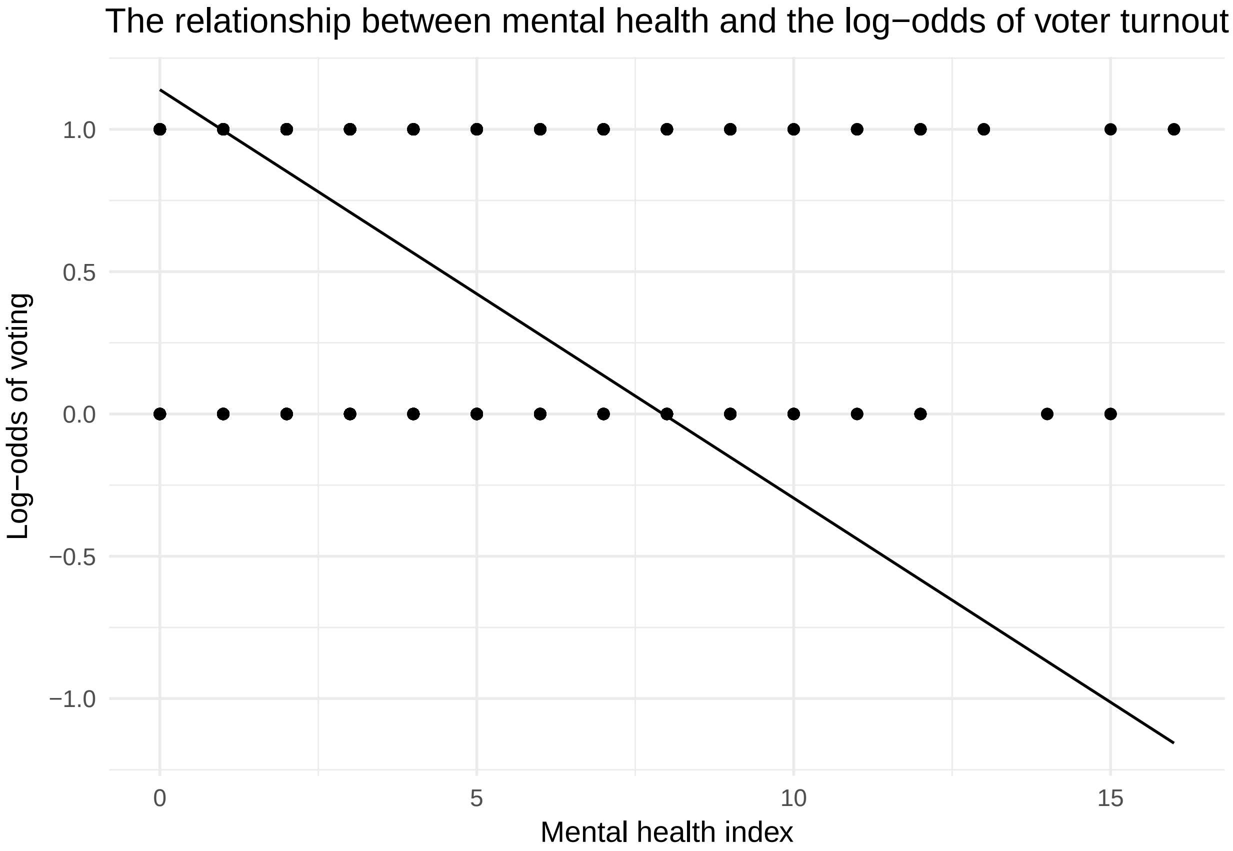
```
##          term   estimate  std.error statistic      p.value
## 1 (Intercept)  1.1392097 0.08444019 13.491321 1.759191e-41
## 2 mhealth_sum -0.1434752 0.01968511 -7.288516 3.133883e-13
```

1. Is the relationship between mental health and voter turnout statistically and/or substantively significant?

The relationship between mental health and voter turnout is statistically significant since the p-value of `mhealth_sum` coefficient in the estimated model above is 3.133883e-13, which shows more than 99% chance of rejecting the null hypothesis.

