Final project

Peter Sullivan

Introduction

Alcohol abuse affects many different populations in the world, and it can ruin lives. Using the "Alcohol" data set from the Wooldridge package, I have created multiple models to help predict the likelihood of an individual abusing alcohol based on multiple predictors or covariates. The alcohol data set came with 33 variables, and 9822 observations. Here is a list of the 30 variables: abuse, status, unemrate, age, educ, married, famsize, white, exhealth, vghealth, goodhealth, fairhealth, northeast, midwest, south, centcity, outercity, qrt1, qrt2, qrt3, beertax, cigtax, ethanol, mothalc, fathalc, livealc, inwf, employ, agesq, beertaxsq, cigtaxsq, ethanolsq, educsq

At first glance I found it quite hard to decide on which variable to use in my model. Which variables have the strongest correlation to predicting whether someone will abuse alcohol? How do we choose? To start off, I choose variables for my models based on criteria that I believe would influence an individual to abuse alcohol. After I created those models, I determined how accurate those models were by using the R squared value for the LPM, and the AIC for the logit and Probit models. To create better models, I developed an automated process using the R^2 values from each variable in the model against the outcome variable (abuse). The process is outlined below. This paper has two objectives: 1. Create a best fit model that will help determine how likely an individual will abuse alcohol. 2. Create an automated process that identifies that the top N variables, and the best fit model to predict likelihood based on an outcome variable.

Method

For my initial models, I've decided that the variables that should influence the likelihood of an individual abusing alcohol are: status, age, education, fathalc, mothalc, beertax, and married. Some of these variables speak for them self. Fathalc and mothalc are used to determine whether the mother and fathers are alcoholics, 1 for yes and 0 for no. Status is used to identify if someone is out of the workforce, unemployed or employed using the 1, 2 and 3 respectively. Below are tables showing the distribution of each variable I have chosen.

Models

I've decided to use three types of models: LPM, Logit and Probit. I will create 5 nested models, and for ease of comparison, the LPM's, Logits and Probits covariates will all match for each nested model. For example lpm1, logit1, and probit1 all use status, age, education, mother alcholic, and father alcoholic as covariates.

Below are the models used for LPM, Logit, and Probits.

	age
	Min. :25.00
	1st Qu.:31.00
	Median :38.00
	Mean :39.18
	3rd Qu.:46.00
	Max. :59.00

	beertax
	Min. :0.045
	1st Qu.:0.145
	Median :0.259
	Mean :0.426
	3rd Qu.:0.446
	Max. :2.370

	educ
	Min. : 0.00
	1st Qu.:12.00
	Median :13.00
	Mean :13.31
	3rd Qu.:16.00
	Max. :19.00

	fathalc
	Min. :0.0000
	1st Qu.:0.0000
	Median :0.0000
	Mean :0.1543
	3rd Qu.:0.0000
	Max. :1.0000

	married
	Min. :0.0000
	1st Qu.:1.0000
	Median :1.0000
	Mean :0.8164
	3rd Qu.:1.0000
	Max. :1.0000

	status
	Min. :1.000
	1st Qu.:3.000
	Median :3.000
	Mean :2.829
	3rd Qu.:3.000
	Max. :3.000

