Problem Set 6: Generalized linear Models

William L. Guzman

February 20, 2017

Part 1: Modeling voter turnout

Describe the data (1 point)

```
#Load the Libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.3.2
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'tibble' was built under R version 3.3.2
```

```
## Warning: package 'tidyr' was built under R version 3.3.2
```

```
## Warning: package 'readr' was built under R version 3.3.2
```

```
## Warning: package 'purrr' was built under R version 3.3.2
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(tidyverse)
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 3.3.2
```

```
library(broom)
```

```
##
## Attaching package: 'broom'
```

```
## The following object is masked from 'package:modelr':
##
##     bootstrap
```

```
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 3.3.2
```

```
library(tidyverse)
library(modelr)
library(broom)
library(forcats)
library(pROC)
```

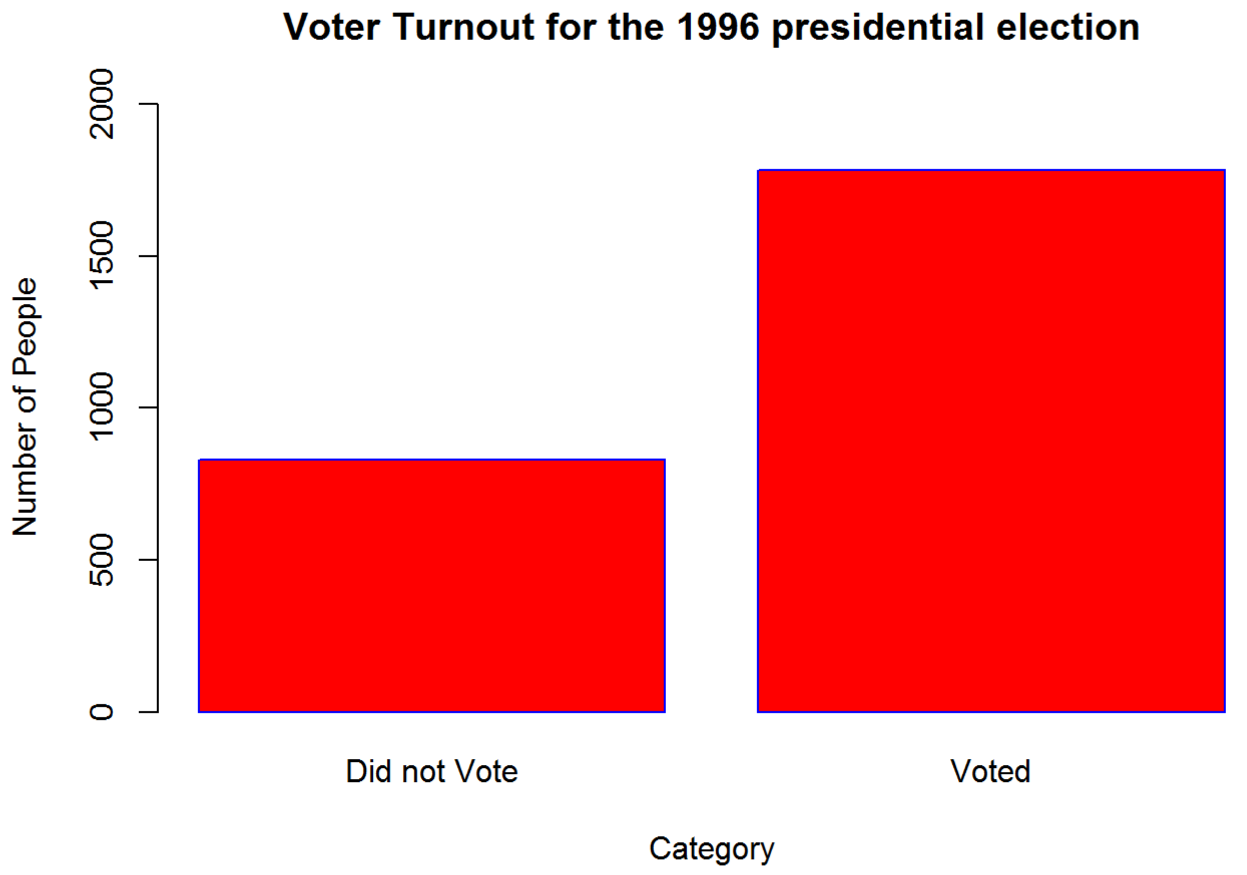
```
## Warning: package 'pROC' was built under R version 3.3.2
```

```
## Type 'citation("pROC")' for a citation.
```

