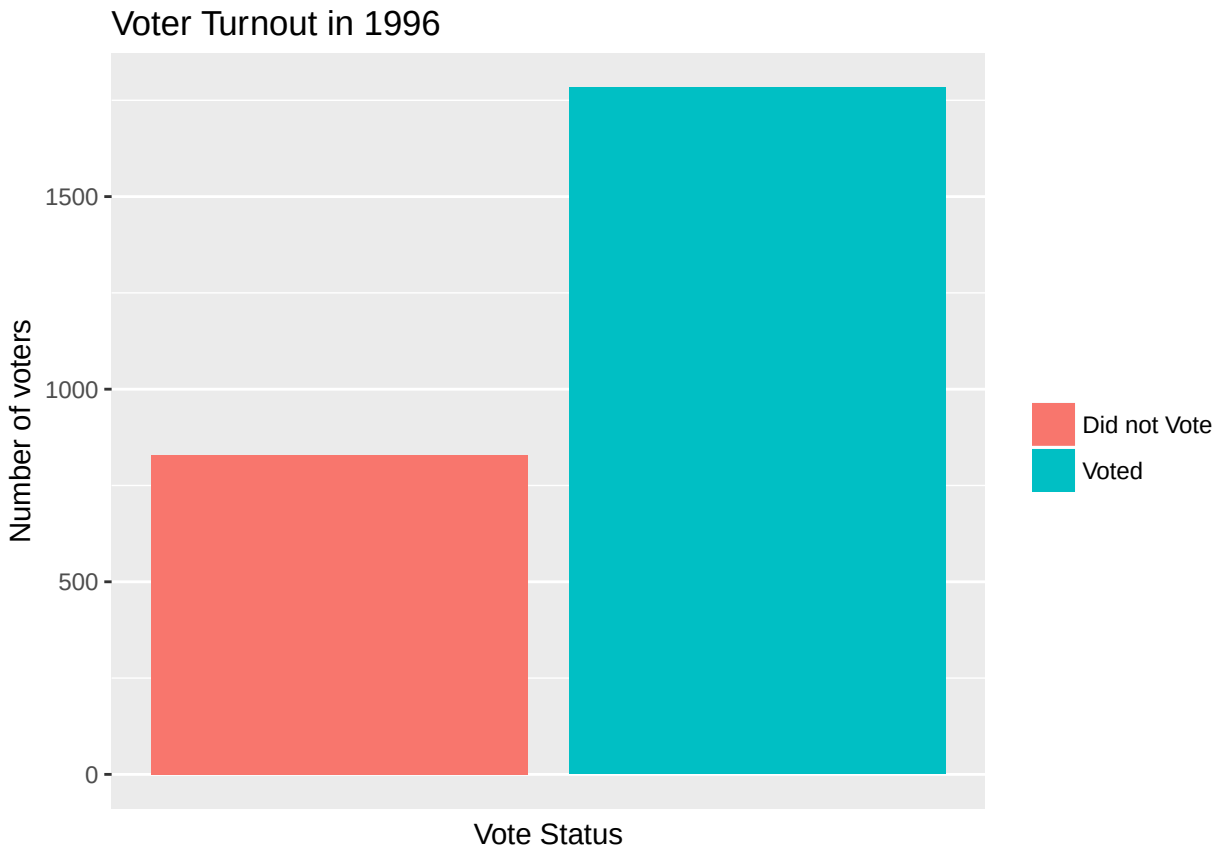
Ps6

Weijia Li

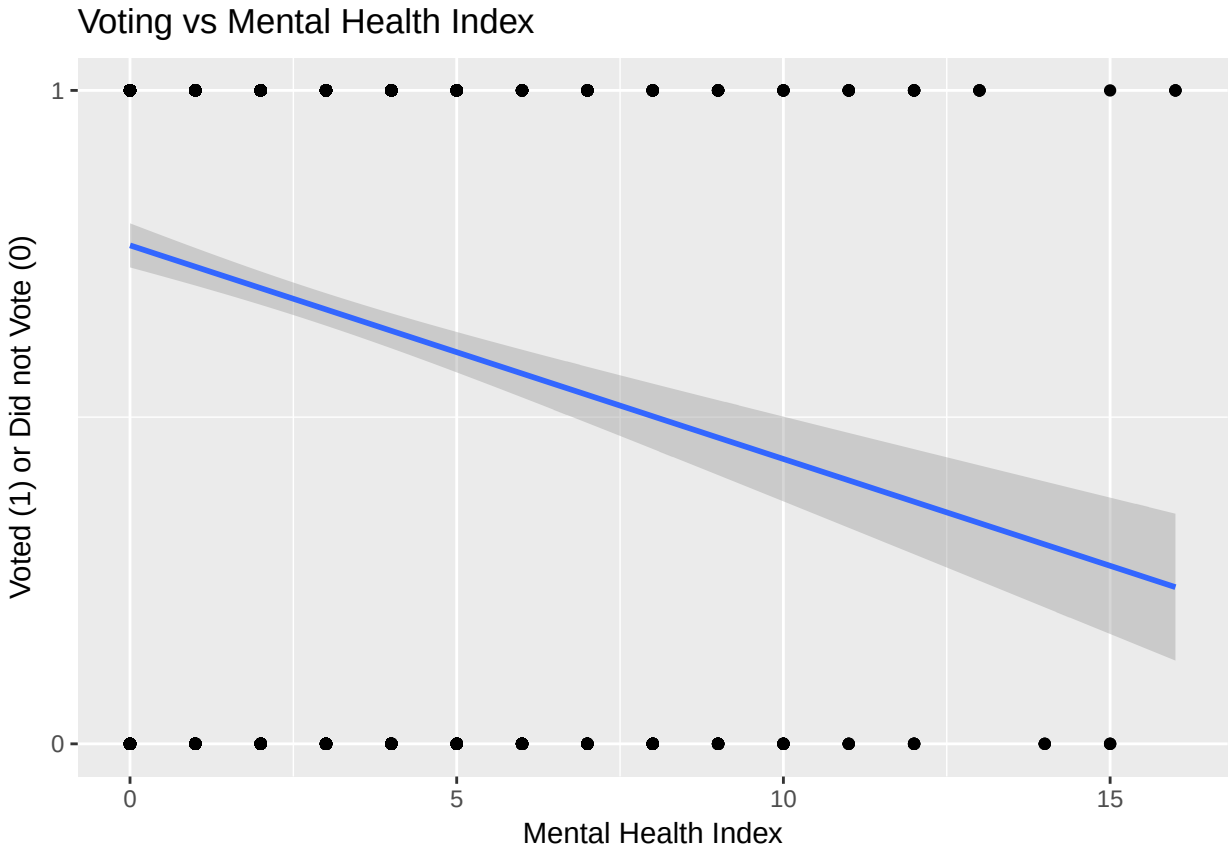
Part 1: Modelling Voter Turnout

Describe the Data




```
## [1] 0.63
```

The unconditional probability is 63%.



The plot shows that the worse the mental health(higher index), the less likely the person will vote.

The plot is problematic in the sense that, the vertical coverage of the linear line is a range in between 0 and 1 while the actual response variables are 0 or 1. By only looking at the line plot, the y value at each end does not explain the voting pattern.