	mothalc
	Min. :0.00000
	1st Qu.:0.00000
	Median :0.00000
	Mean :0.04042
	3rd Qu.:0.00000
	Max. :1.00000

```
#LPM
lpm1 <- lm(abuse ~ status+ age+ educ + fathalc, data = alcohol)
lpm2 <-lm(abuse ~ status+ age+ educ+ fathalc +mothalc, data = alcohol)
lpm3 <- lm(abuse ~ status+ age+ educ+ fathalc+ mothalc + beertax,
           data = alcohol)
lpm4 <- lm(abuse ~ status+ age+ educ+ fathalc+ mothalc + beertax+ married,
           data = alcohol)
lpm5 <- lm(abuse ~ status + age + educ + fathalc+ mothalc + beertax + married +
             fathalc:mothalc, data = alcohol)

# Logit
logit1 <- glm(abuse ~ status+ age+ educ+ fathalc,
              family = binomial(link = logit), data = alcohol)
logit2 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc,
              family = binomial(link = logit), data = alcohol)
logit3 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc + beertax,
              family = binomial(link = logit), data = alcohol)
logit4 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc + beertax+
                married, family = binomial(link = logit), data = alcohol)
logit5 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc + beertax+
                married+ fathalc:mothalc,
              family = binomial(link = logit), data = alcohol)

# Probit
probit1 <- glm(abuse ~ status+ age+ educ+ fathalc,
               family = binomial(link = probit), data = alcohol)
probit2 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc ,
               family = binomial(link = probit), data = alcohol)
probit3 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc +
                 beertax, family = binomial(link = probit), data = alcohol)
probit4 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc +
        beertax+ married, family = binomial(link = probit), data = alcohol)
probit5 <- glm(abuse ~ status+ age+ educ+ fathalc+ mothalc +
                           beertax+ married+ fathalc:mothalc,
               family = binomial(link = probit), data = alcohol)
```

