# PS6

*Yuqing Zhang*

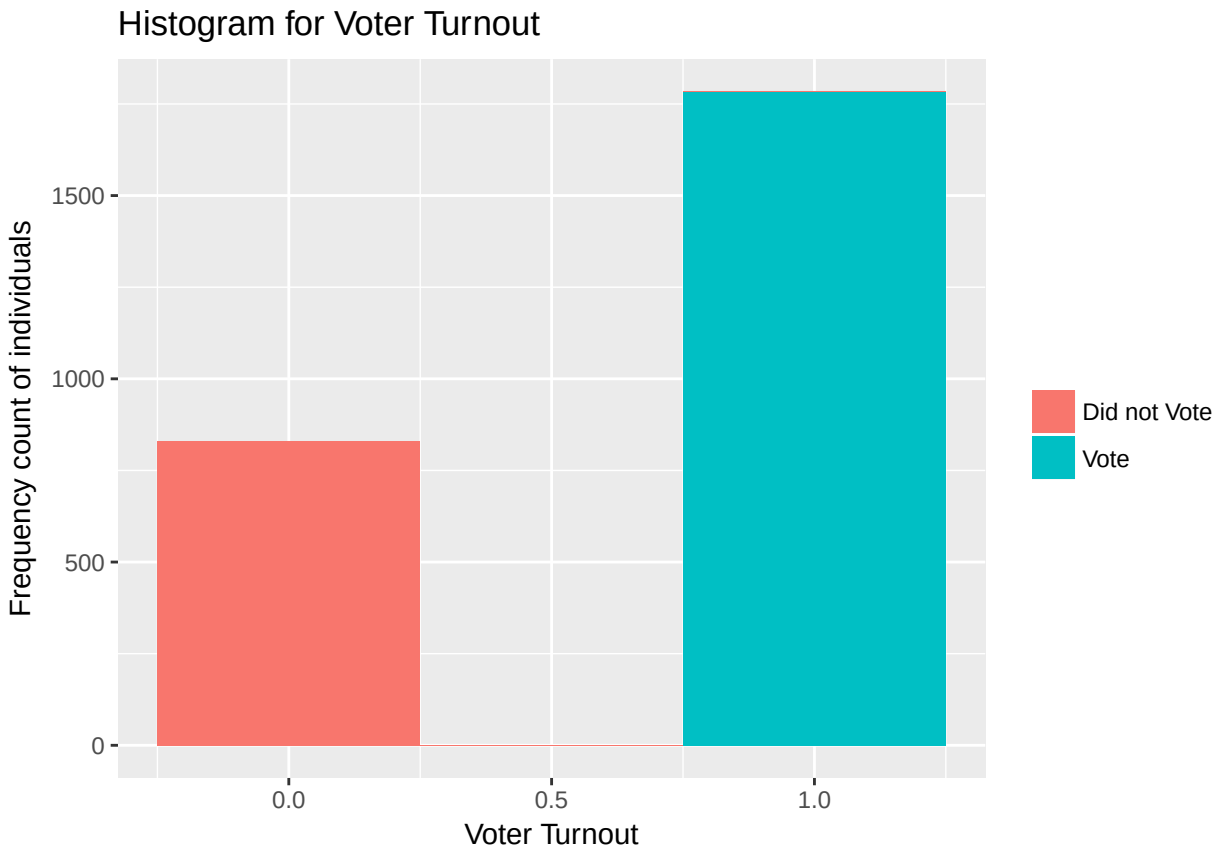
*2/18/2017*

## Modeling voter turnout

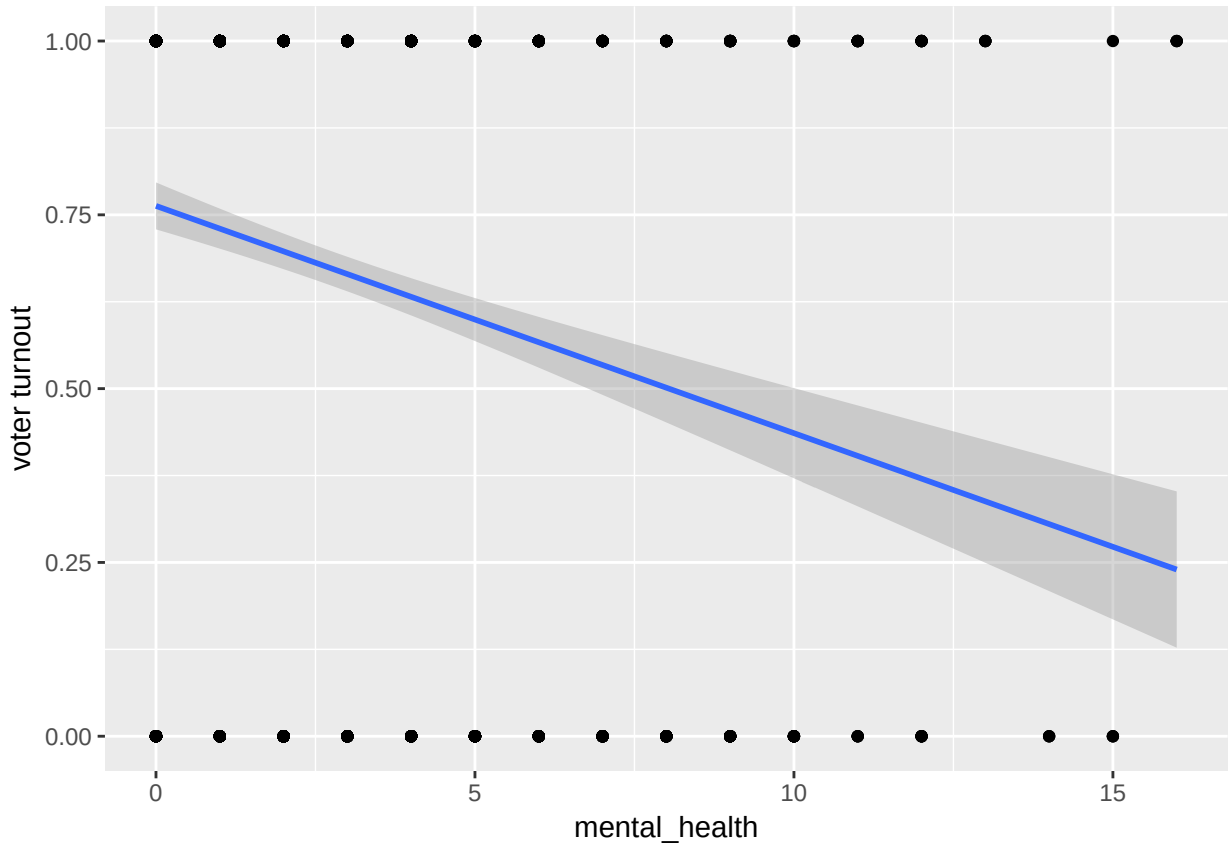
### Describe the data

```
mental_health<-read.csv('mental_health.csv')
```

### Including Plots




1. The unconditional probability of a given individual turning out to vote is: 63%




2. The graph tells us that the worse one person's mental condition is, the less likely he or she is going to vote. The problem with the linear line is that the only possible values for voter turnout are 0 and 1. Yet the linear regression model gives us predicted values such as .75 and .25.

## Basic Model

```
vote_mental <- glm(vote96 ~ mhealth_sum, data = mental_health, family = binomial)
summary(vote_mental)
```

```
##
## Call:
## glm(formula = vote96 ~ mhealth_sum, family = binomial, data = mental_health)
##
## 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
```

1.The relationship between mental health and voter turnout is statistically significant because p-value is almost 0. The coefficient is -.14348, which means increasing by 1 on the mental health scale, decreases the likelihood of voting by almost -86. It indicates the relationship is substantive.