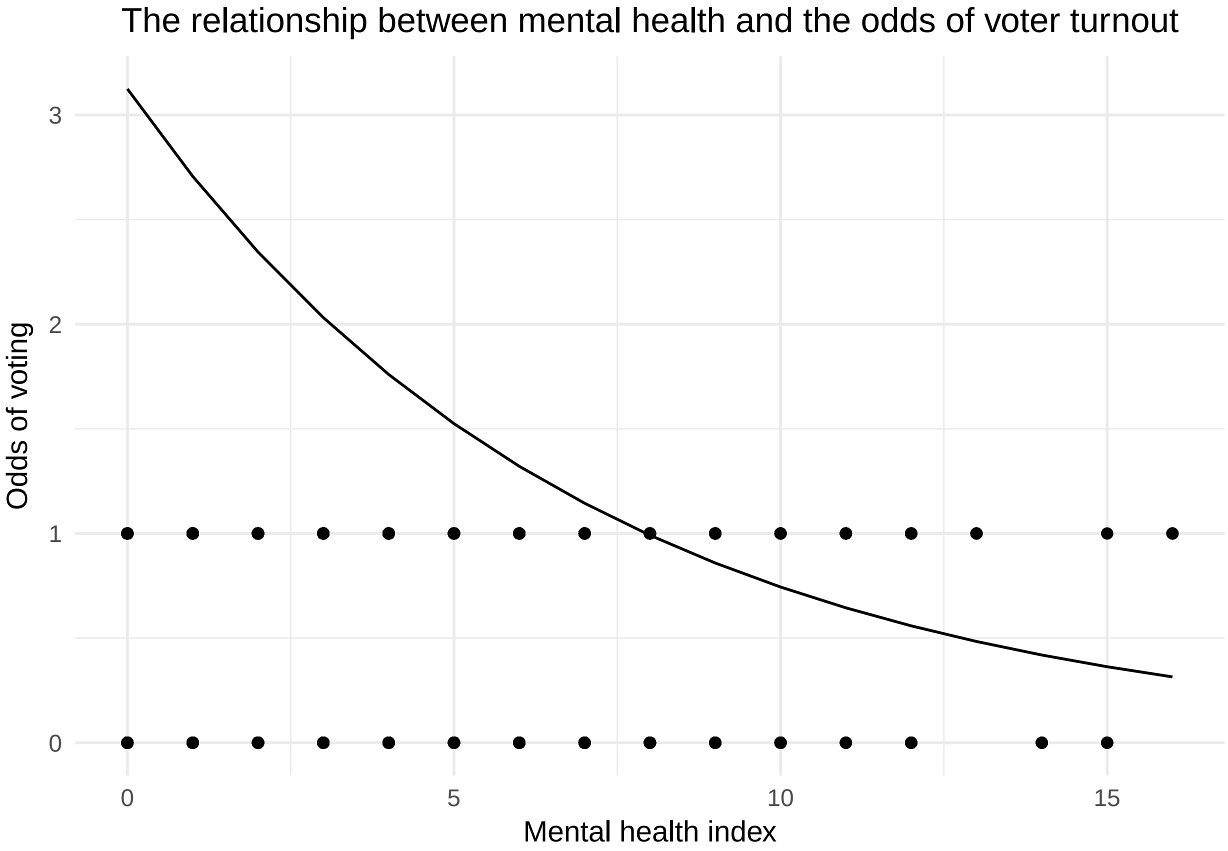
The relationship is substantively significant because the `mhealth_sum` coefficient is -0.1434752, which means that an increase in the mental health index from 1 to 7, there will be 0.8608512 decrease in the log-odds of voting, aka, 1.5623084 decrease in the odds of voting, 0.1965493 decrease in the probability of voting, so the effect of mental health on voting is significant.

2. Interpret the estimated parameter for mental health in terms of log-odds. Generate a graph of the relationship between mental health and the log-odds of voter turnout.



