Problem set #6: Generalized linear models

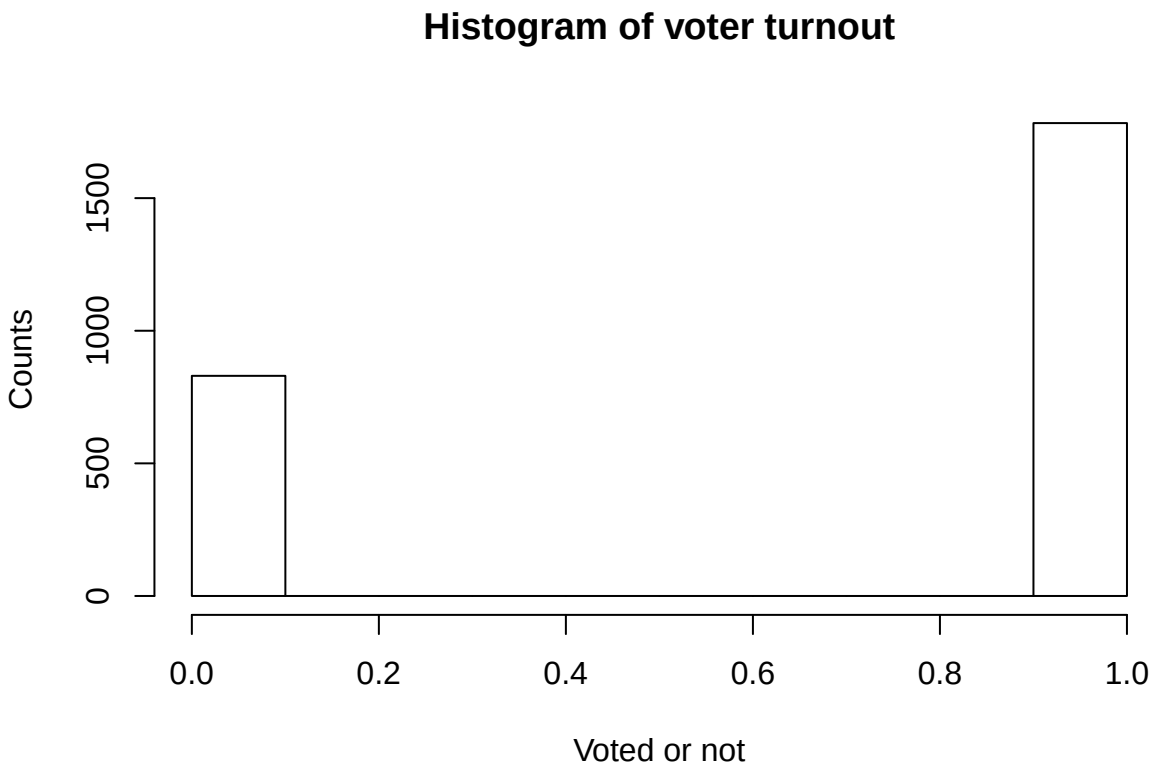
Jingyuan Zhou

2/16/2017

Part I: Modeling voter turnout

Describe the data

```
hist(mh_data$vote96, main="Histogram of voter turnout", xlab="Voted or not",
     ylab = 'Counts')
```