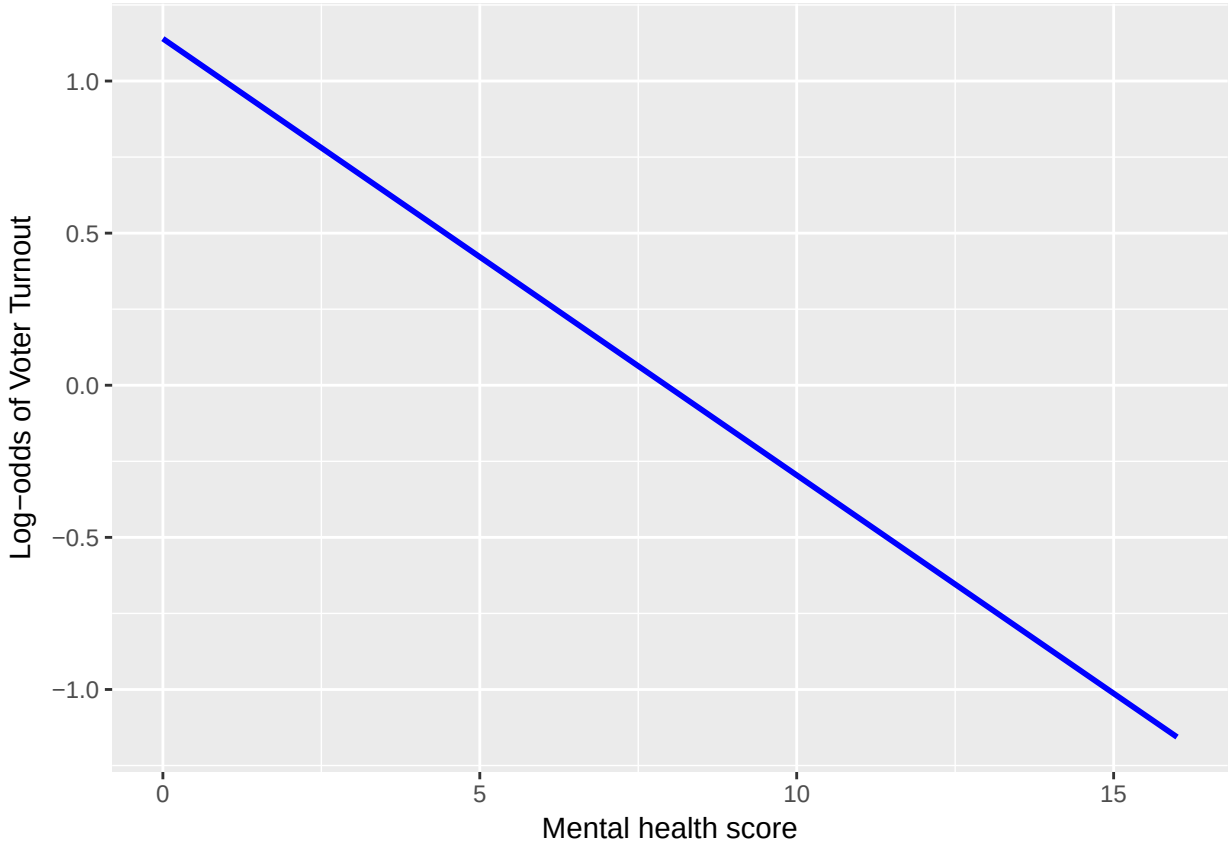
2.

```
tidy(vote_mental)
```

```
##          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
```

```
log_odds <- vote_mental$coefficients[2]
```

For every one-unit increase in mental_health score, we expect the log-odds of voter turnout to decrease by {r param}log_odds



