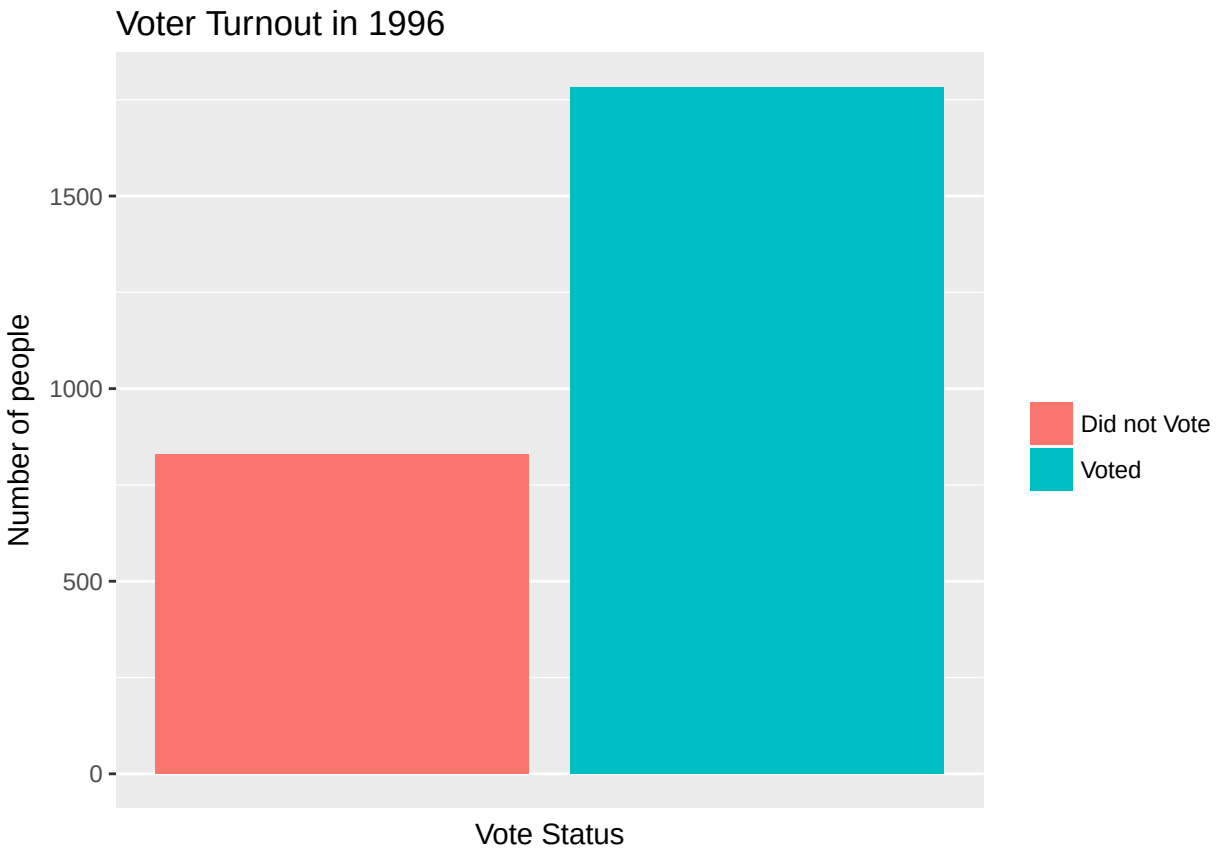
Problem Set 6 | MACS 301

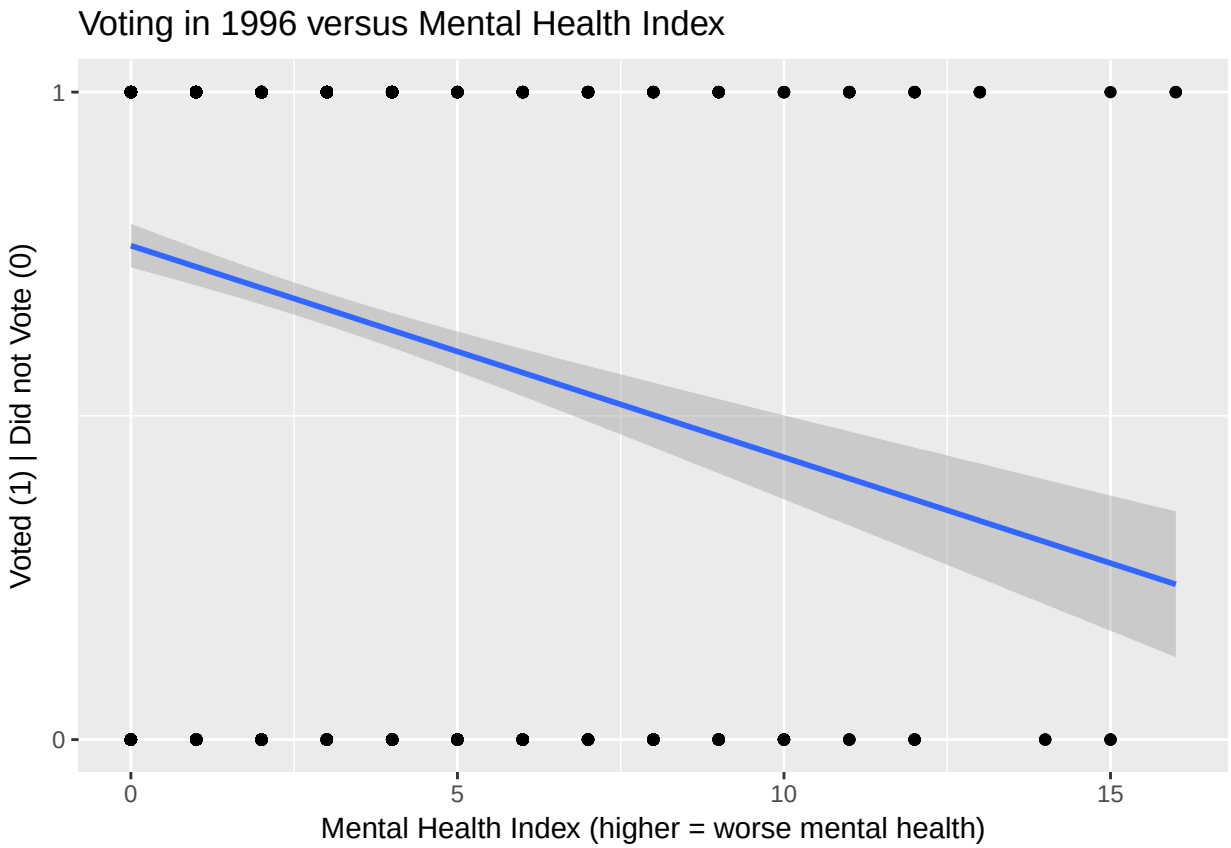
Julian McClellan

Due 2/20/17

Part 1: Modelling Voter Turnout

Describe the Data





The unconditional probability of a given individual turning out to vote is: 62.96%.

The scatterplot with the linear smoothing line tells us that in general, higher values of the mental health index, `mhealth_sum` (worse mental health) are associated with not voting in 1996. However, the problem with this graph is that a smooth fit line assumes the response variable can cover all real numbers. In our case, the response variable is either 1 (voted) or 0, not voted. Thus, interpretation of the line does make sense in the context of 'voting' or 'not voting'.