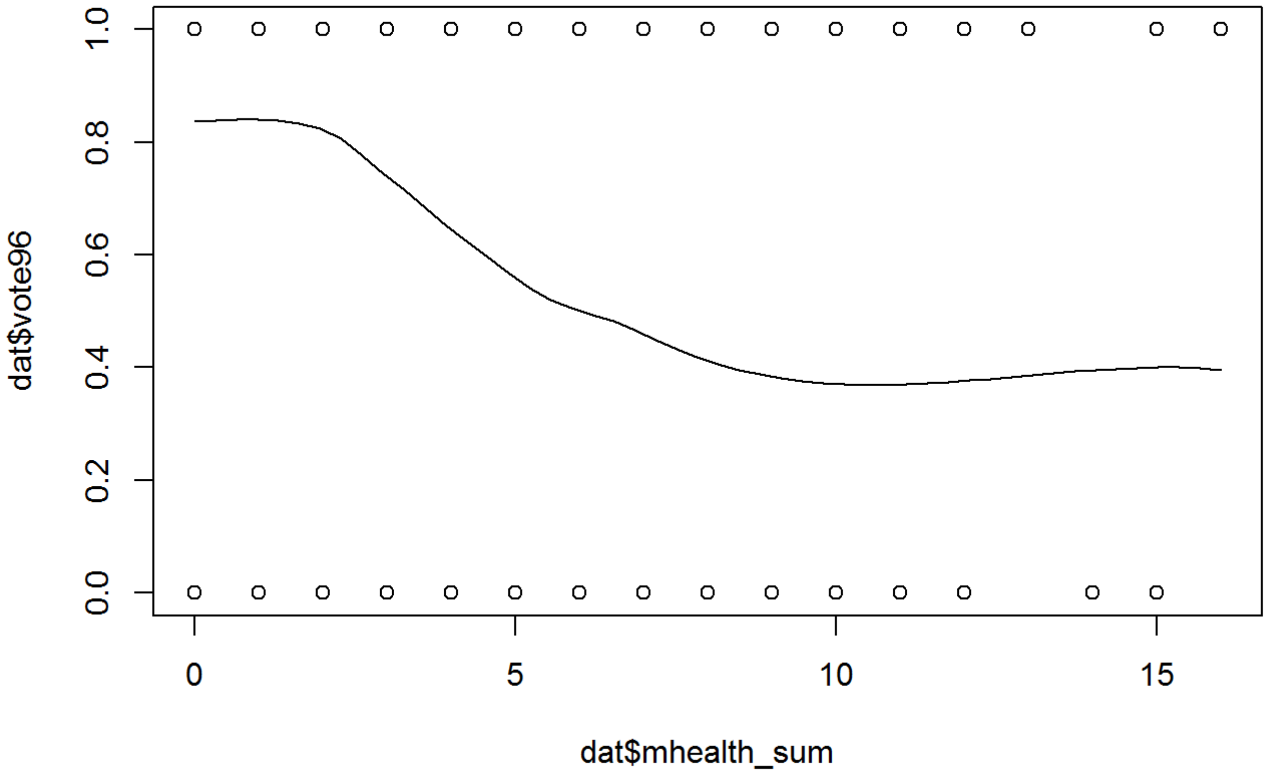
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
#Getting the data
datapath <- "C:/Users/Walle/Documents/RScript/Data"

dat <- read.csv(file=paste(datapath,"mental_health_np.csv",sep="/"))

#Plotting the data
counts <- table(dat$vote96)

barplot(counts, main = "Voter Turnout for the 1996 presidential election", xlab = "Category", yl
ab = "Number of People", border = "blue", col = "red",
        names.arg = c("Did not Vote", "Voted"), ylim = c(0,2000))
```