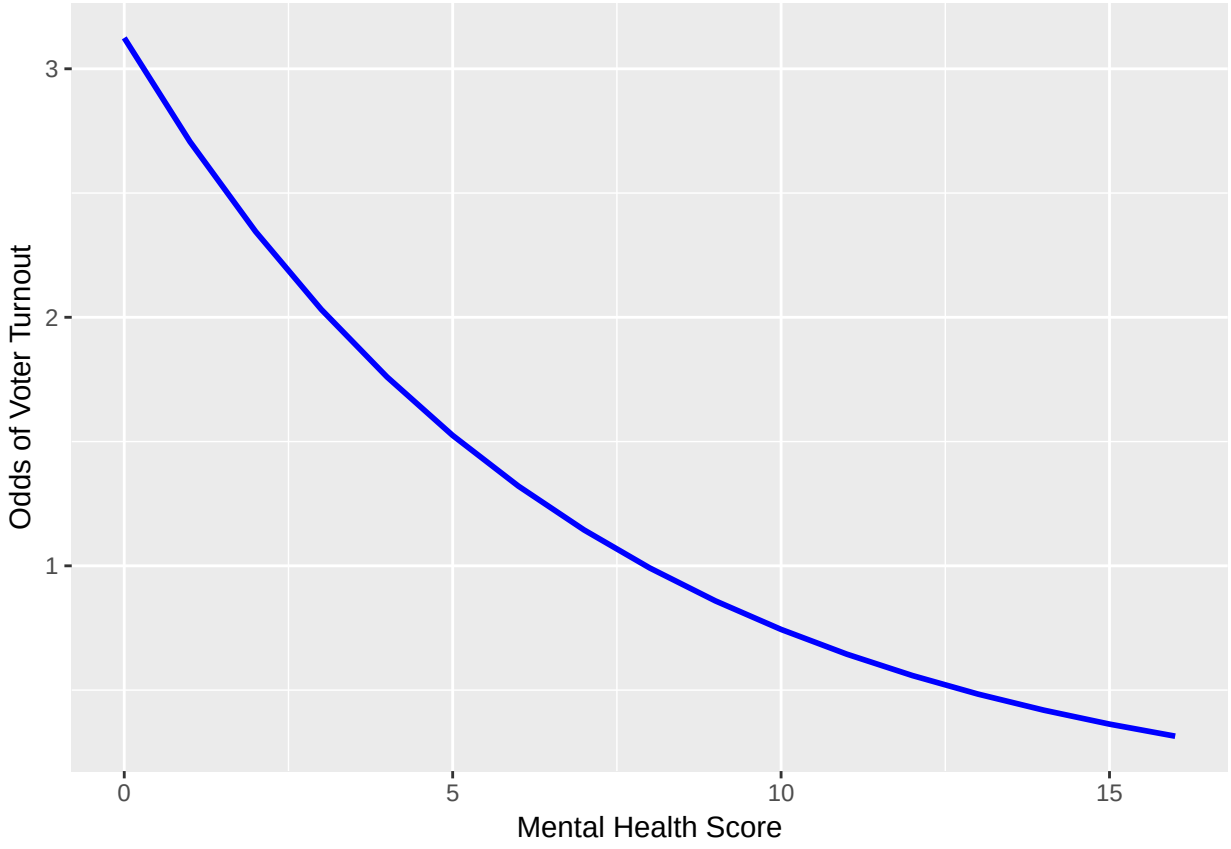

```
mental_health_score <- mental_health %>%
  add_predictions(vote_mental) %>%
  # predicted values are in the log-odds form - convert to probabilities
  mutate(prob = logit2prob(pred))
prob2odds <- function(x){
```

```
  x / (1 - x)
}
mental_health_score <- mental_health_score %>%
  mutate(odds = prob2odds(prob))
```

3.

```
exp_odds = exp(log_odds)
```

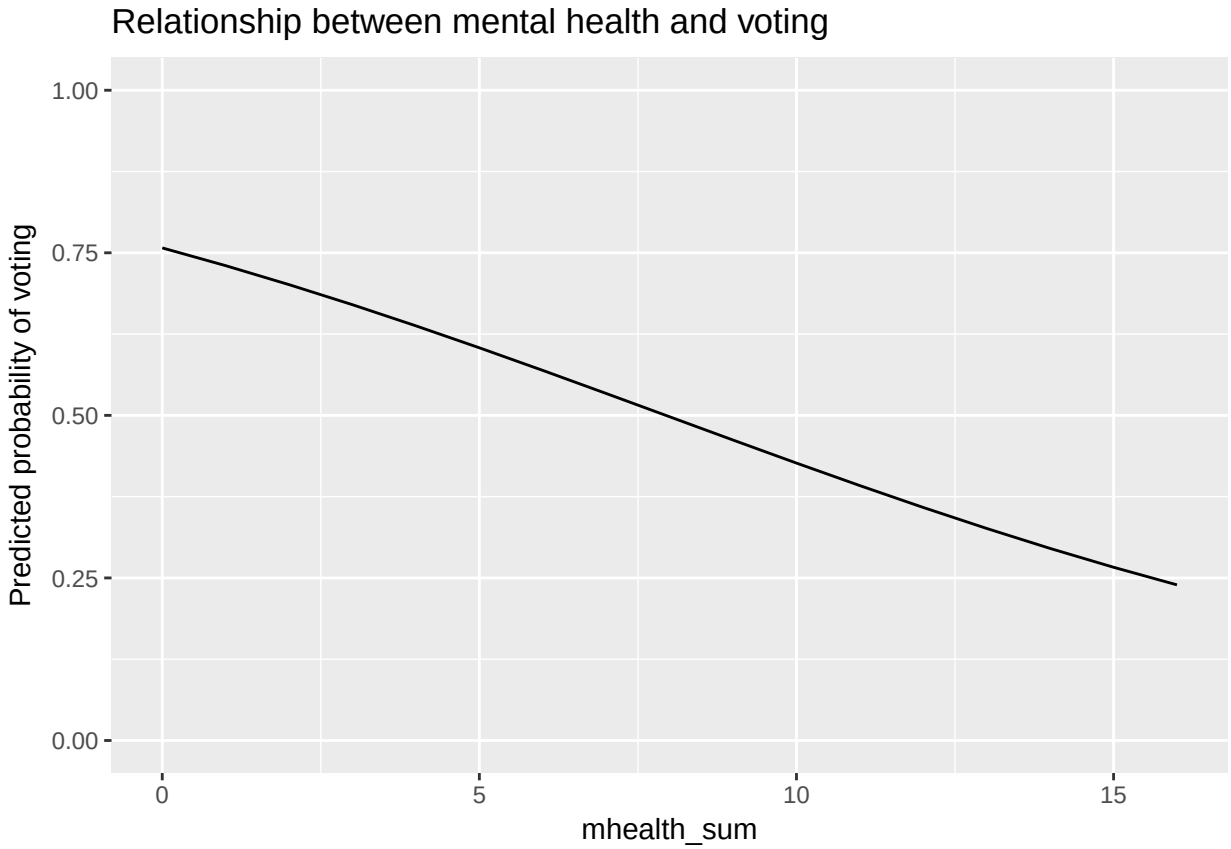
The odds ratio associated with a one unit increase in mhealth_sum is 0.8663423