Basic Model

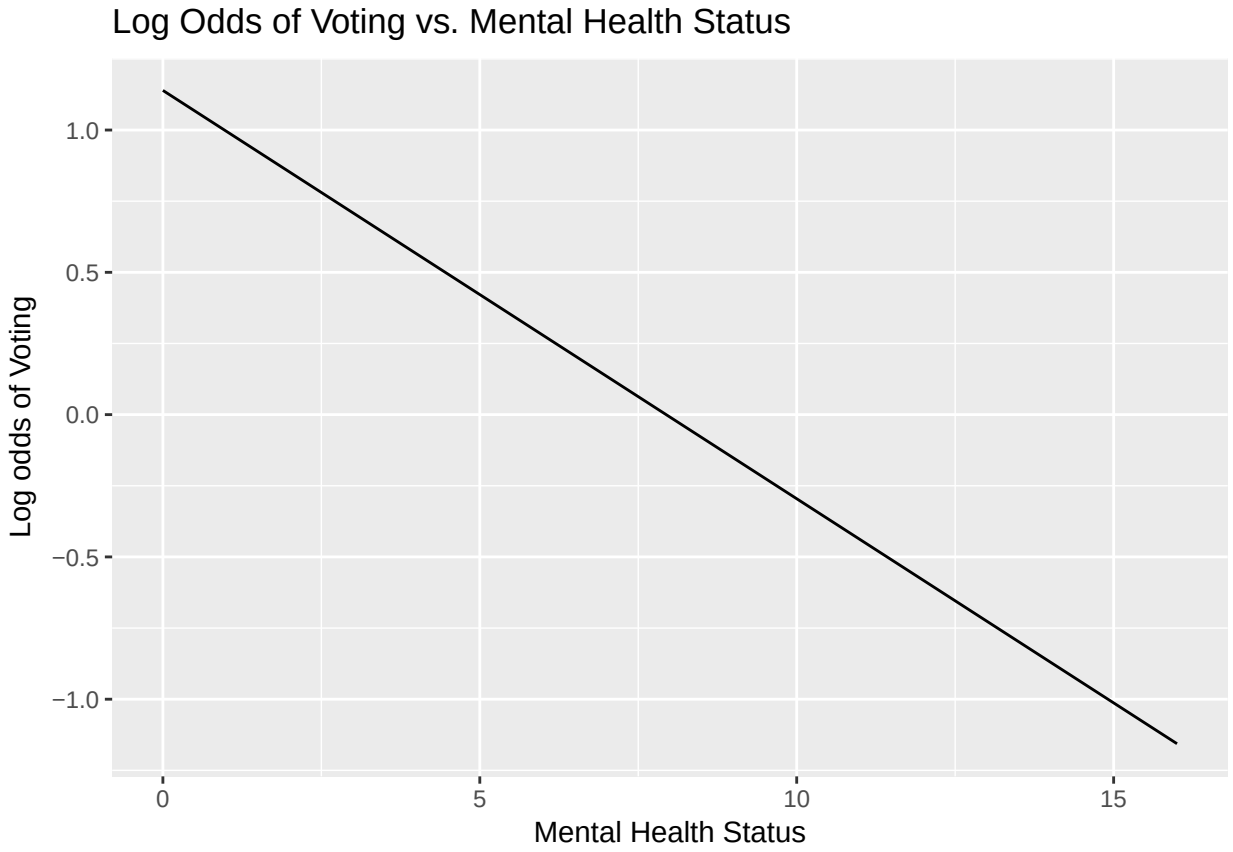
```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  1.1392097 0.08444019 13.491321 1.759191e-41
## mhealth_sum -0.1434752 0.01968511 -7.288516 3.133883e-13
```

1.

The relationship between mental health and voter turnout is stastistically significant since p-value is about 3.133883e-13, which is very small.