The estimated parameter for mental health in terms of log-odds is -0.1434752, which means that for every one-unit increase in mental health index, we expect the log-odds of voting to decrease by 0.1434752.

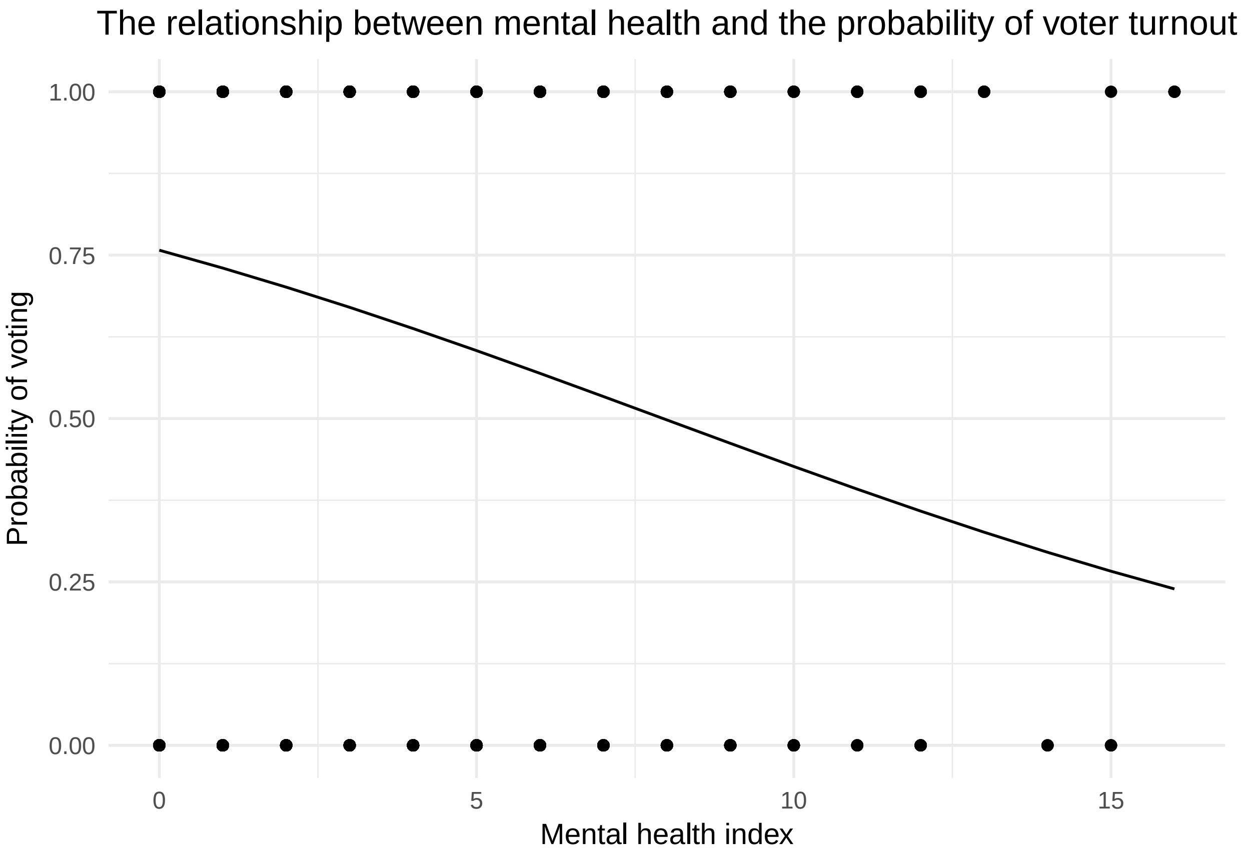
3. Interpret the estimated parameter for mental health in terms of odds. Generate a graph of the relationship between mental health and the odds of voter turnout.



The estimated parameter for mental health in terms of odds is 0.8663423, and it can only be evaluated at a certain mental health index since the relationship between mental health index and odds of voting is not linear.