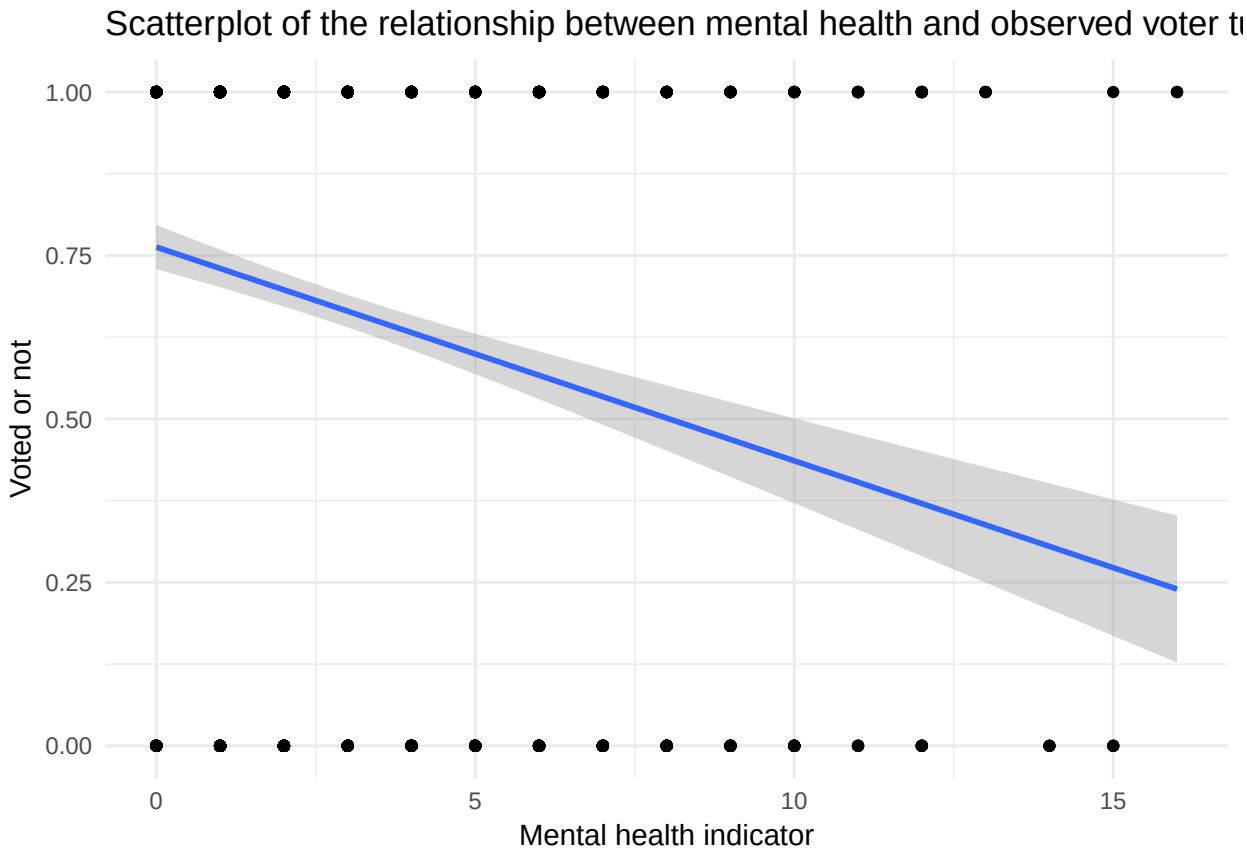



```
sum(mh_data$vote96, na.rm = TRUE)/length(mh_data$vote96)
```

```
## [1] 0.6295904
```

The unconditional probability of a given individual turning out to vote is 0.6295904.

```
ggplot(mh_data, aes(mhealth_sum, vote96)) +
  geom_point() +
  geom_smooth(method = "lm")+
  labs(title = 'Scatterplot of the relationship between mental health and observed voter turnout',
       x = 'Mental health indicator',y = 'Voted or not')
```



1.The line shows that as Mental health indicator increases, meaning people's depression mood becomes more severe, they are less likely to vote.

2.The problem with this line is that the y axis, whether people voted or not, is a discrete variable; however, the smooth line gives us values between 0 and 1. It's hard for us to interpret this result.

Basic model

```
# logistic regression model of the relationship between mental health and voter turnout.
log_model <- glm(formula = vote96 ~ mhealth_sum, family = binomial, data = mh_data)
tidy(log_model)
```

```
##          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
```