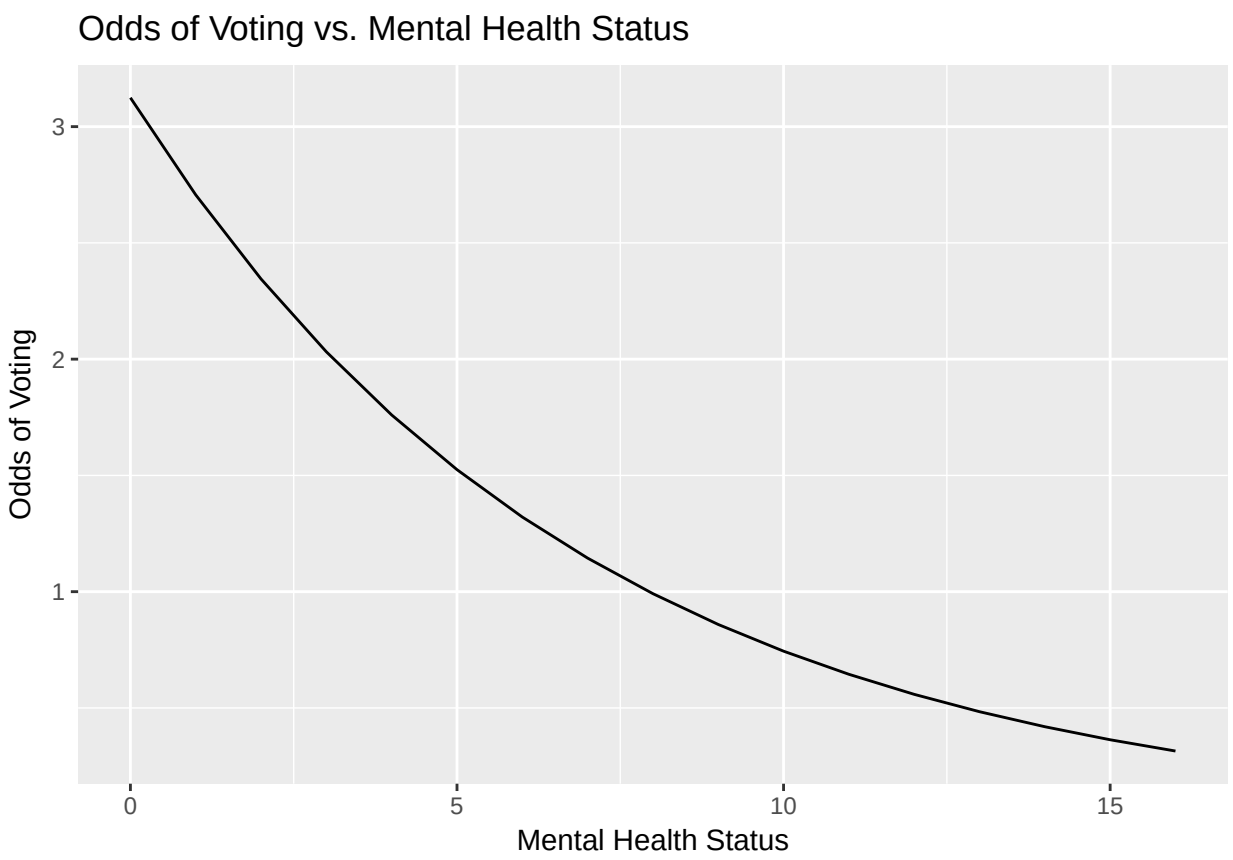
2.

```
## mhealth_sum
##  -0.1434752
```