```
#Plotting the scatterplot
scatter.smooth(x = dat$mhealth_sum, y= dat$vote96)
```



1. What is the unconditional probability of a given individual turning out to vote?

The unconditional propability will be 1,783/2,613 = 0.682357, or, 68.23% of the people are more likely to vote.

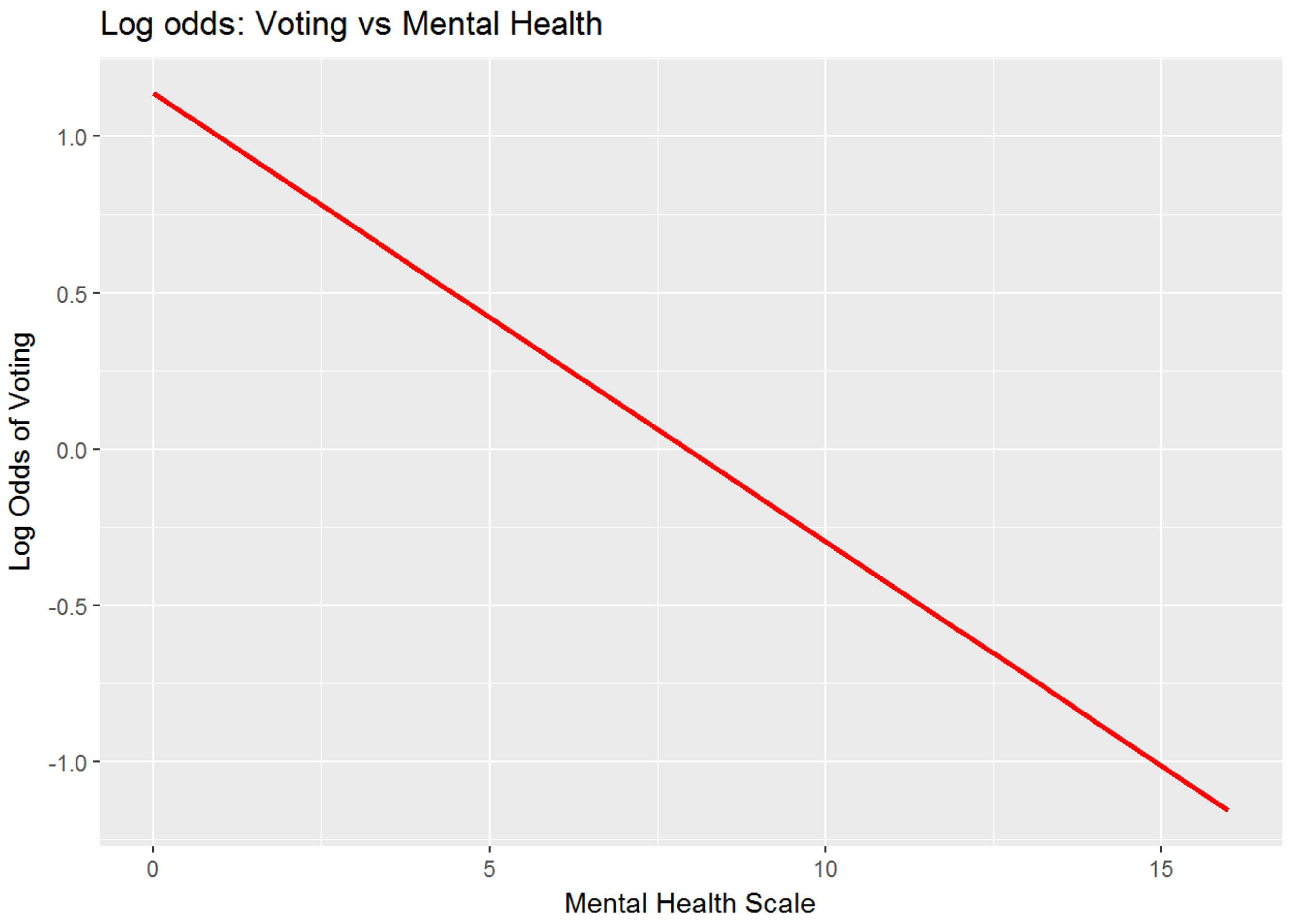
2. What information does this tell us? What is problematic about this linear smoothing line?

This graph does not tells us a lot of information because we are using categorical variables to explain a plot that explain continuos variables. The problem with the linear smoothing line is that none of the values are fitting the line.

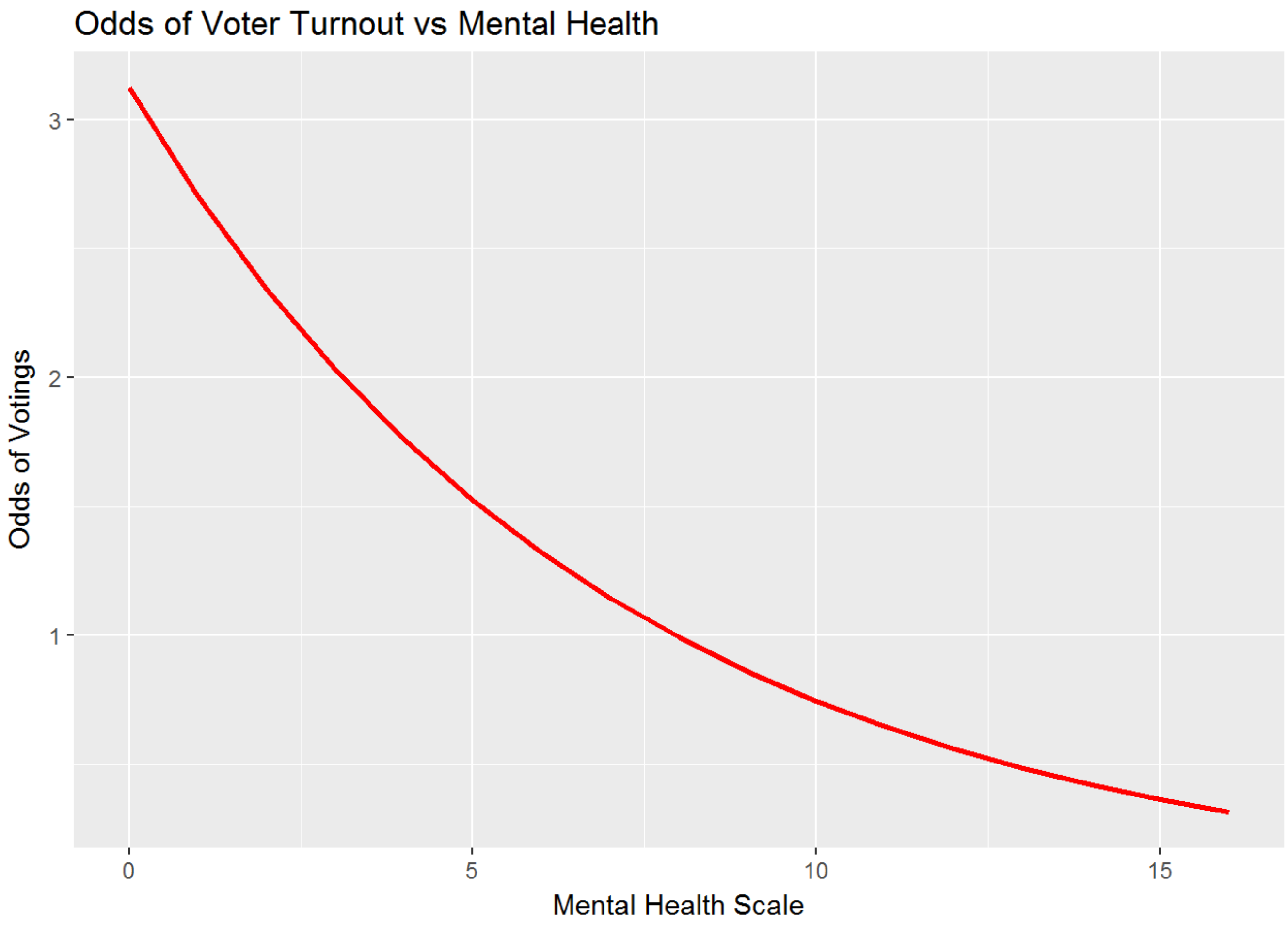
```
#1.2 Generate a graph of the relationship between mental health and the log-odds of voter turnou
t.
#From Notes and Rmarkdown
logit2prob <- function(x){
  exp(x) / (1 + exp(x))}