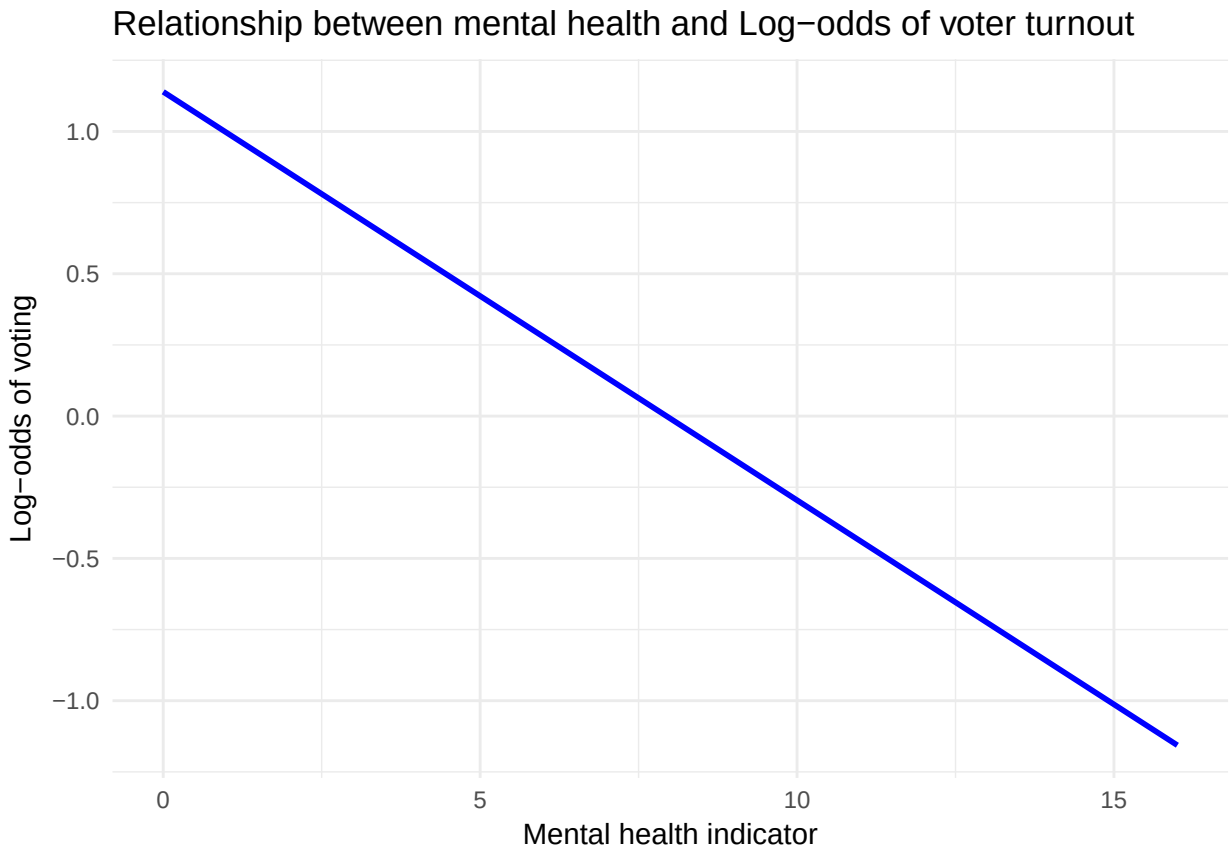
1.The relationship between mental health and voter turnout is statistically significant because p-value of mhealth_sum is 3.133883e-13 which is approximately zero and much smaller than 0.025, the critical level for a two-tail test at a 95% confidence interval.

2.Interpret the estimated parameter for mental health in terms of log-odds: When you estimate a logistic regression model the log-odds function is actually the function for which you are estimating parameters. Since the estimated parameter of mhealth_sum is -0.1434752. This means that for every one-unit increase in depression mood, we expect the log-odds of voting to decrease by 0.143.

Generate a graph of the relationship between mental health and the log-odds of voter turnout.

```
logit2prob <- function(x){
  exp(x) / (1 + exp(x))
```