For example, a respondent with mental health index of 1 would have odds of voting 2.7067116, which means the respondent is 2.7067116 times more likely to vote than not vote.

4. Interpret the estimated parameter for mental health in terms of probabilities. Generate a graph of the relationship between mental health and the probability of voter turnout. What is the first difference for an increase in the mental health index from 1 to 2? What about for 5 to 6?



The estimated parameter for mental health in terms of probabilities is 0.4641926, and just like odds, it can only be evaluated at a certain mental health index since the relationship between mental health index and probabilites of voting is not linear.

For example, a respondent with mental health index of 1 would have probability of voting 0.7302191.

The probability difference for an increase in the mental health index from 1 to 2 is -0.0291782, and for an increase in the mental health index from 5 to 6 is -0.0347782.

5. Estimate the accuracy rate, proportional reduction in error (PRE), and the AUC for this model. Do you consider it to be a good model?

For this model, the accuracy rate is 0.677761, the proportional reduction in error(PRE) is 0.0161663, and the area under the ROC curve (AUC) is 0.6243087.

I don't think this is a good model. Though the accuracy rate seems acceptable, the PRE shows that this model only reduces about 1.62% preduction error, and the AUC indicates that its performance is a liitle better than a random guess, which will have the AUC of 0.5.

Problem 3: Multiple variable model

1. Write out the three components of the GLM for your specific model of interest. This includes the Probability distribution (random component), Linear predictor, Link function.

My specific model of interest includes: the Probability distribution (random component) is the Bernoulli distribution:

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

the Linear predictor:

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

the link function is the logit function:

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

2. Estimate the model and report your results.

The summary of the model of the relationship between voting status and mental health, age, education, race(black or not), marrital status(married or not), family income(in $10,000s) estimated is shown below:

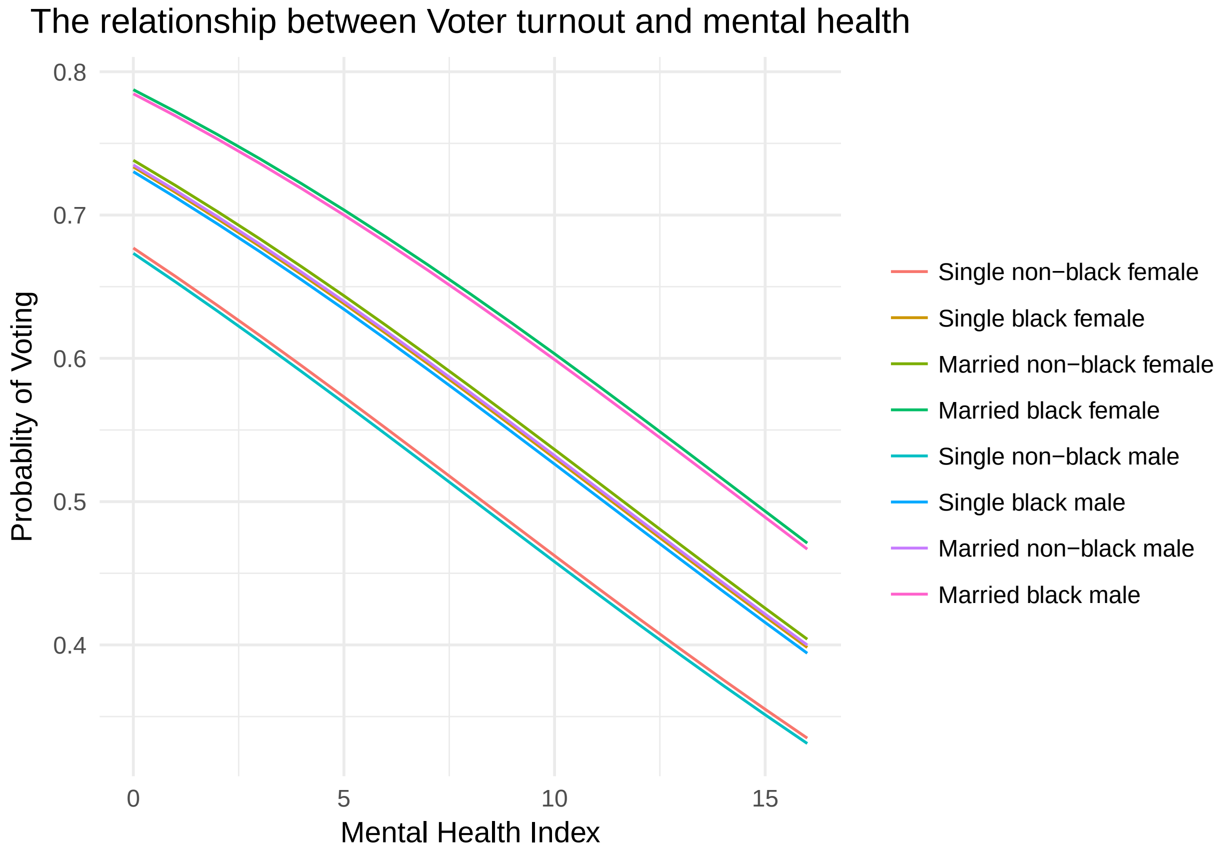
```
##
## Call:
## glm(formula = vote96 ~ ., family = binomial, data = mhealth_nona)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4843  -1.0258   0.5182   0.8428   2.0758
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.304103   0.508103  -8.471  < 2e-16 ***
## mhealth_sum -0.089102   0.023642  -3.769 0.000164 ***
## age          0.042534   0.004814   8.835  < 2e-16 ***
## educ         0.228686   0.029532   7.744 9.65e-15 ***
## black        0.272984   0.202585   1.347 0.177820
## female      -0.016969   0.139972  -0.121 0.903507
## married      0.296915   0.153164   1.939 0.052557 .
## inc10        0.069614   0.026532   2.624 0.008697 **
## ---
## 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: 1241.8  on 1157  degrees of freedom
##   (1448 observations deleted due to missingness)
## AIC: 1257.8
##
## Number of Fisher Scoring iterations: 4
```

```
##          term    estimate   std.error  statistic      p.value
## 1 (Intercept) -4.30410314 0.508103096 -8.4709248 2.434523e-17
## 2 mhealth_sum -0.08910191 0.023642197 -3.7687660 1.640566e-04
## 3         age  0.04253412 0.004814133  8.8352601 9.986562e-19
## 4        educ  0.22868627 0.029531656  7.7437673 9.651356e-15
## 5       black  0.27298352 0.202585333  1.3474989 1.778196e-01
## 6      female -0.01696914 0.139971531 -0.1212328 9.035067e-01
## 7     married  0.29691482 0.153163585  1.9385471 5.255651e-02
## 8       inc10  0.06961381 0.026532274  2.6237407 8.696996e-03
```

3. Interpret the results in paragraph format. This should include a discussion of your results as if you were reviewing them with fellow computational social scientists. Discuss the results using any or all of log-odds, odds, predicted probabilities, and first differences - choose what

makes sense to you and provides the most value to the reader. Use graphs and tables as necessary to support your conclusions.

Given the α level of 0.05, the p-values above indicates that the relationship between voter turnout and mental health, age, education, and family income are statistically siginificant, while the relationship between voter turnout and race, gender, and marrital status are statistically insignificant. We can verify this by the below plot.