Inital Observations

Below are the beta coefficients for the LPM, Logit and Probit models:

```
##
## ================================================================================
##                                          LPM's
##                  ------------------------------------------------------------
##                  Model 1      Model 2      Model 3      Model 4      Model 5
## --------------------------------------------------------------------------------
## (Intercept)        0.15 ***     0.15 ***     0.15 ***     0.16 ***     0.16 ***
##                   (0.03)       (0.03)       (0.03)       (0.03)       (0.03)
## status            -0.01        -0.01        -0.01        -0.01        -0.01
##                   (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## age                0.00         0.00         0.00         0.00         0.00
##                   (0.00)       (0.00)       (0.00)       (0.00)       (0.00)
## educ              -0.00 **     -0.00 **     -0.00 **     -0.00 **     -0.00 **
##                   (0.00)       (0.00)       (0.00)       (0.00)       (0.00)
## fathalc            0.05 ***     0.05 ***     0.05 ***     0.05 ***     0.05 ***
##                   (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## mothalc                         0.04 **      0.04 **      0.05 **      0.04 *
##                                 (0.02)       (0.02)       (0.02)       (0.02)
## beertax                                      -0.01        -0.01        -0.01
##                                              (0.01)       (0.01)       (0.01)
## married                                                   -0.03 ***    -0.03 ***
##                                                           (0.01)       (0.01)
## fathalc:mothalc                                                         0.01
##                                                                        (0.03)
## --------------------------------------------------------------------------------
## R^2                0.01         0.01         0.01         0.01         0.01
## Adj. R^2           0.01         0.01         0.01         0.01         0.01
## Num. obs.        9822         9822         9822         9822         9822
## ================================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
##
## ================================================================================
##                                        Logits
##                  ------------------------------------------------------------
##                  Model 1       Model 2       Model 3       Model 4       Model 5
## --------------------------------------------------------------------------------
## (Intercept)        -0.99 ***     -1.01 ***     -0.99 ***     -0.96 ***     -0.96 ***
##                    (0.14)        (0.14)        (0.14)        (0.14)        (0.14)
## status             -0.06         -0.06         -0.06         -0.04         -0.04
##                    (0.03)        (0.03)        (0.03)        (0.03)        (0.03)
## age                 0.00          0.00          0.00          0.00          0.00
##                    (0.00)        (0.00)        (0.00)        (0.00)        (0.00)
## educ               -0.02 **      -0.02 **      -0.02 **      -0.02 **      -0.02 **
##                    (0.01)        (0.01)        (0.01)        (0.01)        (0.01)
## fathalc             0.28 ***      0.26 ***      0.26 ***      0.26 ***      0.26 ***
##                    (0.04)        (0.04)        (0.04)        (0.04)        (0.05)
## mothalc                           0.21 **       0.21 **       0.22 **       0.22 *
##                                   (0.08)        (0.08)        (0.08)        (0.11)
## beertax                                         -0.03         -0.03         -0.03
##                                                 (0.04)        (0.04)        (0.04)
## married                                                       -0.18 ***     -0.18 ***
##                                                               (0.05)        (0.05)
## fathalc:mothalc                                                             -0.01
##                                                                             (0.16)
## --------------------------------------------------------------------------------
## AIC                6306.43       6301.83       6303.10       6290.39       6292.39
## BIC                6342.39       6344.98       6353.45       6347.93       6357.12
## Log Likelihood   -3148.21      -3144.91      -3144.55      -3137.20      -3137.19
## Deviance           6296.43       6289.83       6289.10       6274.39       6274.39
## Num. obs.          9822          9822          9822          9822          9822
## ================================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05

##
## ================================================================================
##                                        Probits
##                  ------------------------------------------------------------
##                  Model 1       Model 2       Model 3       Model 4       Model 5
## --------------------------------------------------------------------------------
## (Intercept)        -1.64 ***     -1.67 ***     -1.63 ***     -1.57 ***     -1.57 ***
##                    (0.27)        (0.27)        (0.27)        (0.27)        (0.27)
## status             -0.11         -0.10         -0.10         -0.08         -0.08
##                    (0.06)        (0.06)        (0.06)        (0.06)        (0.06)
## age                 0.00          0.00          0.00          0.01          0.01
##                    (0.00)        (0.00)        (0.00)        (0.00)        (0.00)
## educ               -0.03 **      -0.03 **      -0.03 **      -0.04 **      -0.04 **
##                    (0.01)        (0.01)        (0.01)        (0.01)        (0.01)
## fathalc             0.52 ***      0.49 ***      0.49 ***      0.49 ***      0.50 ***
##                    (0.08)        (0.08)        (0.08)        (0.08)        (0.09)
## mothalc                           0.39 **       0.39 **       0.41 **       0.43 *
##                                   (0.15)        (0.15)        (0.15)        (0.20)
## beertax                                         -0.07         -0.06         -0.06
##                                                 (0.08)        (0.08)        (0.08)
## married                                                       -0.34 ***     -0.34 ***
```

```
##                                                                    (0.09)        (0.09)
## fathalc:mothalc                                                                  -0.04
##                                                                                  (0.29)
## ---------------------------------------------------------------------------------------
## AIC                6306.83       6302.30       6303.56       6290.77       6292.76
## BIC                6342.79       6345.46       6353.90       6348.31       6357.49
## Log Likelihood    -3148.42      -3145.15      -3144.78      -3137.39      -3137.38
## Deviance           6296.83       6290.30       6289.56       6274.77       6274.76
## Num. obs.          9822          9822          9822          9822          9822
## =======================================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Initial Results

It seems like the R^2 value did not change no matter how many observations we included. We were only able to show about .01 variation of the model. The initial models are not explaining much variance in the model, so I am now going to attempt to automate the process and explain more variation in the model. We can't really tell too much from the logit and probit models, but we can keep track of the AIC, for when we compare these results to the new models I will develop.

Variables with High Correlation

The first step in the process to pick my variables for my linear models would be to identify the covariates with the highest R squared value when running a linear model against the outcome variable "abuse". Below are the process and results:

```
columns <- (colnames(alcohol))

outcome <- "abuse"
models <- lapply(paste(outcome," ~", columns), as.formula)
y <- NULL


for (model in models){
  linearmodel <- lm(model, data = alcohol)
  x <- summary(linearmodel)
  print(paste(format(model), "R^2 value: ", round(x$r.squared,3)*100, "%"))
  y <- rbind(y, data.frame(variable = as.character(model[3]),
              "Rvalue_Percent" = round(x$r.squared,3)*100))
}