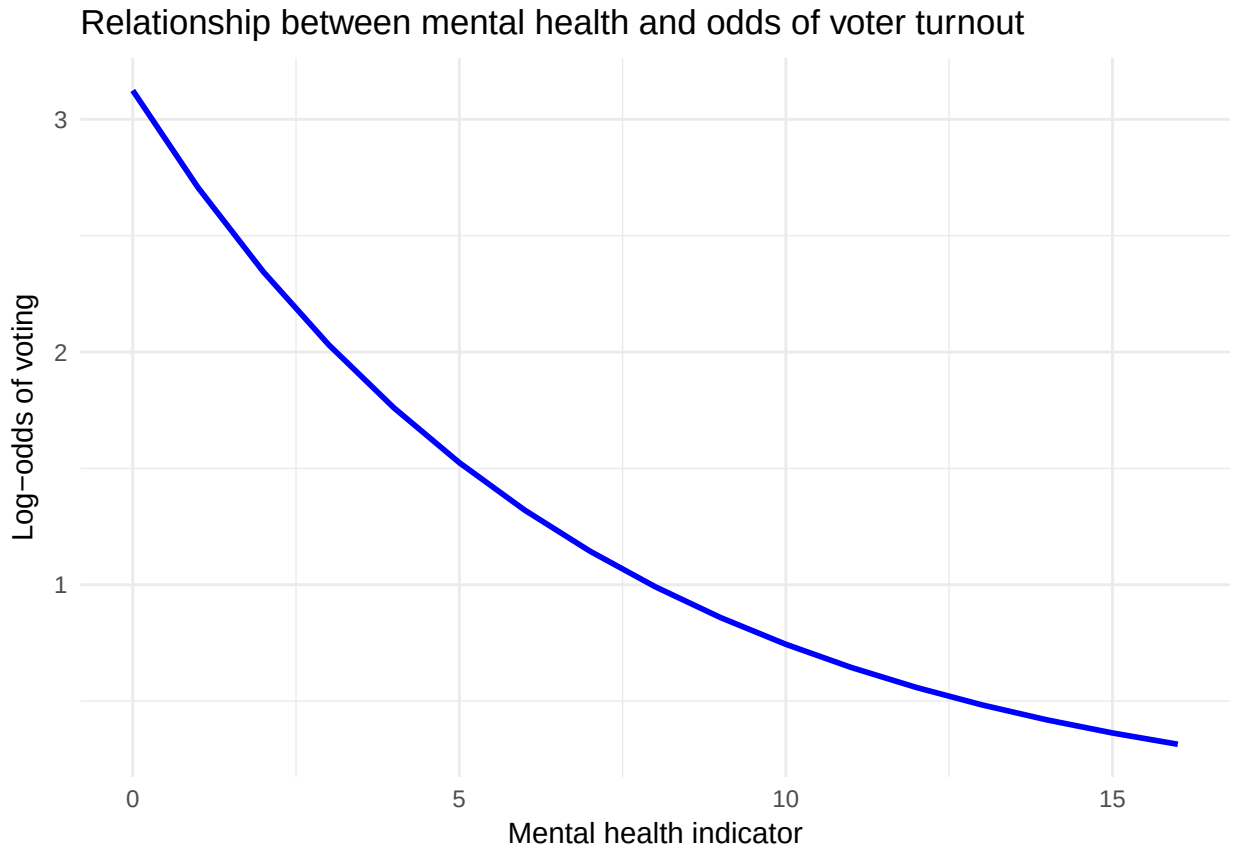
```
}

prob2odds <- function(x){
  x / (1 - x)
}

prob2logodds <- function(x){
  log(prob2odds(x))
}

vm_pred <- mh_data %>%
  add_predictions(log_model,  var = 'pred') %>%
  # predicted values are in the log-odds form - convert to probabilities
  mutate(prob = logit2prob(pred)) %>%
  mutate(odds = prob2odds(prob)) %>%
  mutate(logodds = prob2logodds(prob))

# graph it- logodds
ggplot(vm_pred, aes(x=mhealth_sum)) +
  geom_line(aes(y = logodds), color = "blue", size = 1) +
  labs(title = 'Relationship between mental health and Log-odds of voter turnout',
       x = "Mental health indicator",
       y = "Log-odds of voting")
```



3.Interpret the estimated parameter for mental health in terms of odds: Since the estimated parameter of mhealth_sum is -0.1434752 showing that per unit log-odds is -0.143, we can calculate that per unit odds is 0.8663423. This means that for every unit increase in depression mood, we expect the odds of voting to