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

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

logModel1 = glm(vote96 ~ mhealth_sum, data = dat, family = binomial)

summary(logModel1)
```

```
## 
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = binomial, data = dat)
## 
## 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
##   (1510 observations deleted due to missingness)
## AIC: 1620.7
## 
## Number of Fisher Scoring iterations: 4
```

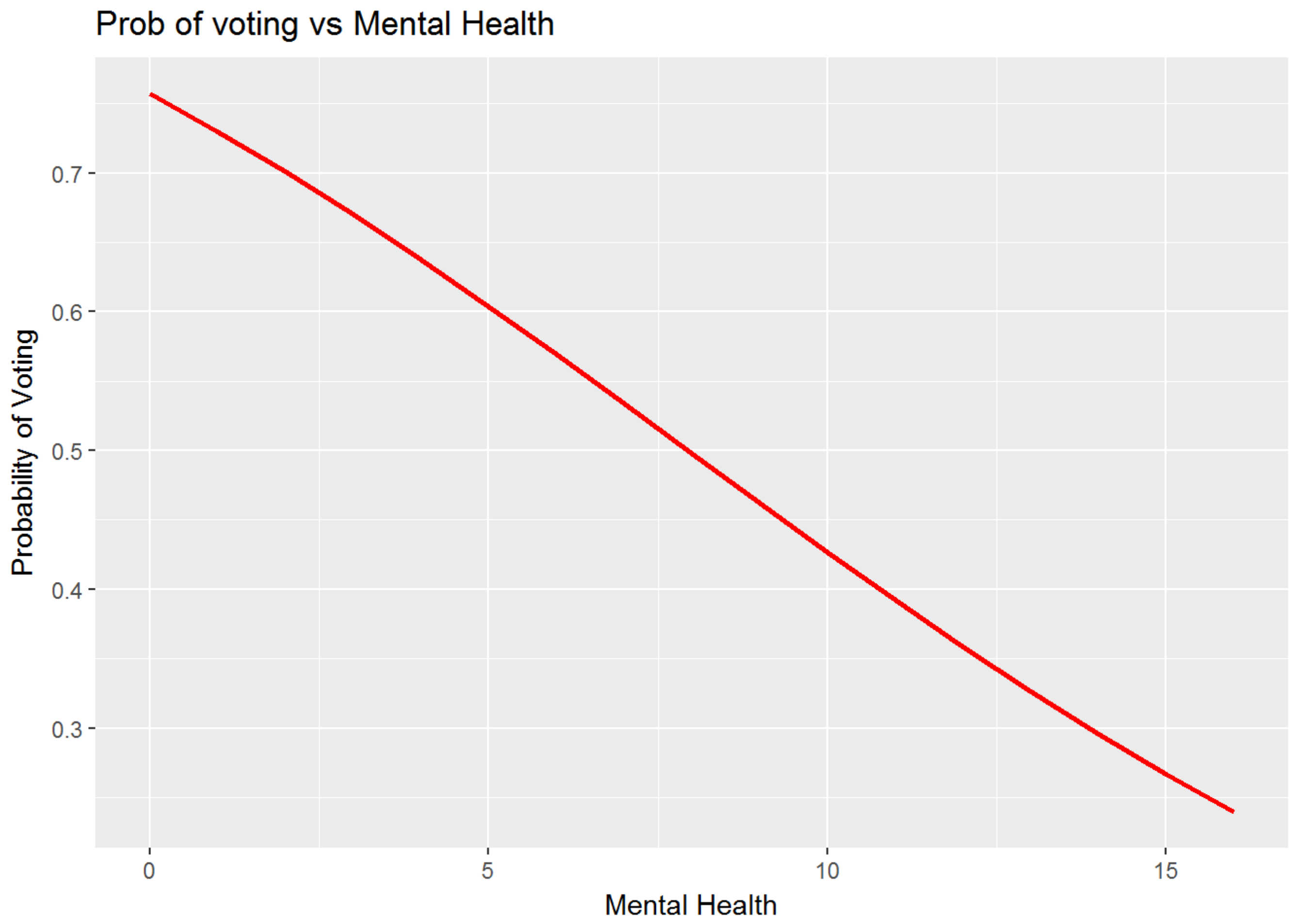
```
votePred <- add_predictions(dat, logModel1)
votePred <- mutate(votePred, prob = logit2prob(pred))
votePred <- mutate(votePred, odds = prob2odds(prob))
votePred <- na.omit(votePred)