new_data <- y[order(-y$Rvalue_Percent),]
new_data %>% slice(1:8)
top_8 <- new_data$variable[1:8]
```

The code above was used to run 32 different linear models with the abuse as the outcome variable. I then created a data frame in a loop that extracted the R^2 as a percentage. I then organized the table from highest to lowest and grabbed the top 8 variables.

The top 8 variables are below:

variable	Rvalue_Percent
famsize	0.5
fathalc	0.4
livealc	0.3
exhealth	0.2
ethanol	0.2
ethanolsq	0.2
educsq	0.2
status	0.1

The R values above are in percentage form. It should be noted that it seems the correlation between abuse and these variables list above is quite low. The highest correlation is family size and that is only .5 %.

Creating New Models

Now that I've identified the covariates with the highest correlation, the next step is to create models for my LPM, Logit, and Probit. Instead of rewriting the code for each model, I have created a framework that can also be applied to other data sets. Building onto the code that was used above to identify the top 8 variables, I then created 8 variables based on the results from above.

For the LM function and the GLM function to run, I first needed to create 5 variables (x1-x5). These variables need to be in the formula format, which was created using the lapply function. I then simply created 5 more models for the LPMS, Logits and Probits below, using the 5 variables I created. See the process below:

```
outcome_variable <- "abuse"

var1 <- top_8[1]
var2 <- top_8[2]
var3 <- top_8[3]
var4 <- top_8[4]
var5 <- top_8[5]
var6 <- top_8[6]
var7 <- top_8[7]
var8 <- top_8[8]


x1 <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",var4)
             , as.formula)
x2 <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",var4,
                   "+",var5), as.formula)
x3 <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",var4
                   ,"+",var5,"+",var6), as.formula)
x4 <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",var4
                   ,"+",var5,"+",var6,"+",var7), as.formula)
x5 <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",var4
                   ,"+",var5,"+",var6,"+",var7,"+",var8), as.formula)
```

```r
#LPMS
attach(alcohol)
lm1 <- lm(x1[[1]])
lm2 <- lm(x2[[1]])
lm3 <- lm(x3[[1]])
lm4 <- lm(x4[[1]])
lm5 <- lm(x5[[1]])


#Logits

logit1.1 <- glm(x1[[1]], family = binomial(link = logit))
logit1.2 <- glm(x2[[1]], family = binomial(link = logit))
logit1.3 <- glm(x3[[1]], family = binomial(link = logit))
logit1.4 <- glm(x4[[1]], family = binomial(link = logit))
logit1.5 <- glm(x5[[1]], family = binomial(link = logit))



#Probits

probit1.1 <- glm(x1[[1]], family = binomial(link = probit))
probit1.2 <- glm(x2[[1]], family = binomial(link = probit))
probit1.3 <- glm(x3[[1]], family = binomial(link = probit))
probit1.4 <- glm(x4[[1]], family = binomial(link = probit))
probit1.5 <- glm(x5[[1]], family = binomial(link = probit))

detach(alcohol)
```

Below are the tables showing our new LPM, Logit and Probit models, using the new variables with the highest R squared values.

```
##
## ============================================================================
##                                        LPM's
##              ------------------------------------------------------------
##              Model 1      Model 2      Model 3      Model 4      Model 5
## ----------------------------------------------------------------------------
## (Intercept)    0.14 ***     0.07 ***     0.05         0.07         0.09
##                (0.01)       (0.02)       (0.04)       (0.04)       (0.05)
## famsize       -0.01 ***    -0.01 ***    -0.01 ***    -0.01 ***    -0.01 ***
##                (0.00)       (0.00)       (0.00)       (0.00)       (0.00)
## fathalc        0.04 ***     0.05 ***     0.05 ***     0.05 ***     0.05 ***
##                (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## livealc        0.01         0.01         0.01         0.01         0.01
##                (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## exhealth      -0.03 ***    -0.03 ***    -0.03 ***    -0.02 ***    -0.02 ***
##                (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## ethanol                     0.03 ***     0.05         0.06         0.06
##                             (0.01)       (0.04)       (0.04)       (0.04)
## ethanolsq                                -0.00        -0.01        -0.01
##                                          (0.01)       (0.01)       (0.01)
## educsq                                                -0.00 ***    -0.00 ***
```