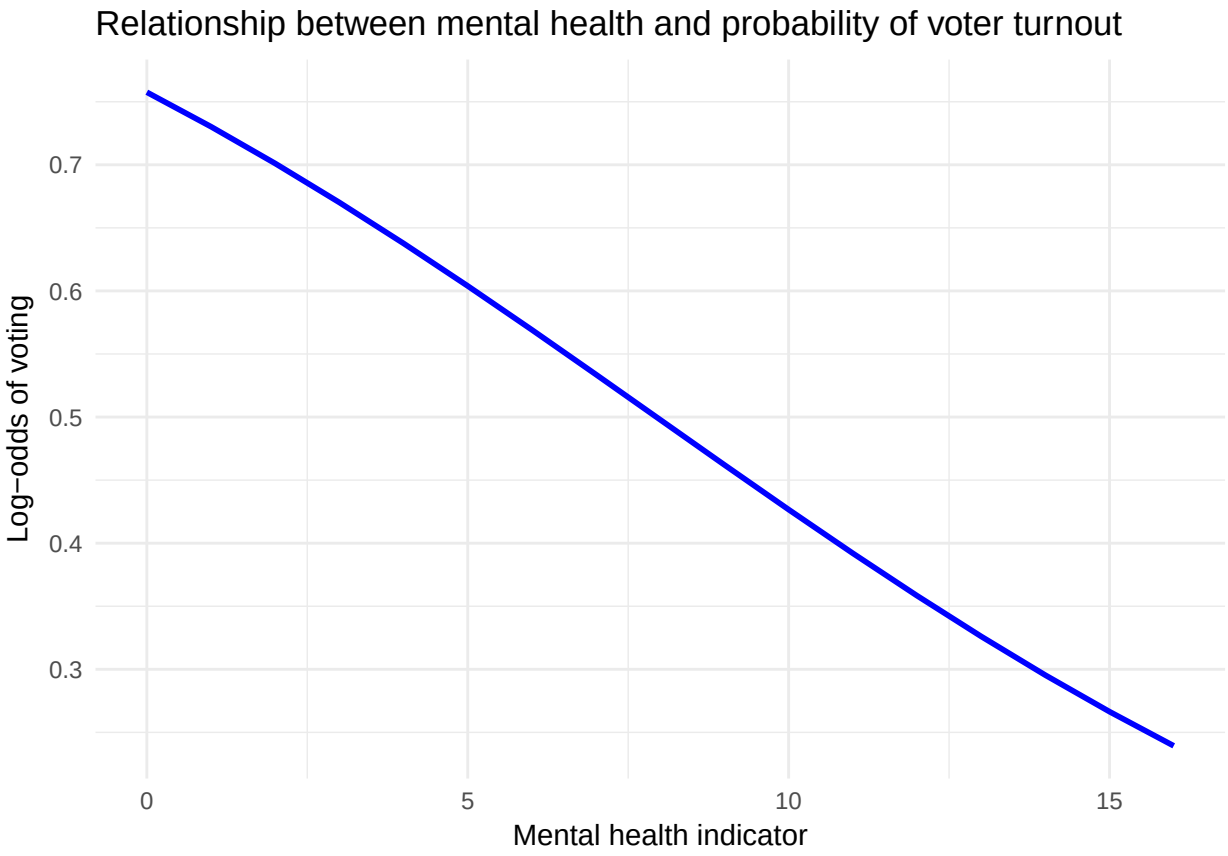
increase by 0.8663423.

```
ggplot(vm_pred, aes(x=mhealth_sum)) +
  geom_line(aes(y = odds), color = "blue", size = 1) +
  labs(title = 'Relationship between mental health and odds of voter turnout',
       x = "Mental health indicator",
       y = "Log-odds of voting")
```



4.Interpret the estimated parameter for mental health in terms of probabilities: with each unit increase in depression mood, the probability of voting is increasing by 0.4641914.

```
ggplot(vm_pred, aes(x=mhealth_sum)) +
  geom_line(aes(y = prob), color = "blue", size = 1) +
  labs(title = 'Relationship between mental health and probability of voter turnout',
       x = "Mental health indicator",
       y = "Log-odds of voting")
```




```
b0 <- 1.1392097
b1 <- -0.1434752
df12 <- exp(b0 + (2 * b1)) / (1 + exp(b0 + (2 * b1)))- exp(b0 + (1 * b1)) / (1 + exp(b0 + (1 * b1)))
df56 <- exp(b0 + (6 * b1)) / (1 + exp(b0 + (6 * b1)))- exp(b0 + (5 * b1)) / (1 + exp(b0 + (5 * b1)))
df12
```

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

```
df56
```

```
## [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.

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?

```
x_accuracy <- mh_data %>%
  add_predictions(log_model) %>%
  mutate(pred = logit2prob(pred),
         prob = pred,
         pred = as.numeric(pred > .5))

mean(x_accuracy$vote96 == x_accuracy$pred, na.rm = TRUE)
```

```
## [1] 0.677761
```

```
# function to calculate PRE for a logistic regression model
PRE <- function(model){
  # get the actual values for y from the data
```