#Graph the Model
ggplot(votePred, aes(mhealth_sum, pred)) +
    geom_line(color = "red", size = 1) +
    labs(
        title = "Log odds: Voting vs Mental Health", x = "Mental Health Scale", y = "Log Odds of
 Voting")
```



```
#1.3 Generate a graph of the relationship between mental health and the odds of voter turnout.
ggplot(votePred, aes(mhealth_sum, odds)) +
   geom_line(color = "red", size = 1) +
   labs(
        title = "Odds of Voter Turnout vs Mental Health", x = "Mental Health Scale", y = "Odds o
f Votings")
```



```
#1.4 Generate a graph of the relationship between mental health and the probability of voter turnout
ggplot(votePred, aes(mhealth_sum, prob)) +
   geom_line(color = "red", size = 1) +
   labs(
     title = "Prob of voting vs Mental Health",x = "Mental Health", y = "Probability of Voting")
```



```
#Interpret the estimated parameter for mental health in terms of probabilities
datSet <- data_grid(dat, mhealth_sum)
datSet<-  add_predictions(datSet, logModel1)
datSet <- mutate(datSet, prob = logit2prob(pred))

#Difference in 1 to 2
incr12 <- datSet[3,] - datSet[2,]
incr12 <- incr12$prob

#Difference in 5 to 6
incr56 <- datSet[7,] - datSet[6,]
incr56 <- incr56$prob

incr12
```

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

```
incr56
```

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

```
#Estimate the accuracy rate, proportional reduction in error (PRE), and the AUC for this model.
PRE <- function(model){
  y <- model$y
  y.hat <- round(model$fitted.values)

  E1 <- sum(y != median(y))
  E2 <- sum(y != y.hat)

  PRE <- (E1 - E2) / E1

  return(PRE)
}

voteAccuracy <- add_predictions(dat, logModel1)

voteAccuracy <- mutate(voteAccuracy, pred = as.numeric(logit2prob(pred) > .5))

accRate <- mean(voteAccuracy$vote96 == voteAccuracy$pred, na.rm = TRUE)

proPRE <- PRE(logModel1)

aucValues <- auc(voteAccuracy$vote96, voteAccuracy$pred)

accRate
```

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

```
proPRE
```

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

```
aucValues
```

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

Basic model (3 points)

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

After looking at the basic model from the logistic regression of voter turnout dependent on the mental health, we can see that by looking at the p-value of 0.000000000000313, there is an statistically relationship between voter turnout and mental health.

4. What is the first difference for an increase in the mental health index from 1 to 2? What about for 5 to 6?

The difference for an increase in the mental health index from 1 to 2 is -002917824, and from 5 to 6 is -0.03477821.

5. Do you consider it to be a good model?

With an accuracy rate of 0.677 we can say that we have a good model, but it can be improve.

Part 1.2: Multiple Variable Model (3 points)

```
#Using all the variables to interpret the model.
logModel2 <- glm(vote96 ~ . , data = dat, family = binomial)

summary(logModel2)
```

```
##
## Call:
## glm(formula = vote96 ~ ., family = binomial, data = dat)
##
## 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
```