Basic model

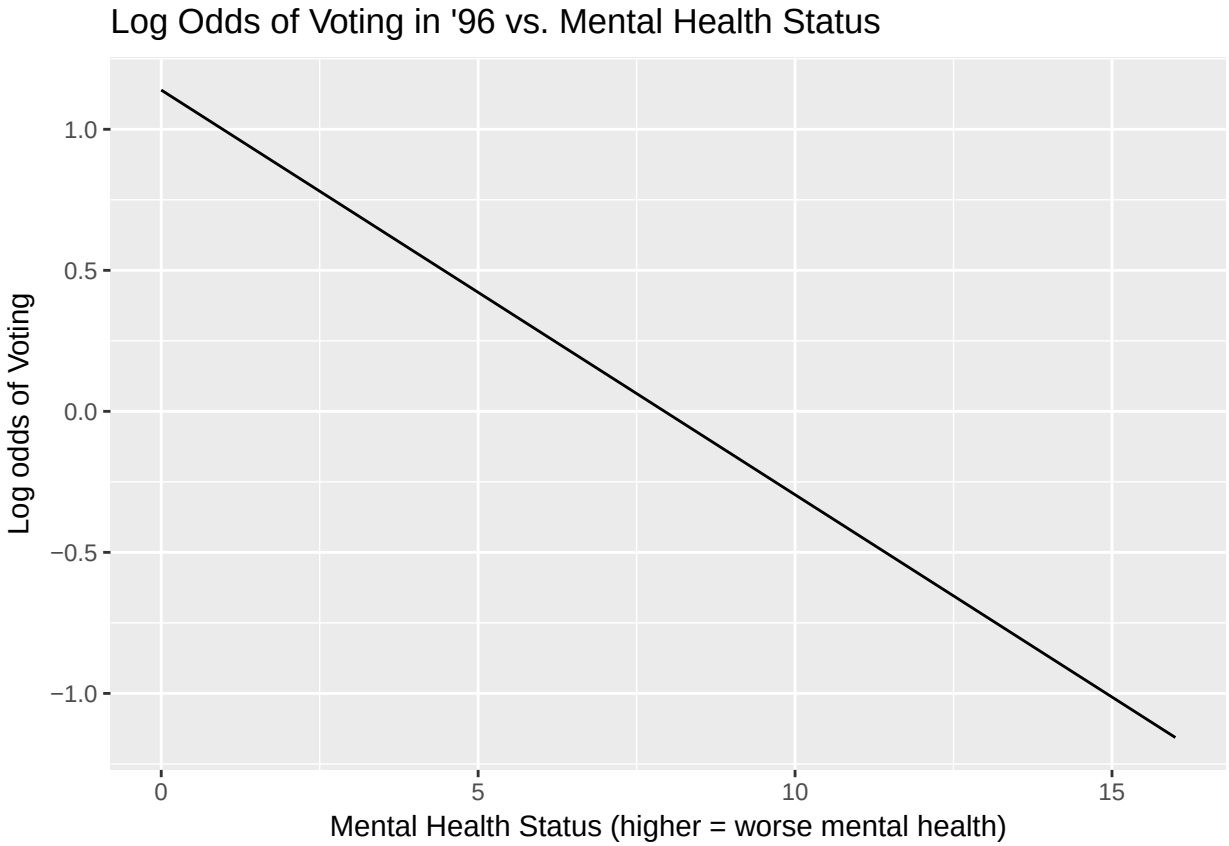
Table 1: Summary of Voting Status Regressed on Mental Health

	Dependent variable:
	vote96
mhealth_sum	−0.143***
	(0.020)
Constant	1.139***
	(0.084)
Observations	1,322
Log Likelihood	−808.360
Akaike Inf. Crit.	1,620.720
Note:	*p<0.1; **p<0.05; ***p<0.01

1.