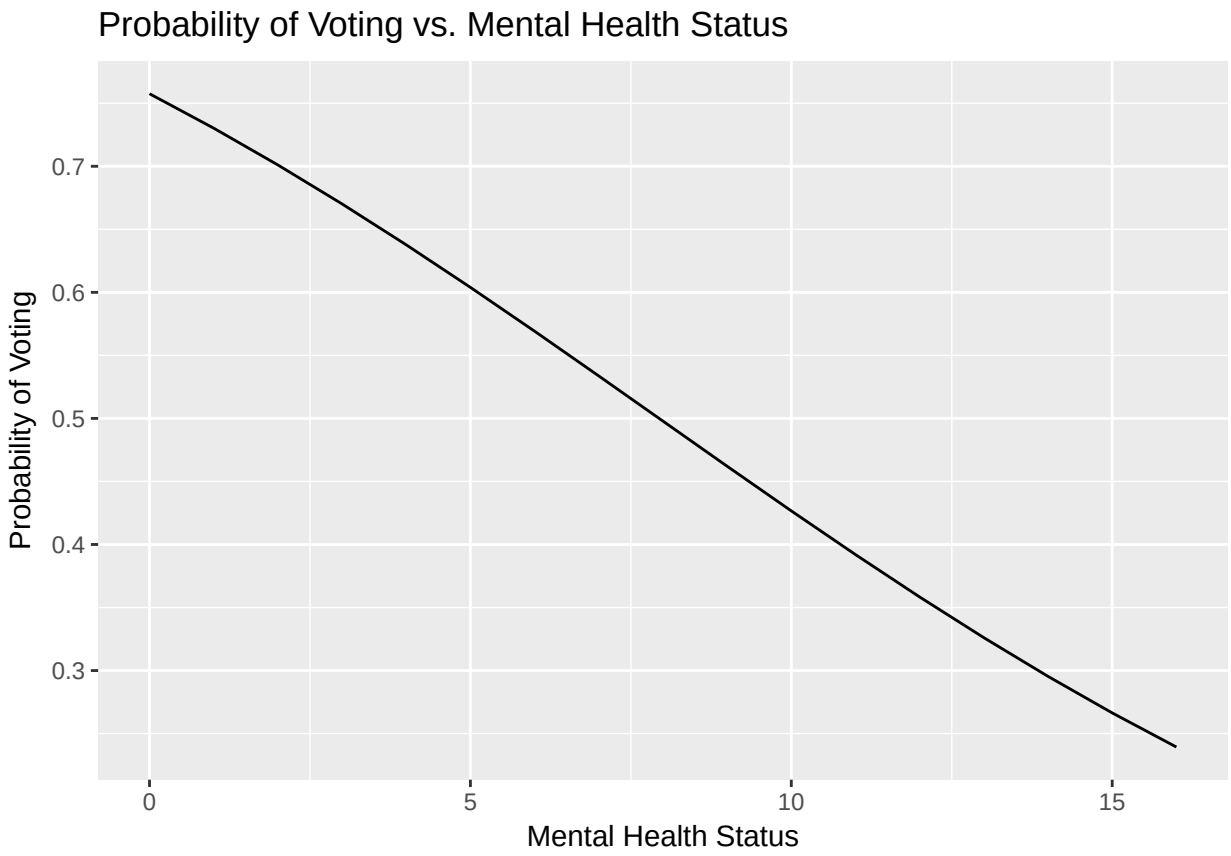


The estimated parameter for mental health is -0.1434752, this means the change in the log-odds associated with a one unit increase in `mhealth_sum` is -0.1434752.

3.


Odds of Voting vs. Mental Health Status
Odds of Voting
3
2
1
0
5
10
15
Mental Health Status

4.




```
## [1] -0.02917824
```

```
## [1] -0.03477821
```

The first difference for an increase in the mental health index from 1 to 2 is: -0.02917824.
The first difference for an increase in the mental health index from 5 to 6 is: -0.03477821.

```
## [1] 0.68
```

```
## [1] 0.01616628
```

```
## Area under the curve: 0.5401
```

Given a threshhold of .5, the accuracy rate is: 0.68 and the proportional reduction in error is: 0.0162; the AUC is 0.5401.

This is not very good model. The proportional reduction in error is `1.62%`, which is a tiny increase to the baseline rate. More than this, the AUC score only have increase 0.04 in AUC score than the baseline 0.5.

Multiple Variable Model

1.

- The random component of the probability distribution: Bernoulli distribution

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

* The linear predictor:

$$\eta_i = \beta_0 + \beta_1 X_{mhealth_sum,i} + \beta_2 X_{age,i} + \beta_3 X_{educ,i} + \beta_4 X_{black,i} + \beta_5 X_{female,i} + \beta_6 X_{married,i} + \beta_7 X_{inc10,i}$$

- Link function is:

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

2. Estimate and report results

```
##
## Call:
## glm(formula = vote96 ~ ., family = binomial(), data = data.mhealth)
##
## 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
##   (1667 observations deleted due to missingness)
## AIC: 1257.8
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:  glm(formula = vote96 ~ mhealth_sum + age + educ + married + inc10,
##     family = binomial(), data = data.mhealth)
##
## Coefficients:
## (Intercept)  mhealth_sum          age         educ      married
##    -4.20013     -0.08833      0.04211      0.22527      0.29386
##       inc10
##     0.06624
##
## Degrees of Freedom: 1164 Total (i.e. Null);  1159 Residual
##   (1667 observations deleted due to missingness)
```

```
## Null Deviance:       1468
## Residual Deviance: 1244  AIC: 1256
```

3.