```
tidy(logModel2)
```

```
##          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
```

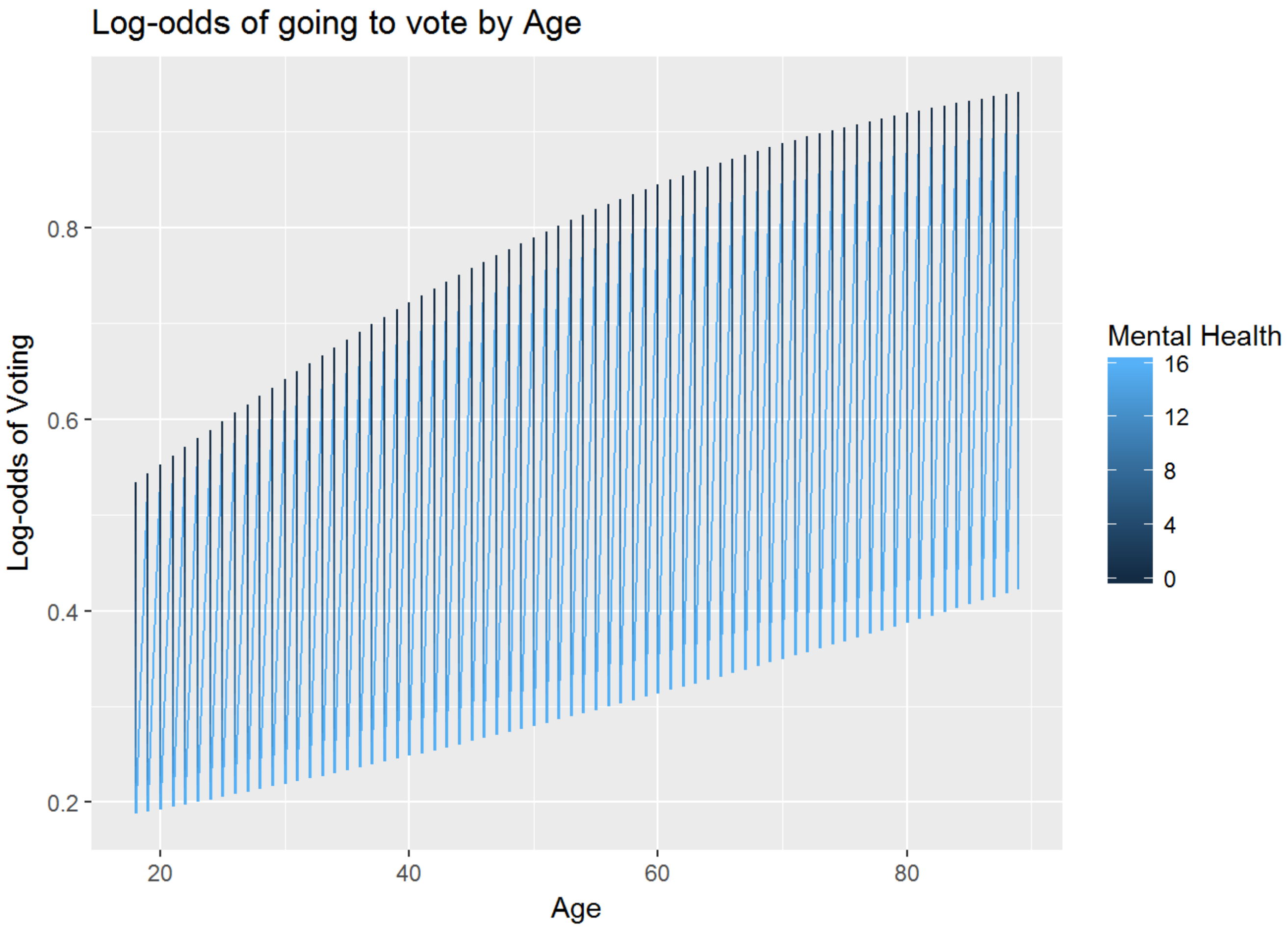
```
#Logmodel with only mhealth_sum and education
logModel3 <- glm(vote96 ~ mhealth_sum * age, data = dat, family = binomial)

summary(logModel3)
```

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum * age, family = binomial,
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3210  -1.1883   0.6535   0.8919   1.7207
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.532227   0.269000  -1.979   0.0479 *
## mhealth_sum     -0.076588   0.057368  -1.335   0.1819
## age              0.037126   0.005873   6.322 2.58e-10 ***
## mhealth_sum:age -0.001306   0.001226  -1.065   0.2868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1667.6  on 1319  degrees of freedom
## Residual deviance: 1538.4  on 1316  degrees of freedom
##   (1512 observations deleted due to missingness)
## AIC: 1546.4
##
## Number of Fisher Scoring iterations: 4
```

```
#Getting the Log Odds
dat_age_mental <- dat %>%
  data_grid(mhealth_sum, age) %>%
  add_predictions(logModel3) %>%
  mutate(pred = logit2prob(pred))
dat_age_mental
```

```
## # A tibble: 1,224 × 3
##    mhealth_sum   age      pred
##          <dbl> <dbl>     <dbl>
## 1            0    18 0.5339557
## 2            0    19 0.5431815
## 3            0    20 0.5523779
## 4            0    21 0.5615386
## 5            0    22 0.5706575
## 6            0    23 0.5797287
## 7            0    24 0.5887463
## 8            0    25 0.5977048
## 9            0    26 0.6065985
## 10           0    27 0.6154220
## # ... with 1,214 more rows
```

```
#Ploting the Interactive results
ggplot(dat_age_mental, aes(age, pred, color = mhealth_sum)) +
  geom_line() +
  labs(title = "Log-odds of going to vote by Age",
       x = "Age",
       y = "Log-odds of Voting",
       color = "Mental Health")
```



```
#Finding the accuracy of our model.
voteAccuracy2 <- add_predictions(dat, logModel2)

voteAccuracy2 <- mutate(voteAccuracy2, pred = as.numeric(logit2prob(pred) > .5))

accRate2 <- mean(voteAccuracy2$vote96 == voteAccuracy2$pred, na.rm = TRUE)

proPRE2 <- PRE(logModel2)

aucValues2 <- auc(voteAccuracy2$vote96, voteAccuracy2$pred)

accRate2
```

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

```
proPRE2
```

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

```
aucValues2
```

```
## Area under the curve: 0.6394
```

6. Write out the three components of the GLM for your specific model of interest.

1. Random Component

Our random component is Y = Voter Turnout(vote96) and is binomial.