When we hold the age, education, and family income constant, race, gender and marrital status don't change the probability difference of voting, rather, they only change the range of probability of voing. This can be caused by the correlation between race, gender, marital status and age, education, family income, for example, marital status might be strongly correlated with age and family income. Therefore, race, gender, and marrital status are not necessarily included in the model. I adjusted my original model and estimated a new model only with variables mental health, age, education, and family income. And here are the summary of the new model:

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum + age + educ + inc10, family = binomial,
##     data = mhealth_nona)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5650  -1.0385   0.5264   0.8428   2.1159
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.053955   0.490844  -8.259  < 2e-16 ***
## mhealth_sum -0.091129   0.023555  -3.869 0.000109 ***
## age          0.042576   0.004793   8.883  < 2e-16 ***
## educ         0.217151   0.028987   7.491 6.82e-14 ***
## inc10        0.085883   0.024440   3.514 0.000441 ***
## ---
## 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: 1247.3  on 1160  degrees of freedom
##   (1448 observations deleted due to missingness)
## AIC: 1257.3
##
## Number of Fisher Scoring iterations: 4
```

```
##          term    estimate   std.error statistic      p.value
## 1 (Intercept) -4.05395538 0.490844387 -8.259146 1.467186e-16
## 2 mhealth_sum -0.09112867 0.023554935 -3.868772 1.093850e-04
## 3         age  0.04257589 0.004792922  8.883077 6.503616e-19
## 4        educ  0.21715119 0.028986978  7.491336 6.817609e-14
## 5       inc10  0.08588256 0.024440354  3.513966 4.414699e-04
```

Now every variable is statistically siginificant. We can evaluate the accuracy of the model by counting the accuracy rate: 0.7167382, the proportional reduction in error(PRE): 0.1269841, the area under the ROC curve (AUC): 0.7562255. Compared to the single-variable model, this model seems work better. The accuracy rate is higher, the PRE shows that this model reduces about 12.7% preduction error, and the AUC indicates that its performance is better than a random guess.

We can interpret the results as: when holding other variables constant, for every one-unit increase in mental health index, we expect the log-odds of voting to decrease by 0.09112867; when holding other variables constant, for every one-unit increase in age, we expect the log-odds of voting to increase by 0.04257589; when holding other variables constant, for every one-unit increase in education, we expect the log-odds of voting to increase by 0.21715119; when holding other variables constant, for every one-unit increase in family income, we expect the log-odds of voting to increase by 0.08588256.

Part 2: Modeling tv consumption

1. Write out the three components of the GLM for your specific model of interest. This includes the Probability distribution (random component), Linear predictor, Link function

My specific model of interest includes:

the Probability distribution (random component) is the Poisson distribution:

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

the Linear predictor:

$$\eta_i = \beta_0 + \beta_1 hrsrelax_i + \beta_2 socialconnect_i$$

the link function is the log function:

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

2. Estimate the model and report your results.

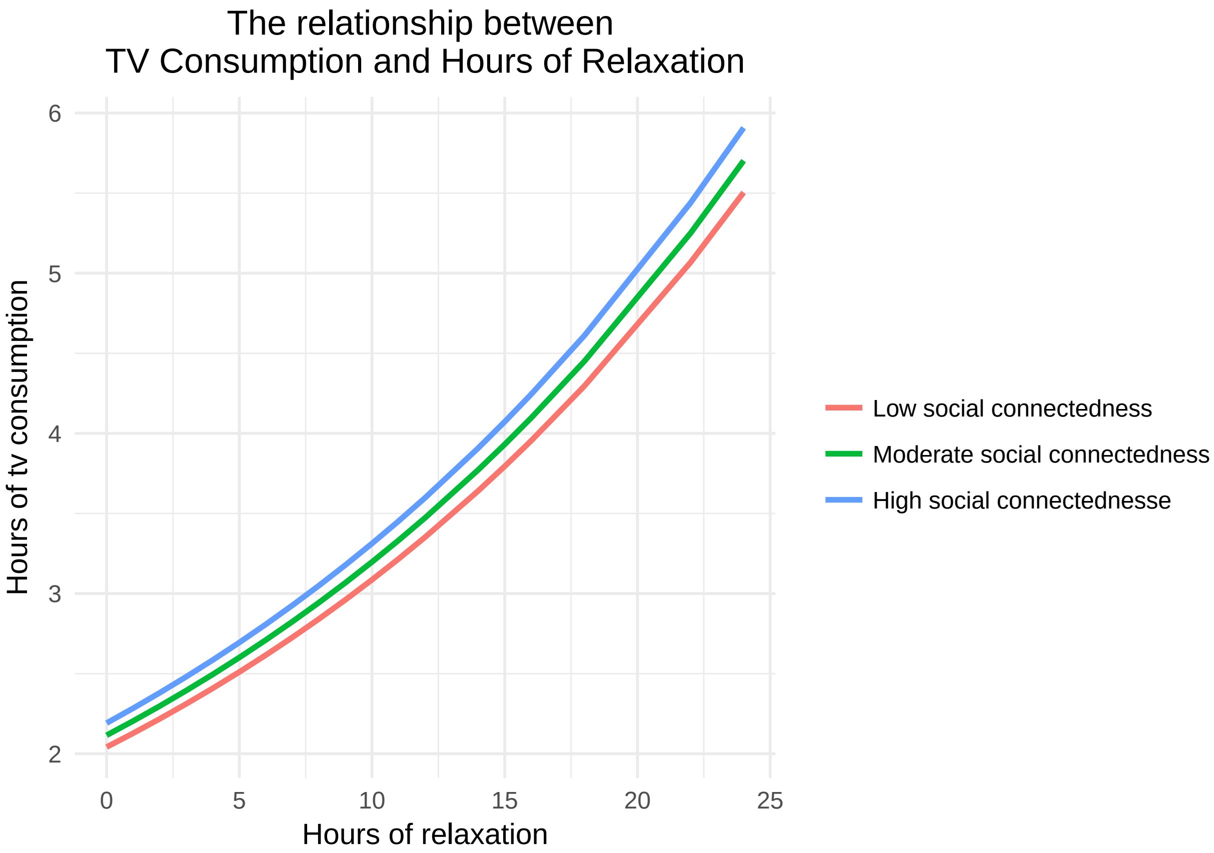
The summary of the model of the relationship between tv consumption and hours and relaxtion , social connectedness estimated is shown below:

```
##
## Call:
## glm(formula = tvhours ~ hrsrelax + social_connect, family = poisson,
##     data = gss_nona)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8626  -0.9178  -0.2095   0.3977   6.5348
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.713770   0.036076  19.785  < 2e-16 ***
## hrsrelax       0.041321   0.006123   6.748  1.5e-11 ***
## social_connect 0.035386   0.023896   1.481    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1346.9  on 1119  degrees of freedom
## Residual deviance: 1301.8  on 1117  degrees of freedom
##   (867 observations deleted due to missingness)
## AIC: 4177.1
##
## Number of Fisher Scoring iterations: 5
```

```
##             term   estimate   std.error statistic      p.value
## 1    (Intercept) 0.71377025 0.036076328 19.785003 4.008764e-87
## 2       hrsrelax 0.04132135 0.006123309  6.748206 1.496847e-11
## 3 social_connect 0.03538642 0.023895942  1.480855 1.386452e-01
```

3. Interpret the results in paragraph format. This should include a discussion of your results as if you were reviewing them with fellow computational social scientists. Discuss the results using any or all of log-counts, predicted event counts, and first differences - choose what makes sense to you and provides the most value to the reader. Is the model over or under-dispersed? Use graphs and tables as necessary to support your conclusions.

We find that the variable hours of relaxation is statistically significant since its p-value is 1.496847e-11, while the variable social connectedness is not statistically significant since its p-value is 1.386452e-01. The variable hours of relaxation has a coefficient of 0.04132135, which means when holding other variables constant, for every one-unit increase in relaxation hours, we expect the log-count of TV watching hours to increase by 0.04132135, aka, TV watching hours to increase by 1.042187. We can visually observe this relationship by the below plot.


The relationship between
TV Consumption and Hours of Relaxation
Hours of tv consumption
2
3
4
5
6
0
5
10
15
20
25
Hours of relaxation
Low social connectedness
Moderate social connectedness
High social connectednesse