```
  y <- model$y

  # get the predicted values for y from the model
  y.hat <- round(model$fitted.values)

  # calculate the errors for the null model and your model
  E1 <- sum(y != median(y))
  E2 <- sum(y != y.hat)

  # calculate the proportional reduction in error
  PRE <- (E1 - E2) / E1
  return(PRE)
}

PRE(log_model)
```

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

```
library(pROC)
auc <- auc(x_accuracy$vote96, x_accuracy$prob)
auc
```

```
## Area under the curve: 0.6243
```

The accuracy rate of this model is 0.677761. The proportional reduction in error (PRE) is 0.01616628. The AUC for this model is 0.6243. It's not a very good model because the accuracy rate is not significantly larger than 0.5, and its PRE is very little. The AUC also shows that it's only 12% higher than the baseline, which would be 50%.

Multiple variable model

1.Three components of the GLM

- Probability distribution (random component): the conditional distribution of vote96 given information of age, education, black, female, married and inc10 is a Bernoulli distribution. $\mathrm{P}(Y_i = y_i|p) = p^{y_i}(1-p)^{1-y_i}$
- Linear predictor: $\eta_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_3 black_i + \beta_4 female_i + \beta_5 married_i + \beta_6 inc10_i$
- Link function: $p_i = e^{\eta_i}/(1+e^{\eta_i})$

2.Estimate the model and report your results.

```
mglm_model <- glm(vote96 ~ age+educ+black+female+married+inc10, data=mh_data, family=binomial())
tidy(mglm_model)
```

```
##          term    estimate   std.error   statistic      p.value
## 1 (Intercept) -4.41114892 0.343913259 -12.8263416 1.167481e-37
## 2         age  0.04303838 0.003362445  12.7997289 1.645246e-37
## 3        educ  0.21891013 0.020762110  10.5437322 5.430025e-26
## 4       black  0.18929436 0.138718329   1.3645951 1.723803e-01
## 5      female  0.05235039 0.098180190   0.5332073 5.938901e-01
## 6     married  0.23224748 0.107010030   2.1703338 2.998157e-02
## 7       inc10  0.06800413 0.018324401   3.7111240 2.063410e-04
```

Looking at the p-values of these variables, we can see that *black*, *female* and *married* have p-values larger than 0.025. Their p-values are 1.723803e-01, 5.938901e-01 and 2.998157e-02. So I'll rebuild the model by removing these variables. The linear predictor then becomes $\eta_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \beta_4 inc10_i$

```
mglm_model <- glm(vote96 ~ age+educ+inc10, data=mh_data, family=binomial())
tidy(mglm_model)
```

```
##          term    estimate   std.error  statistic      p.value
## 1 (Intercept) -4.21540098 0.331885800 -12.701360 5.811109e-37
## 2         age  0.04321149 0.003357197  12.871302 6.529256e-38
## 3        educ  0.21207984 0.020510609  10.340007 4.645023e-25
## 4       inc10  0.07962321 0.016773752   4.746893 2.065653e-06
```

```
#summary(mglm_model)

m_accuracy <- mh_data %>%
  add_predictions(mglm_model) %>%
  mutate(pred = logit2prob(pred),
         prob = pred,
         pred = as.numeric(pred > .5))

mean(m_accuracy$vote96 == m_accuracy$pred, na.rm = TRUE)
```

```
## [1] 0.7170056
```

```
PRE(mglm_model)
```

```
## [1] 0.09292649
```

```
#library(pROC)
auc <- auc(m_accuracy$vote96, m_accuracy$prob)
auc
```

```
## Area under the curve: 0.7331
```

3.Compare to the basic model, the model has considerably better performance. The proportional reduction in error (PRE) increases from 0.01616628 to 0.09292649; accuracy rate also increases from 0.677761 to 0.7170056; auc increases from 0.6243 to 0.7331.

Interpreting the model, we can see that all three variables, *age*, *inc10* and *educ* have p-values that are approximating zeo, which shows that they are all statistically significant. Out of these three, *educ* has the most significant effect because its estimate parameter is 0.21207984, which is more than three times as effect of each of the other two variables. Intuitively, it's reasonable because as people have more years of education, they have more desire to participate in politics and try to make a difference with their votes. Thus, they are more likely to vote than people with less education. Estimate parameters of age and income show that as people have more income and grow older, they are more likely to vote. These findings all correspond to our observation in real life and common sense.

Part II: Modeling TV consumption

Estimate a regression model

1.Three components of the GLM

- Probability distribution (random component):the conditional distribution of vote96 given information of other variables follows Poisson distribution. $\mathrm{P}(tvhours = k|\lambda) = \lambda^k e^{-\lambda}/k!$
- Linear predictor: $tvhours_i = \beta_0 + \beta_1 age_i + \beta_2 childs_i + \beta_3 educ_i + \beta_4 female_i + \beta_5 grass_i + \beta_6 hrsrelax_i + \beta_7 black + \beta_8 socialconnect_i + \beta_9 voted04_i + \beta_{10} xmovie_i + \beta_{11} zodiac + \beta_{12} dem_i + \beta_{13} rep_i + \beta_{14} ind_i$
- Link function: $g(\lambda) = log(tvhours_i)$