2. Linear Predictor

$\eta = \beta_0 + \beta_1 mhealthsum + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 female + \beta_6 married + \beta_7 inc10$
The linear predictor’s variables are: Categorical Variables: Mental Health Index (mhealth_sum), Color (black), Gender (Female) and Social Status (Married). Continuos Variables: Age (age) and Income (inc10)

3. Link Function

$log(\mu) = \eta_i$
The function is Logit

7. Interpret the results in paragraph format.

After doing a summary of the model with all the variables in our dataset, we can see that the variable of color, gender and social status are not statistically significant for trying to predict the 1996 voter turnout. We can see that the accuracy of the models explain 72.24% of the model. In our last model, using only a single predictor variable (mhealth_sum) our model was explained by 67.7%. We can decide to create a model with only the variables that are statistically significant to see and observe how much our model improves. When we create an interactive log norm of mental health and age, we can see that the older you are with less mental health issues, the more likely you will show up to vote.

Part 2. Modeling tv consumption

```r
dat2 <- read.csv(file=paste(datapath,"gss2006.csv",sep="/"))

#Create the model with all the variables
tvlogModel1 <- glm(tvhours ~ ., data = dat2, family = poisson())

summary(tvlogModel1)
```

```
##
## Call:
## glm(formula = tvhours ~ ., family = poisson(), data = dat2)
##
## 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
##   (4069 observations deleted due to missingness)
## AIC: 1600.4
##
## Number of Fisher Scoring iterations: 5
```

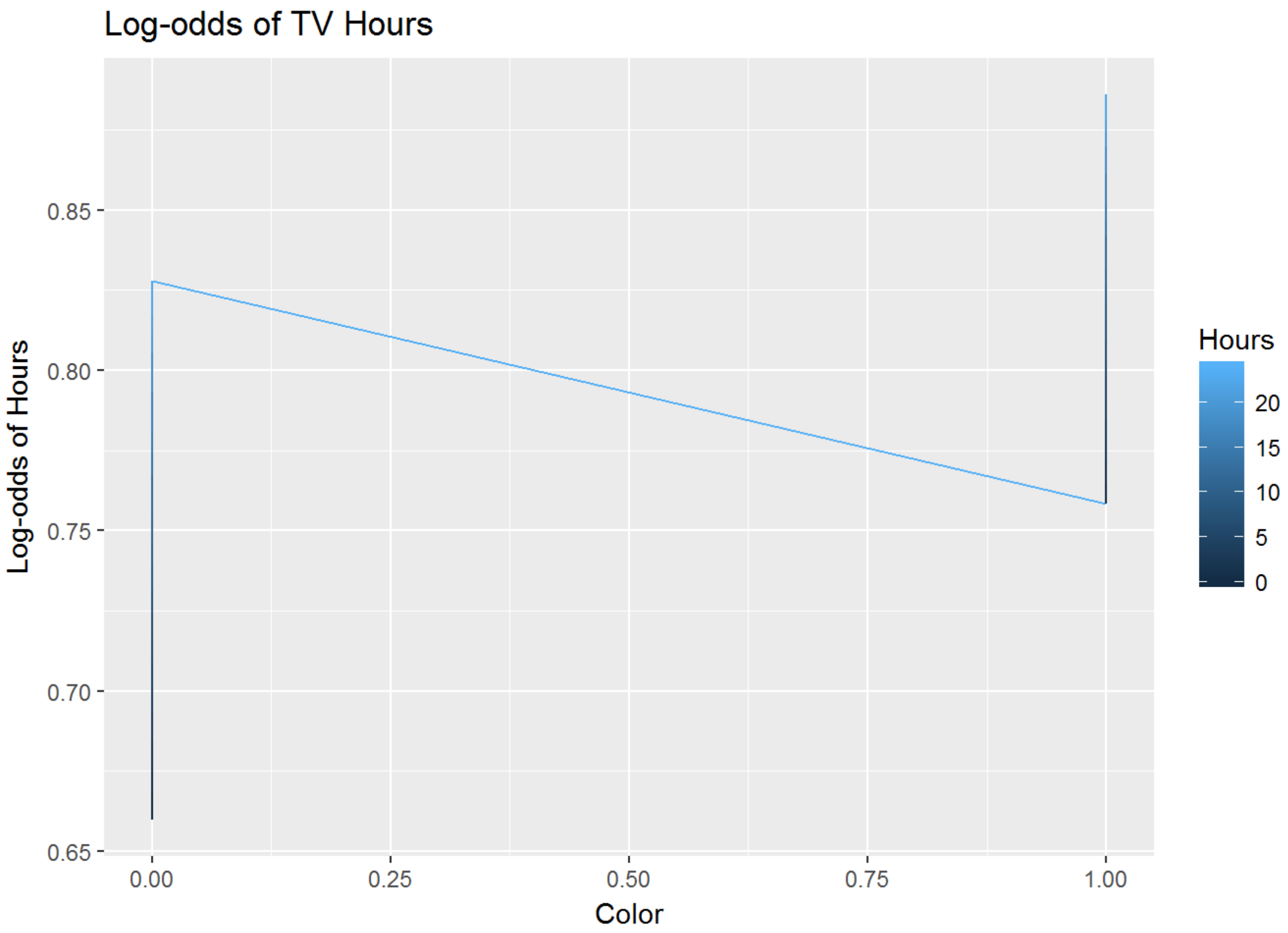
```
#Model with only the Significant variables.
tvlogModel2 <- glm(tvhours ~  hrsrelax + black + educ , data = dat2, family = poisson)

summary(tvlogModel2)
```

```
##
## Call:
## glm(formula = tvhours ~ hrsrelax + black + educ, family = poisson,
##     data = dat2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9673  -0.7848  -0.1372   0.4014   5.6439
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.256586   0.111310  11.289  < 2e-16 ***
## hrsrelax     0.037015   0.006243   5.929 3.05e-09 ***
## black        0.446314   0.046871   9.522  < 2e-16 ***
## educ        -0.042084   0.007550  -5.574 2.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1229.2  on 1005  degrees of freedom
## Residual deviance: 1062.5  on 1002  degrees of freedom
##   (3504 observations deleted due to missingness)
## AIC: 3647.8
##
## Number of Fisher Scoring iterations: 5
```

```
tvlogModel3 <- glm(tvhours ~  hrsrelax + black , data = dat2, family = poisson)

#Getting the Log Odds
dat2_hrs_black <- dat2 %>%
  data_grid(hrsrelax, black) %>%
  add_predictions(tvlogModel3) %>%
  mutate(pred = logit2prob(pred))
dat2_hrs_black
```

```
## # A tibble: 42 × 3
##    hrsrelax black      pred
##       <dbl> <dbl>     <dbl>
## 1         0     0 0.6597695
## 2         0     1 0.7582646
## 3         1     0 0.6682182
## 4         1     1 0.7651381
## 5         2     0 0.6765599
## 6         2     1 0.7718749
## 7         3     0 0.6847908
## 8         3     1 0.7784744
## 9         4     0 0.6929073
## 10        4     1 0.7849363
## # ... with 32 more rows
```

```
#Ploting the Interactive results
ggplot(dat2_hrs_black, aes(black, pred, color = hrsrelax)) +
  geom_line() +
  labs(title = "Log-odds of TV Hours",
       x = "Color",
       y = "Log-odds of Hours",
       color = "Hours")
```



```
#Finding the accuracy of our model.
hoursAccuracy <- add_predictions(dat2, tvlogModel3)

hoursAccuracy <- mutate(hoursAccuracy, pred = as.numeric(logit2prob(pred) > .5))

accRateTV <- mean(hoursAccuracy$tvhours == hoursAccuracy$pred, na.rm = TRUE)

proPRETV <- PRE(tvlogModel3)

aucValuesTV <- auc(hoursAccuracy$tvhours, hoursAccuracy$pred)
```

```
## Warning in roc.default(response, predictor, auc = TRUE, ...): 'response'
## has more than two levels. Consider setting 'levels' explicitly or using
## 'multiclass.roc' instead
```

```
accRateTV
```

```
## [1] 0.2408325
```

```
proPRETV
```

```
## [1] -0.008915305
```

```
aucValuesTV
```

```
## Area under the curve: 0.5
```