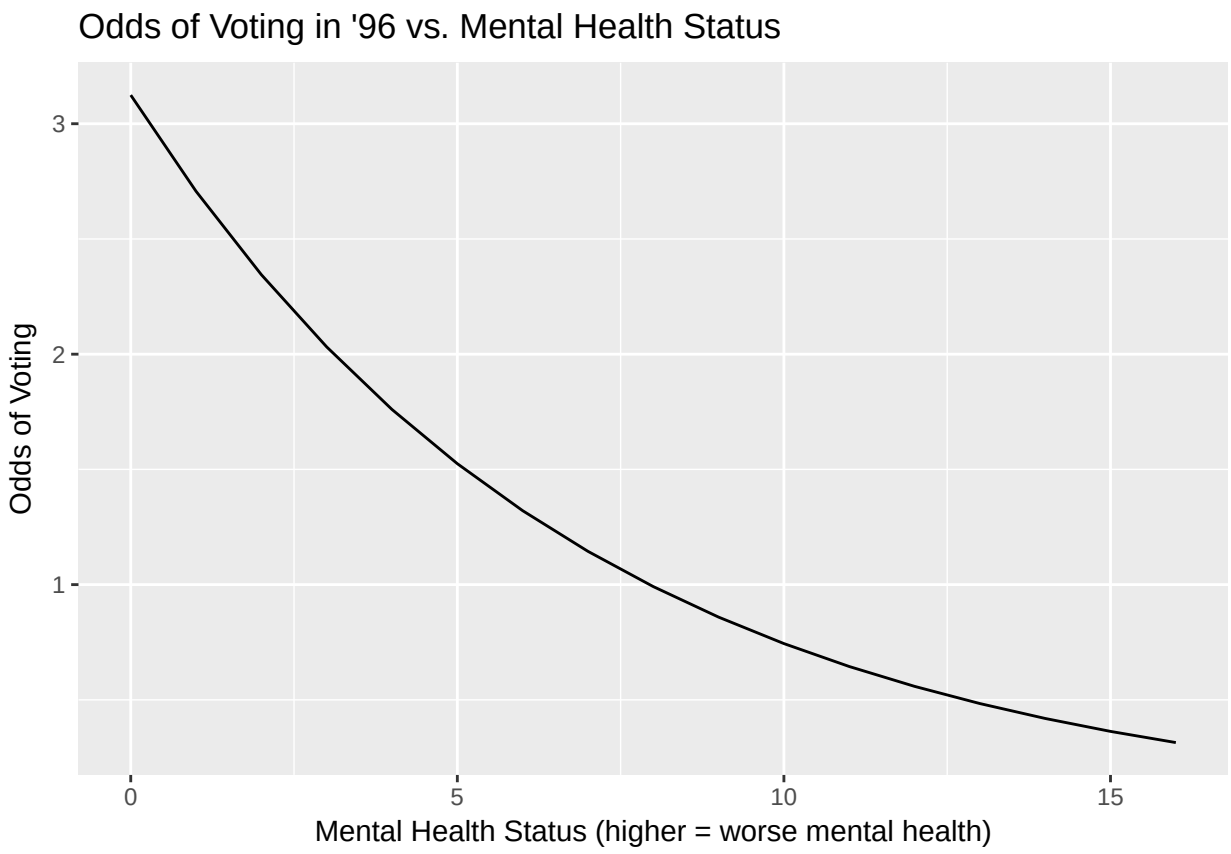
The relationship between mental health (`mhealth_sum`) and voter turnout is stastistically significant, with a p-value approaching 0 (3.1338829×10^{-13}). Additionally, we see that the change in the log-odds associated with a one unit increase in `mhealth_sum` (worse mental health) is -0.1434752. In other words, the odds ratio associated with a one unit increase in `mhealth_sum` is 0.8663423. This appears to be substantively significant as well in the negative direction. However, we will confirm this with the following graphs.

2.