```
tv_model <- glm(tvhours ~ ., data=gss_data, family=poisson)
tidy(tv_model)
```

```
##                 term      estimate  std.error   statistic      p.value
## 1        (Intercept)  1.0795865332 0.24197937  4.46148172 8.139489e-06
## 2                age  0.0016521563 0.00283970  0.58180660 5.606970e-01
## 3             childs -0.0003896381 0.02387285 -0.01632139 9.869780e-01
## 4              educ -0.0292174017 0.01263513 -2.31239477 2.075594e-02
## 5             female  0.0457000419 0.06529870  0.69986145 4.840138e-01
## 6              grass -0.1002725928 0.06861458 -1.46138902 1.439087e-01
## 7           hrsrelax  0.0468472156 0.01027902  4.55755697 5.175205e-06
## 8              black  0.4657923645 0.08416286  5.53441700 3.122653e-08
## 9     social_connect  0.0437348968 0.04079985  1.07193760 2.837481e-01
## 10           voted04 -0.0994787227 0.07856798 -1.26614844 2.054599e-01
## 11            xmovie  0.0708407795 0.07734198  0.91594210 3.596973e-01
## 12       zodiacAries -0.1011363820 0.15082478 -0.67055546 5.025038e-01
## 13      zodiacCancer  0.0267776138 0.14515566  0.18447516 8.536407e-01
## 14   zodiacCapricorn -0.2155760173 0.16570344 -1.30097493 1.932670e-01
## 15      zodiacGemini  0.0285894938 0.14811434  0.19302313 8.469409e-01
## 16         zodiacLeo -0.1515676052 0.15532153 -0.97583129 3.291481e-01
## 17       zodiacLibra -0.0392537020 0.13791025 -0.28463223 7.759259e-01
## 18         zodiacNaN -0.2985239737 0.21261263 -1.40407452 1.602967e-01
## 19      zodiacPisces -0.1446730925 0.16498953 -0.87686227 3.805615e-01
## 20 zodiacSagittarius -0.2177845756 0.15776382 -1.38044690 1.674491e-01
## 21     zodiacScorpio  0.0225910524 0.15384596  0.14684202 8.832567e-01
## 22      zodiacTaurus -0.1273890642 0.16447992 -0.77449616 4.386374e-01
## 23       zodiacVirgo -0.1240441866 0.15644951 -0.79287040 4.278533e-01
## 24               dem  0.0103275813 0.09170546  0.11261686 9.103343e-01
## 25               rep  0.0148615484 0.09276621  0.16020433 8.727201e-01
```

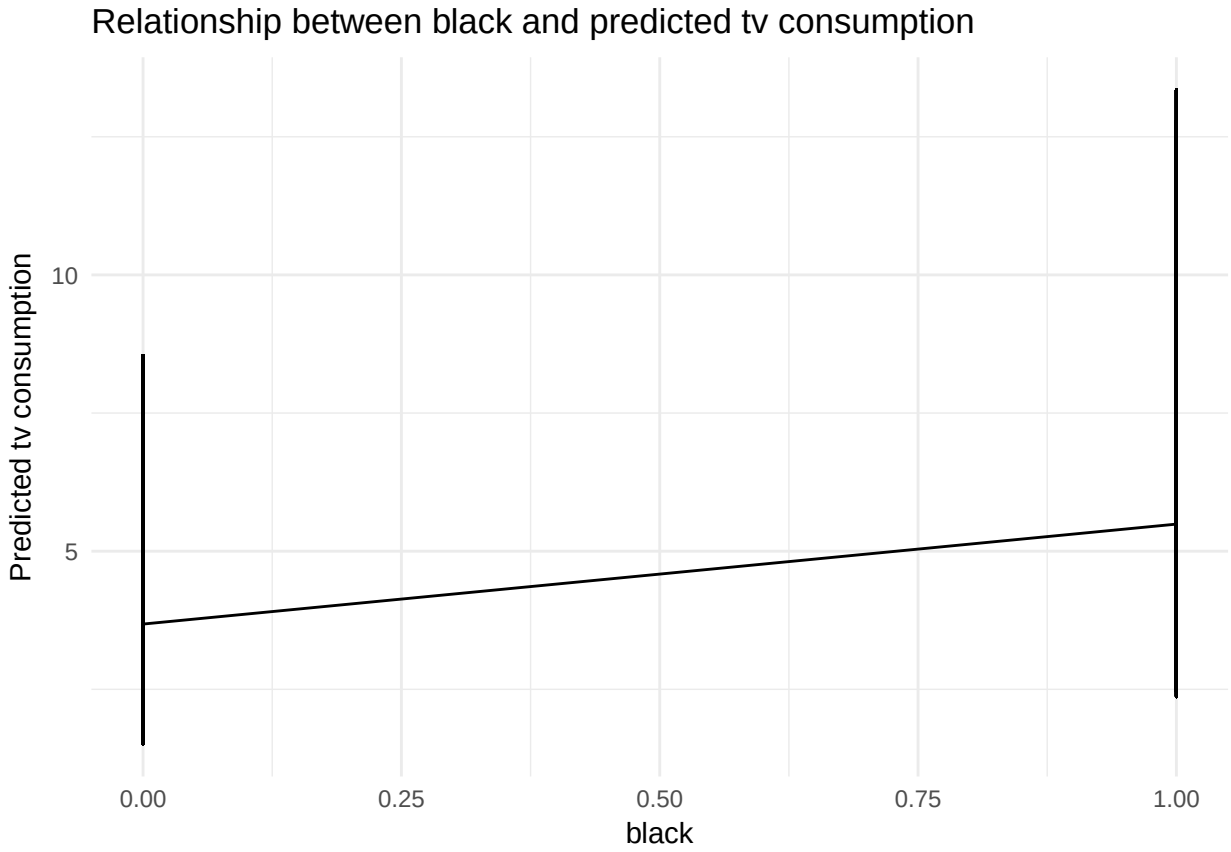
After including all variables, it seems that only three variables, namely *educ*, *hrsrelax* and *black*, are statistically significant because their p-values, 2.075594e-02, 5.175205e-06 and 3.122653e-08 are all less than 0.025. Thus, I rebuilt the model with only these three variables. The linear predictor then becomes: $tvhours_i = \beta_0 + \beta_1 educ_i + \beta_2 hrsrelax_i + \beta_3 black_i$

2.Model and results