2.1 EStimate a regression Model (3)

2.1 Write out the three components of the GLM for your specific model of interest.

1. Random Component

Our random component is Y = Hours Watched TV and is familly is poisson.

2. Linear Predictor

$$\eta = \beta_0 + \beta_1 hrsrelax + \beta_2 black + \beta_3 educ$$

The linear predictor’s variables are: Categorical Variables: color (black) Continuos Variables: Education (edu) and Hours to relax (hrsrelax)

3. Link Function

$$log(\mu) = \eta_i$$

The function is Logit

2.2 Estimate the model and report your results.

```
summary(tvlogModel3)
```

```
##
## Call:
## glm(formula = tvhours ~ hrsrelax + black, family = poisson, data = dat2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0763  -0.8411  -0.1209   0.4710   5.5984
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.662267   0.033687  19.659  < 2e-16 ***
## hrsrelax    0.037870   0.006183   6.125 9.08e-10 ***
## black       0.480921   0.046358  10.374  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1242.9  on 1008  degrees of freedom
## Residual deviance: 1102.7  on 1006  degrees of freedom
##   (3501 observations deleted due to missingness)
## AIC: 3696
##
## Number of Fisher Scoring iterations: 5
```

2.3 Interpret the results in paragraph format After doing a summary of the model with all the variables in our dataset of the tv hours, we can see that the variable of color, education and hours of relaxation in the day status are statistically significant for trying to predict the tv hours spend by and individual. We can see that the accuracy of the models explain 24..08 of the model. By choosing only the variables that are significant to our test, we can see that we dont have a good model. When we create an interactive log norm of hours of relaxiation and color, we can see that the more hours you have of relaxiation and the your color is black, the more hours you will have for watching tv.