```
##                                                      (0.00)       (0.00)
## status                                                            -0.01
##                                                                   (0.01)
## ------------------------------------------------------------------------
## R^2             0.01         0.01         0.01         0.01         0.01
## Adj. R^2        0.01         0.01         0.01         0.01         0.01
## Num. obs.    9822         9822         9822         9822         9822
## ========================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
##
## ============================================================================
##                                        Logits
##                 ------------------------------------------------------------
##                 Model 1       Model 2       Model 3       Model 4       Model 5
## ----------------------------------------------------------------------------
## (Intercept)        -1.77 ***     -2.47 ***     -2.95 ***     -2.75 ***     -2.61 ***
##                    (0.08)        (0.20)        (0.51)        (0.51)        (0.53)
## famsize            -0.17 ***     -0.16 ***     -0.16 ***     -0.17 ***     -0.16 ***
##                    (0.02)        (0.02)        (0.02)        (0.02)        (0.02)
## fathalc             0.43 **       0.45 **       0.44 **       0.44 **       0.44 **
##                    (0.14)        (0.14)        (0.14)        (0.14)        (0.14)
## livealc             0.12          0.10          0.10          0.09          0.08
##                    (0.13)        (0.13)        (0.13)        (0.13)        (0.13)
## exhealth           -0.29 ***     -0.30 ***     -0.30 ***     -0.25 ***     -0.25 ***
##                    (0.07)        (0.07)        (0.07)        (0.07)        (0.07)
## ethanol                           0.34 ***      0.76          0.85 *        0.85 *
##                                   (0.09)        (0.43)        (0.43)        (0.43)
## ethanolsq                                       -0.09         -0.11         -0.11
##                                                 (0.09)        (0.09)        (0.09)
## educsq                                                        -0.00 ***     -0.00 ***
##                                                               (0.00)        (0.00)
## status                                                                      -0.05
##                                                                             (0.06)
## ----------------------------------------------------------------------------
## AIC              6253.86       6240.92       6241.87       6230.89       6232.08
## BIC              6289.82       6284.08       6292.22       6288.43       6296.81
## Log Likelihood  -3121.93      -3114.46      -3113.94      -3107.45      -3107.04
## Deviance         6243.86       6228.92       6227.87       6214.89       6214.08
## Num. obs.        9822          9822          9822          9822          9822
## ============================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
##
## ============================================================================
##                                        Probits
##                 ---------------------------------------------------------
##                 Model 1       Model 2       Model 3       Model 4       Model 5
## ----------------------------------------------------------------------------
## (Intercept)       -1.07 ***     -1.44 ***     -1.67 ***     -1.57 ***     -1.49 ***
##                   (0.04)        (0.10)        (0.26)        (0.26)        (0.28)
## famsize           -0.08 ***     -0.08 ***     -0.08 ***     -0.08 ***     -0.08 ***
##                   (0.01)        (0.01)        (0.01)        (0.01)        (0.01)
## fathalc            0.22 **       0.23 **       0.23 **       0.23 **       0.23 **
##                   (0.07)        (0.07)        (0.07)        (0.07)        (0.07)
## livealc            0.07          0.06          0.06          0.05          0.05
##                   (0.07)        (0.07)        (0.07)        (0.07)        (0.07)
## exhealth          -0.15 ***     -0.15 ***     -0.16 ***     -0.13 ***     -0.13 ***
##                   (0.04)        (0.04)        (0.04)        (0.04)        (0.04)
## ethanol                          0.18 ***      0.39          0.44 *        0.44 *
##                                  (0.05)        (0.22)        (0.22)        (0.22)
## ethanolsq                                      -0.05         -0.06         -0.06
##                                                (0.05)        (0.05)        (0.05)
## educsq                                                       -0.00 ***     -0.00 ***
##                                                              (0.00)        (0.00)
## status                                                                     -0.03
##                                                                            (0.03)
## ----------------------------------------------------------------------------
## AIC              6254.48       6241.18       6242.18       6231.09       6232.06
## BIC              6290.44       6284.33       6292.53       6288.63       6296.79
## Log Likelihood  -3122.24      -3114.59      -3114.09      -3107.55      -3107.03
## Deviance         6244.48       6229.18       6228.18       6215.09       6214.06
## Num. obs.        9822          9822          9822          9822          9822
## ============================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Results

It looks like the LPM R^2 values are still at .01. I believe this is due to the originally values being very small and .01 is the smallest value we can visualize on screen reg. After we identify the LPM model to use, I will extract the actual R^2 value and compare those to the actuals in the orginal LPM's to see how much variance we reduced by using the new models. I will perform a similar method for the Logit and Probits, but with comparing the AIC and BIC values. Now lets identify the best fitting models.