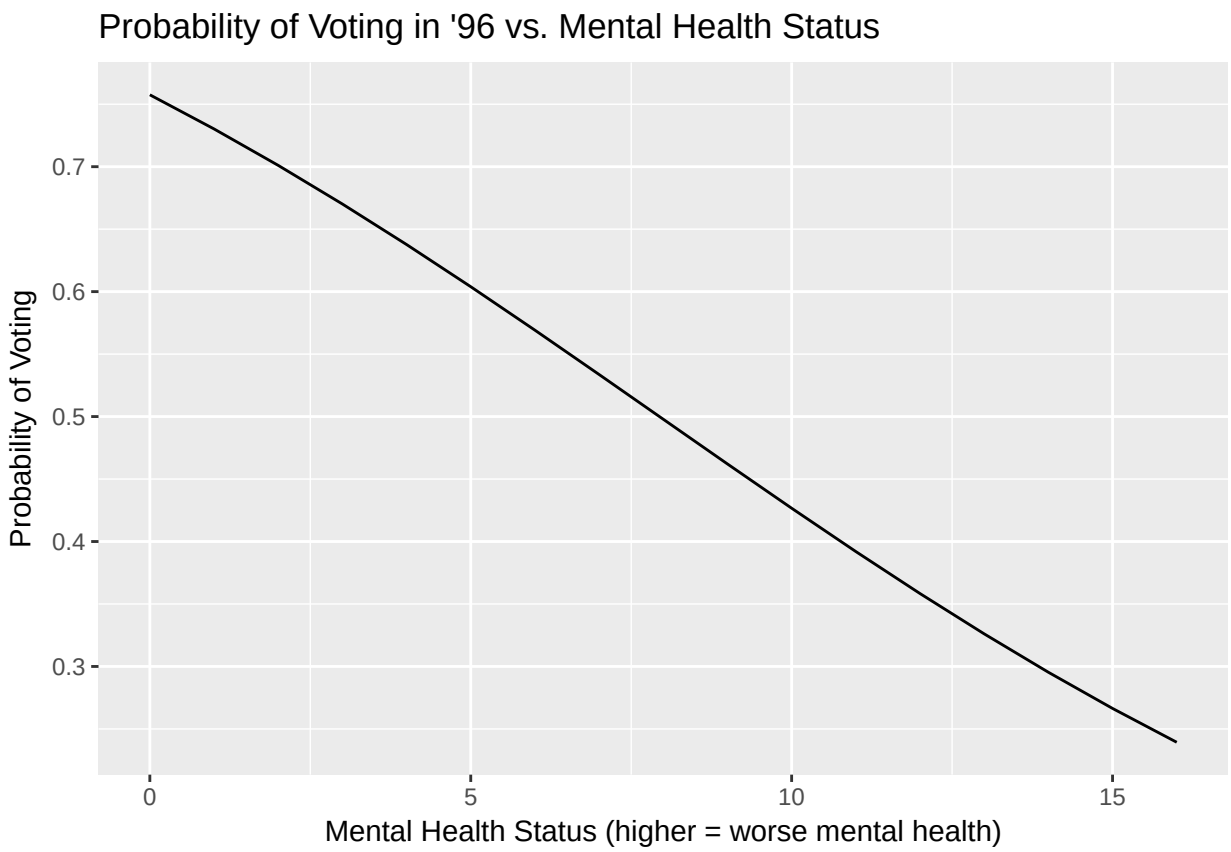


Looks linear, that's good. The estimated parameter, is given by default in terms of log odds, and was previously stated as: -0.1434752.

3.


Odds of Voting in '96 vs. Mental Health Status
Odds of Voting
3
2
1
0
5
10
15
Mental Health Status (higher = worse mental health)

4.



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

5.

Given a threshhold of .5, the accuracy rate is: 67.78% and the proportional reduction in error is: 1.62%. The AUC is 0.5400676, and the AUC score takes into account all possible threshhold values.

I don't think this is a very good model. The proportional reduction in error is a good indicator of this. A proportional reduction in error of `1.62%` is a pretty negligible increase over the useless classifier. Additionally, we see that the AUC score only provides a 0.0400676 increase in AUC score over the useless classifier.

Multiple Variable Model

1.

- The random component of the probability distribution, `vote96` is distributed as a binomial random variable. Each individual $vote96_i$ (each row of our dataframe) is a Bernoulli Trial and thus the sum of all individual $vote96_i$ $'s$ (i.e. the entire column `vote96`) is distributed as a binomial random variable.

$$Pr(\sum_{i=1}^{n} vote96_i = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- In our case, the linear predictor is:

$$vote96_i = \beta_0 + \beta_1 mhealth\_sum + \beta_2 age + \beta_3 educ + \beta_4 black + \beta_5 female + \beta_6 married + \beta_7 inc10$$

  - Note that this is the linear predictor for a model utilizing *all* possible explanatory variables. The model I utilize may or may not use all of the explanatory variables.