4.a

```
prob = logit2prob(log_odds)
```

A one-unit increase in mental health index is associated with 0.4641926 decrease in probability of voting on average.




4.b

```
fd_1_2 = logit2prob(2) - logit2prob(1)
fd_5_6 = logit2prob(6) - logit2prob(5)
```

The first difference for an increase in the mental health index from 1 to 2 is: 0.1497385. The first difference for an increase in the mental health index from 5 to 6 is: 0.0042202.

5.

```
mh_accuracy <- mental_health %>%
  add_predictions(vote_mental) %>%
  mutate(pred = logit2prob(pred),
         pred = as.numeric(pred > .5))

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

```
# create a function to calculate the modal value of a vector
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
# 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 = PRE(vote_mental)
```

```
mh_accuracy <- mental_health %>%
  add_predictions(vote_mental) %>%
  mutate(pred = logit2prob(pred),
         prob = pred,
         pred = as.numeric(pred > .5))
auc_x <- auc(mh_accuracy$vote96, mh_accuracy$prob)
```

The accuracy rate is: 67.78% and the proportional reduction in error is: 1.62%. The AUC is 0.6243087.

# Multiple variable model

## 1. Three components

a) A random component specifying the conditional distribution of the response variable, Yi, given the values of the predictor variables in the model. Each individual vote turnout is either 0(not vote) or 1(vote), so each one is a bernoulli trial. Thus the response variable vote96, ,which is a collection of each individual vote turnout, is distributed as a binomial random variable.

b) The linear predictor is:

$$vote96_i = \beta_0 + \beta_1 mhealth\_sum_i + \beta_2 age_i + \beta_3 educ_i + \beta_4 black_i + \beta_5 female_i + \beta_6 married_i + \beta_7 inc10_i$$

c) The link function is

$$g(vote96_i) = \frac{e^{vote96_i}}{1 + e^{vote96_i}}$$

## 2,3 Estimate the model and report your results.

```
vote_all <- glm(vote96 ~ ., data = mental_health,
                          family = binomial)
vote_all_grid <- mental_health %>%
  #data_grid(.) %>%
  add_predictions(vote_all) %>%
  mutate(pred = logit2prob(pred))

summary(vote_all)
```

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

```
## 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
```

The results table shows that four predictors are statistically significant. Mental health score and coefficient is -0.089102, meaning that for every one-unit increase in mental_health score, we expect the log-odds of voter turnout to decrease by 0.089102; age and coefficient is 0.042534 for age, meaning that for every one-unit increase in mental_health score, we expect the log-odds of voter turnout to increase by0.042534; education and coefficient is 0.228686,meaning that for every one-unit increase in education, we expect the log-odds of voter turnout to increase by 0.228686; income and coefficient is 0.069614,meaning that for every one-unit increase in income, we expect the log-odds of voter turnout to increase by 0.069614.