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

```
## [1] 0.7236052
```

```
## Area under the curve: 0.7596
```

```
## (Intercept) mhealth_sum         age        educ       black      female
##   0.0135130   0.9147523   1.0434517   1.2569476   1.3138786   0.9831740
##     married       inc10
##   1.3457007   1.0720941
```

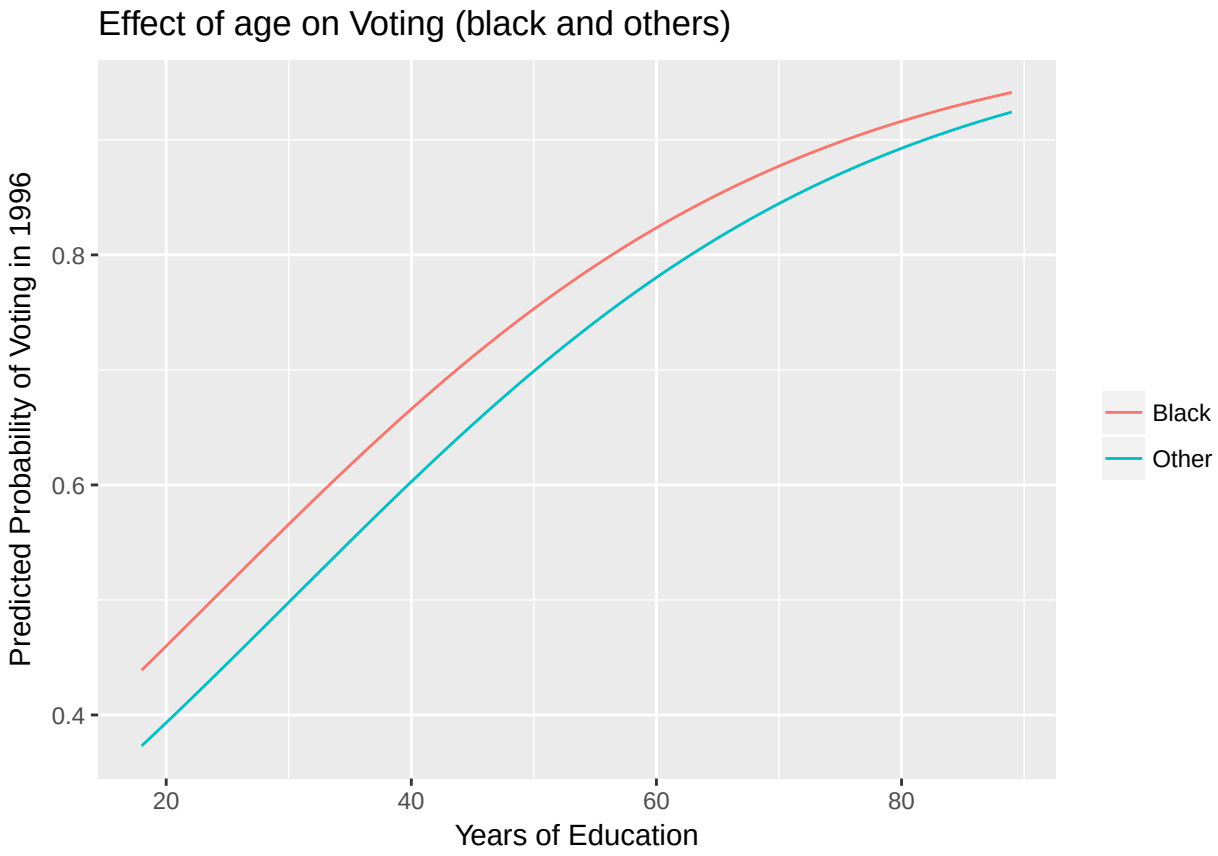
From the results above, we can say the model performs well. First of all, five of the indicators are significant: 'mhealth_sum', 'age' and 'education', 'married' and 'income' with coefficients (log-odds) -0.08833, 0.04211, 0.22527, 0.29386, and 0.06624 respectively. Secondly, the model fits the real values fairly well with accuracy rate 72.4% , proportional reduction in error (PRE) 14.8%, and AUC 0.7596. Comparing to the simple model, all three indicators has been improved.