- Our link function is:

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

2.

```
# Define a full and a null model.
logit.mh_all <- glm(vote96 ~ ., data = df.mhealth,
                    family = binomial)
logit.mh_none <- glm(vote96 ~ 1, data = df.mhealth, family = binomial)

# We will use backward stepwise AIC selection to select a model
# In simple terms, AIC offers a tradeoff between model parsimony and log likelihood.
logit.mh_bselect <- stepAIC(logit.mh_all, trace = 0)
stargazer(logit.mh_bselect, type = 'latex', title = 'Results of Backwards AIC selected Model (Logit)', l
```

Table 2: Results of Backwards AIC selected Model (Logit)

	Dependent variable:
	vote96
mhealth_sum	-0.088^{***}
	(0.024)
age	0.042^{***}
	(0.005)
educ	0.225^{***}
	(0.029)
married	0.294^{*}
	(0.153)
inc10	0.066^{**}
	(0.026)
Constant	-4.200^{***}
	(0.498)
Observations	1,165
Log Likelihood	−621.808
Akaike Inf. Crit.	1,255.616
Note:	*p<0.1; **p<0.05; ***p<0.01

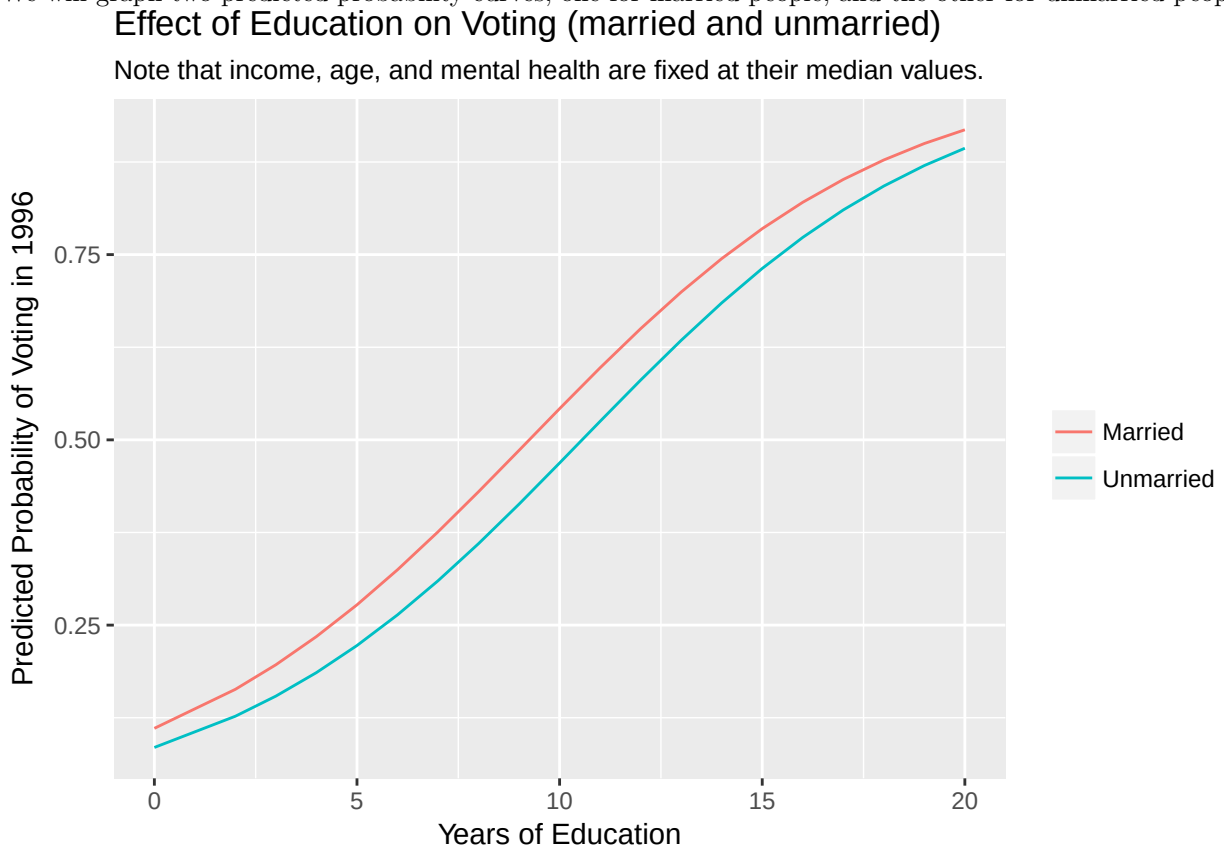
3.