```
mh_accuracy_all <- mental_health %>%
  add_predictions(vote_all) %>%
  mutate(pred = logit2prob(pred),
         prob = pred,
         pred = as.numeric(pred > .5))
accr_rate_all = mean(mh_accuracy_all$vote96 == mh_accuracy_all$pred, na.rm = TRUE)
auc_all <- auc(mh_accuracy_all$vote96, mh_accuracy_all$prob)
pre_all <- PRE(vote_all)
```

The accuracy rate,72.36%, proportional reduction in error (PRE),14.81% and area under the curve (AUC),0.759624 of the current model indicate that the model is better than the "simple" logistic regression model. Nonetheless, even with more predictors, the current logistic regression model shows a rather poor performance.

# Estimate a regression model

## 1. Three components

a) The response variable tvhours is distributed as a poisson random variable.

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

b) The linear predictor is:

$$tvhours_i = \beta_0 + \beta_1 age + \beta_2 childs + \beta_3 educ + \beta_4 female + \beta_5 grass + \beta_6 hrsrelax + \beta_7 black + \beta_8 social_connect + \beta_9 voted04 + \beta_{10}$$

c) The link function is

$$\mu_i = \ln(tvhours_i)$$

## 2,3 Estimate the model and report your results. In this model, the number of hours watching TV per day is the response variable. I chose to estimate these predictors: age, number of children, education, social_connect.

```
tv_consumption<-read.csv('gss2006.csv')
tv_pred <- glm(tvhours ~ age+childs+educ+hrsrelax, data = tv_consumption,
                         family = 'poisson')
tv_pred_grid <- tv_consumption %>%
  #data_grid(.) %>%
  add_predictions(tv_pred) %>%
  mutate(pred = logit2prob(pred))

summary(tv_pred)
```

```
##
## Call:
## glm(formula = tvhours ~ age + childs + educ + hrsrelax, family = "poisson",
##     data = tv_consumption)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0735  -0.8218  -0.1922   0.4025   6.5412
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.4014163  0.1195231  11.725  < 2e-16 ***
## age         -0.0003867  0.0016140  -0.240    0.811
## childs      -0.0010728  0.0144411  -0.074    0.941
## educ        -0.0462130  0.0072351  -6.387 1.69e-10 ***
## hrsrelax     0.0417919  0.0061889   6.753 1.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1328.9  on 1111  degrees of freedom
## Residual deviance: 1239.9  on 1107  degrees of freedom
##   (3398 observations deleted due to missingness)
## AIC: 4094.9
##
## Number of Fisher Scoring iterations: 5
```

At the p<.001 level,the results table shows that of the four predictors I chose, two of them are statistically significant. Education and coefficient is -0.0429390, meaning that for every one-unit increase in education, we expect the log-odds of hours of watching tv to decrease by 0.0429390; education and coefficient is 0.228686,meaning that for every one-unit increase in education, we expect the log-odds of voter turnout to increase by 0.228686; hours of relaxing and coefficient is 0.0417919,meaning that for every one-unit increase in social_connect, we expect the log-odds of hours of watching tv to increase by 0.0417919.

This model pretty makes sense. More educated one person is, less time they spend on watching tv. This may because of the fact that instead of 'wasting time on watching tv',they have more of other things to do.The relationship between hours of relax make sense too. As hours of relax increases, number of hours watching television also increases.

```
tv_accuracy <- tv_consumption %>%
  add_predictions(tv_pred) %>%
  mutate(pred = logit2prob(pred),
         prob = pred,
         pred = as.numeric(pred > .5))
accr_rate_tv = mean(tv_accuracy$tvhours == tv_accuracy$pred, na.rm = TRUE)
auc_tv <- auc(tv_accuracy$tvhours, tv_accuracy$prob)
pre_tv <- PRE(tv_pred)
```

The accuracy rate,25%, proportional reduction in error (PRE),-4.56% and area under the curve (AUC),0.483906 of the current model indicate that the model is not a very good model.

```
tv_pred_quasi <- glm(tvhours ~ age+childs+educ+hrsrelax, data = tv_consumption,
                         family = 'quasipoisson')
summary(tv_pred_quasi)
```

```
##
## Call:
## glm(formula = tvhours ~ age + childs + educ + hrsrelax, family = "quasipoisson",
##     data = tv_consumption)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0735  -0.8218  -0.1922   0.4025   6.5412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4014163  0.1351933  10.366  < 2e-16 ***
## age         -0.0003867  0.0018256  -0.212    0.832
## childs      -0.0010728  0.0163345  -0.066    0.948
## educ        -0.0462130  0.0081837  -5.647 2.07e-08 ***
## hrsrelax     0.0417919  0.0070003   5.970 3.19e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.2794)
##
##     Null deviance: 1328.9  on 1111  degrees of freedom
## Residual deviance: 1239.9  on 1107  degrees of freedom
##   (3398 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

I used quasipoisson model to see if the model is under or over-dispersion. From the table summary above, dispersion parameter for quasipoisson family is 1.2794,higher than 1, which indicates that the model is over-dispersed (the true variance of the distribution is greater than its mean).