Given the model is relatively valid, we argue that all seven factors have positive relationships with voting. Holding other factors constant, one unit index worse in mental health and one unit increase in age, education, race, gender, marriage and income will lead to on average increase of 0.9147523, 1.0434517, 1.2569476, 1.3138786, 0.9831740, 1.3457007 and 1.0720941 units in odds of voting.

```
## [1] 0.0141544
```

```
## [1] 0.01717635
```

When comparing the first differences in mental health index between multivariable model and the simple model, I hold the indices of 1 to 2 and 5 to 6; and to keep other variables constant, I chose 30-old-year single black female with university level of education(15 years) whose income is $50,000 as sample. Both differences are small than that of from the simple model.



Taking the most statistically significant effect on voting 'age', I am graphing two predicted probability curves of black and other races to see their influences on voting pattern, holding other non-binary variables as constants.

From the plot above, we can see that education indeed have a remarkable effect on voting decisions and the blackness also shift predicted probability upwards. Also, notice the difference between races decreases as years of education increases.

Part 2: Modeling TV Consumption

Estimate a Regression Model

1.

- The random component probability distribution: Poisson distribution

$$Pr(Y_i = yi|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- The linear predictor:

$$\eta_i = \beta_0 + \beta_1 X_{age,i} + \beta_2 X_{childs,i} + \beta_3 X_{educ,i} + \beta_4 X_{female,i} + \beta_5 X_{grass,i} + \beta_6 X_{hrsrelax,i} +$$
$$\beta_7 X_{black,i} + \beta_8 X_{social\_connect,i} + \beta_9 X_{voted04} + \beta_10 X_{xmovie,i} + \beta_11 X_{zodiac,i} + \beta_12 X_{dem,i} + \beta_13 X_{rep,i} + \beta_14 X_{ind,i}$$

- Link function:

$$\lambda_i = ln(\eta_i)$$

2. Estimate model and report results

```
##
## Call:
## glm(formula = tvhours ~ ., family = poisson, data = data.gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1120  -0.6741  -0.1144   0.4224   4.9257
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.0795865  0.2419794   4.461 8.14e-06 ***
## age                 0.0016522  0.0028397   0.582   0.5607
## childs             -0.0003896  0.0238729  -0.016   0.9870
## educ               -0.0292174  0.0126351  -2.312   0.0208 *
## female              0.0457000  0.0652987   0.700   0.4840
## grass              -0.1002726  0.0686146  -1.461   0.1439
## hrsrelax            0.0468472  0.0102790   4.558 5.18e-06 ***
## black               0.4657924  0.0841629   5.534 3.12e-08 ***
## social_connect      0.0437349  0.0407999   1.072   0.2837
## voted04            -0.0994787  0.0785680  -1.266   0.2055
## xmovie              0.0708408  0.0773420   0.916   0.3597
## zodiacAries        -0.1011364  0.1508248  -0.671   0.5025
## zodiacCancer        0.0267776  0.1451557   0.184   0.8536
## zodiacCapricorn    -0.2155760  0.1657034  -1.301   0.1933
## zodiacGemini        0.0285895  0.1481143   0.193   0.8469
## zodiacLeo          -0.1515676  0.1553215  -0.976   0.3291
## zodiacLibra        -0.0392537  0.1379102  -0.285   0.7759
## zodiacNaN          -0.2985240  0.2126126  -1.404   0.1603
## zodiacPisces       -0.1446731  0.1649895  -0.877   0.3806
## zodiacSagittarius  -0.2177846  0.1577638  -1.380   0.1674
## zodiacScorpio       0.0225911  0.1538460   0.147   0.8833
## zodiacTaurus       -0.1273891  0.1644799  -0.774   0.4386
## zodiacVirgo        -0.1240442  0.1564495  -0.793   0.4279
## dem                 0.0103276  0.0917055   0.113   0.9103
## rep                 0.0148615  0.0927662   0.160   0.8727
## ind                        NA         NA      NA       NA
## ---
## 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: 429.42  on 416  degrees of freedom
## AIC: 1600.4
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] -0.006825939
```