From the above table, we see the backwards AIC selection has resulted in a model with 5 predictor variables, not including the intercept. The binary predictors `black` and `female` were left out. AIC selection involves

the calculation of maximum likelihood, and thus other predictors likely provided greater maximum likelihood than `black` and `female`. All of these predictors are significant at the `.05` level, except `married`, which is only significant at the `.1` level. With a .5 threshold, the accuracy rate of this model is 72.02%, the proportional reduction in error over the useless classifier is 13.76%, and the AUC (over all threshold values) is 0.758913.

All of the predictors, with the exception of `mhealth_sum`, have positive coefficients. These coefficients represent the effect a one unit increase of the predictor, with all other predictors held constant, has on the log-odds of vote participation. However, it is beneficial to visualize the effect of our predictors on actual predicted probabilities.

Let's focus on one of the more statistically and substantively significant predictors, `educ` (years of education) and look at its affect on predicted probability. Note, that we cannot simply graph predicted probability against `educ`, as there are other predictors to take into account. Thus, for the non-binary variables of `age`, `mhealth_sum`, and `inc10` we simply hold those values constant at their median values within the dataset. We will graph two predicted probability curves, one for married people, and the other for unmarried people.



Thus, we see the effect of `educ` the most impactful non-binary predictor on predicted voting probability, and `married` also results in a shift upward in predicted probability.

Part 2: Modeling TV Consumption

Estimate a Regression Model

1.

- The random component of the probability distribution, `tvhours` is distributed as a poisson random variable.

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

- The linear predictor of a model utilizing all of the possible explanatory variables in the GSS survey is:

$$tvhours_i = \beta_0 + \beta_1 age + \beta_2 childs + \beta_3 educ + \beta_4 female + \beta_5 grass + \beta_6 hrsrelax + \beta_7 black$$

$$+\beta_8 social_connect + \beta_9 voted04 + \beta_{10} xmovie + \beta_{11} zodiac + \beta_{12} dem + \beta_{13} rep + \beta_{14} ind$$

- Our link function is:

$$g(vote96_i) = \log(tvhours_i)$$

2.

```
# Define a full and a null model.
df.gss <- na.omit(df.gss)

pois.gss_all <- glm(tvhours ~ ., data = df.gss,
                    family = poisson)
pois.gss_none <- glm(tvhours ~ 1, data = df.gss, family = poisson)

# We will use backward stepwise AIC selection to select a model
# In simple terms, AIC offers a tradeoff between model parsimony and log likelihood.
pois.gss_bselect <- stepAIC(pois.gss_all, trace = 0)
stargazer(pois.gss_bselect, type = 'latex', title = 'Results of Backwards AIC selected Model (GLM | Pois
```

3.

Looking at the table below, we see the backwards AIC selection has resulted in a model with 4 predictor variables, not including the intercept. Thankfully, backwards AIC selection left out variables that wouldn't make much sense to affect TV hours watched, including the Zodiac symbols and whether or not someone saw an X-rated movie or voted in 2004. All of the predictors are significant at the `.05` level, except `grass`, which is only significant at the `.1` level. With a .5 threshold, the accuracy rate of this model is 22.68%, the proportional reduction in error over the useless classifier is 0%, and the AUC (over all threshold values) is 0.54875.