```
tv_model <- glm(tvhours ~ educ+ hrsrelax + black, data=gss_data, family=poisson)
tidy(tv_model)
```

```
##          term    estimate   std.error statistic      p.value
## 1 (Intercept)  1.25658633 0.111309838 11.289086 1.485736e-29
## 2        educ -0.04208422 0.007549591 -5.574370 2.484267e-08
## 3    hrsrelax  0.03701533 0.006243058  5.929038 3.047150e-09
## 4       black  0.44631367 0.046871343  9.522101 1.697123e-21
```

```
tv_accuracy <- gss_data %>%
  data_grid(tvhours, educ, hrsrelax, black, .model = tv_model)%>%
  add_predictions(tv_model)%>%
  mutate(pred = exp(pred))

ggplot(tv_accuracy, aes(x = black, y = pred))+
  #geom_point(aes(y = tvhours), alpha = 0.5)
  geom_line(aes(x = black, y = pred))+
  labs(title = 'Relationship between black and predicted tv consumption',
    x = 'black', y = 'Predicted tv consumption')
```



3.Inspecting the estimate parameters for different variables, we can see that *black* has the most significant effect on tv consumption among all variables. The estimate parameter, 0.44631367, shows that being black will on average increase log of the tv consumption by 0.44631367 unit. We can visualize that effect from the graph.

Interpretting the other aspects of our model, we can realize that age, gender, number of children and party affilication all have no statistically significant influcence on TV consumption. One could argue that hours of relaxation and years of education could reflect some of these variables, but they are also intuitively directly related to tv consumption. Note that *educ* has a negatvie estimate parameter, so as years of education increase by one, the log of amount of TV assumption will decrease by around 0.0420 on avergae.

What's absurd/interesting about this model is that it shows the incredible effect of *black*. Comparing its estimate parameter to those of educ and hrsrelax, we can see that it has more than 10 times of the effect on TV consumption as each of the other two. Is there an systematic error in the data collection? Are there other variables that result in this result not included in the survey? What might be the statistically or sociological reasons behind this finding could be an interesting question to answer.