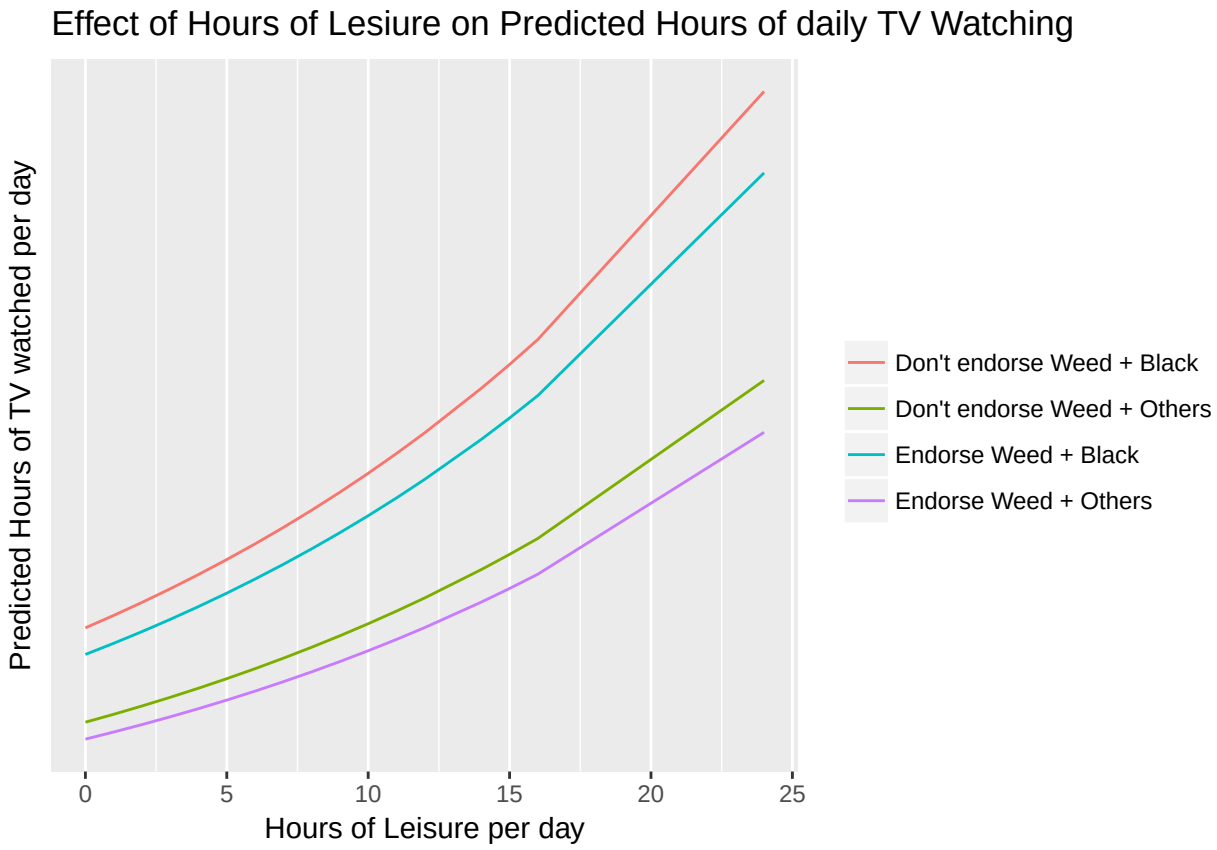
```
## [1] 0.2267574
```

```
## Area under the curve: 0.5488
```

3.

From the result above,the backwards AIC selection only return 4 predictor variables, not including the intercept: education(-0.03897), grass(-0.10787), hours of relax(0.04663) and race(0.45064). The value of each coefficient represent on average the change in log-count in TV watching hours due to one unit increase of the given variable. However, the model is not performing so well with PRE -0.006825939, accuracy 22.7% and AUC 54.88%.

When visualising the effect of 'hrsrelax' on predicted count, non-binary variable 'educ' need to be hold constant to plot 4 different combinations of 'grass' and 'black'.



Surprisingly, we see that there is a larger upward shift in predicted TV watching hour per day if the individual have a preference not to legalize marijuana.