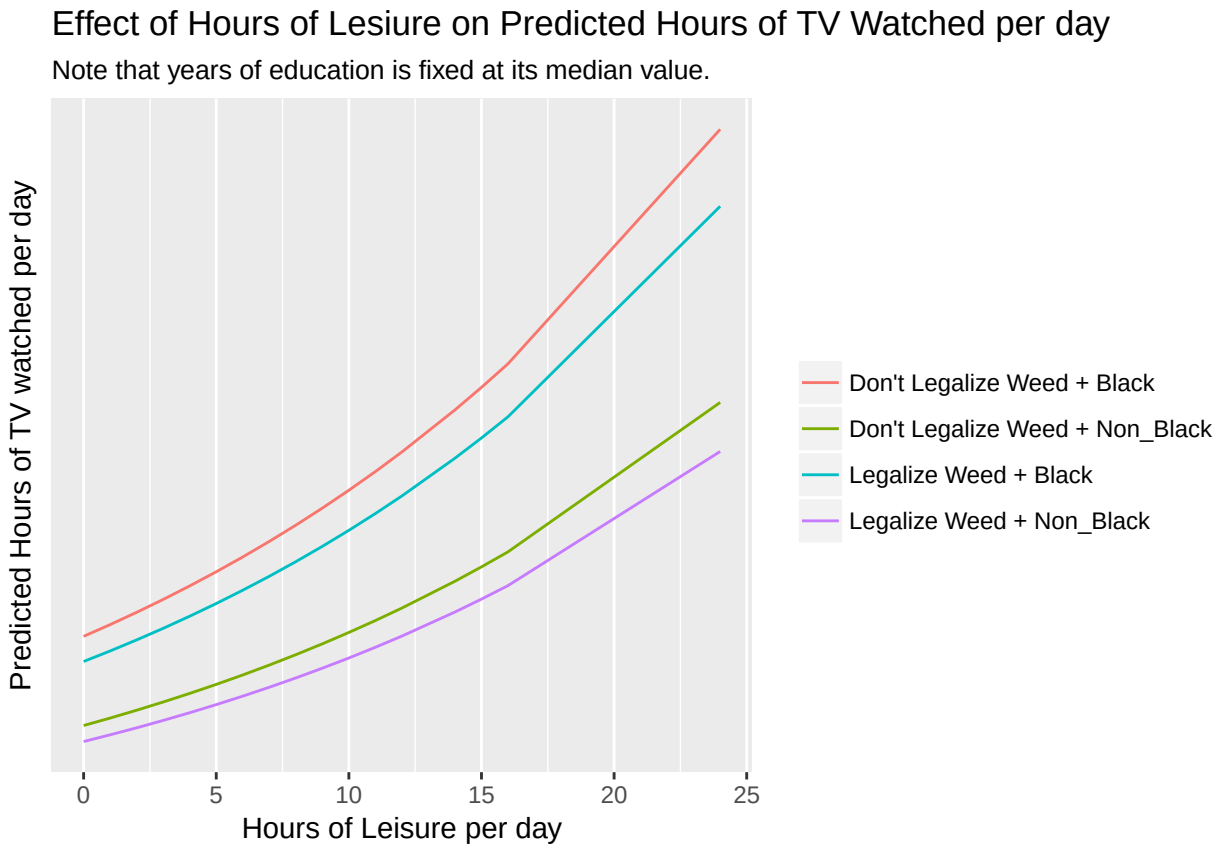
Two of the predictors, `educ` (years of education), and `grass` (believes marijuana should be legalized), have negative coefficients, and the other two, `hrsrelax` and `black` have positive coefficients. These coefficients represent the effect a one unit increase of the predictor, with all other predictors held constant, has on the log count of `tvhours`. However, as before it is beneficial to visualize the effect of our predictors on actual predicted probabilities.

Let's focus on one of the more substantively significant non-binary predictors: `hrsrelax` (the hours in a day one has to relax), and look at its affect on predicted count. Note, that we cannot simply graph predicted count against `hrsrelax`, as there are other predicted variables to take into account. Thus, for the non-binary

Table 3: Results of Backwards AIC selected Model (GLM | Poisson)

	Dependent variable:
	tvhours
educ	-0.039^{***}
	(0.011)
grass	-0.108^{*}
	(0.062)
hrsrelax	0.047^{***}
	(0.009)
black	0.451^{***}
	(0.072)
Constant	1.225^{***}
	(0.169)
Observations	441
Log Likelihood	−781.721
Akaike Inf. Crit.	1,573.442
Note:	*p<0.1; **p<0.05; ***p<0.01

variable `educ` we simply hold this value constant at the median values within the dataset. We will plot 4 different lines for the interactions of the two-binary variables `grass` and `black`.



As one could have seen from the table, but what is now evident from this table, is that if one is Black, and with all other predictors held constant, that there is a larger jump in predicted hours of TV watched per day than there is if one believes in legalizing marijuana. To be honest this seems to go against the stoner stereotype.