Determining the best Fit models

LPM's

To identify the best fitting LPM model, I will use the LinearHypothesis function and compare the F values determine which models are the best fit, and whether they are statistically significant. To automate this process, I first created vectors with the corresponding variables in each model, which can be seen in m1 through m5. When using the Linear Hypothesis function, we also need to identify the difference in variables when comparing those models. For example, if we are comparing m5 to m1, then we would need the variables that model 5 and model 1 don't share. We would need to identify the difference in variables for each model. To automate this process, I am using the function setdiff. Setdiff allows me to quickly identify the difference between vectors of strings. Once I identified the difference, I know can run the LinearHypothesis function. Below is the code used to perform these tasks and the results of each LinearHypothesis:

```
#Hypothesis Tests

m1 <- c(top_8[1:4])
m2 <- c(top_8[1:5])
m3 <- c(top_8[1:6])
m4 <- c(top_8[1:7])
m5 <- c(top_8[1:8])


m5m1 <- setdiff(m5,m1)
m5m2 <- setdiff(m5,m2)
m5m3 <- setdiff(m5,m3)
m5m4 <- setdiff(m5,m4)
m4m3 <- setdiff(m4,m3)


linearHypothesis(lm5,m5m1)
```

```
## Linear hypothesis test
##
## Hypothesis:
## ethanol = 0
## ethanolsq = 0
## educsq = 0
## status = 0
##
## Model 1: restricted model
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   9817 867.84
## 2   9813 865.21  4    2.6262 7.4463 5.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(lm5,m5m2)
```

```
## Linear hypothesis test
```

```
##
## Hypothesis:
## ethanolsq = 0
## educsq = 0
## status = 0
##
## Model 1: restricted model
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   9816 866.44
## 2   9813 865.21  3    1.2246 4.6296 0.003073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(lm5,m5m3)
```

```
## Linear hypothesis test
##
## Hypothesis:
## educsq = 0
## status = 0
##
## Model 1: restricted model
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   9815 866.42
## 2   9813 865.21  2    1.2026 6.8196 0.001097 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(lm5,m5m4)
```

```
## Linear hypothesis test
##
## Hypothesis:
## status = 0
##
## Model 1: restricted model
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   9814 865.32
## 2   9813 865.21  1   0.10315 1.1699 0.2794
```

```
linearHypothesis(lm4,m4m3)
```

```
## Linear hypothesis test
```

```
##
## Hypothesis:
## educsq = 0
##
## Model 1: restricted model
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   9815 866.42
## 2   9814 865.32  1    1.0994 12.469 0.0004156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1: Vifs for Each Variable

famsize	1.010065
fathalc	2.493831
livealc	2.501981
exhealth	1.042810
ethanol	24.059350
ethanolsq	24.032920
educsq	1.055115

LPM Results

Above are the LinearHypothesis results from comparing the LPM models. From the Linear Hypthoesis tests, I have determined that model 4 is more parsimonious than model 5 (F = 1.17, p >.05). Now that I have identified the Model 4 as the LPM model, I will now perform a series of tests to determine if there is any non-linearity in the model and if we can trust the covariates in the model for predictions.

After running the vif test, I have identified that Ethanol and Ethanolsq both have vifs over 24, therefore we will drop them from the model. Now we will run the reset test in order to test for linearity in the model.

```
##
##  RESET test
##
## data:  lm4
## RESET = 4.1787, df1 = 2, df2 = 9812, p-value = 0.01534
```

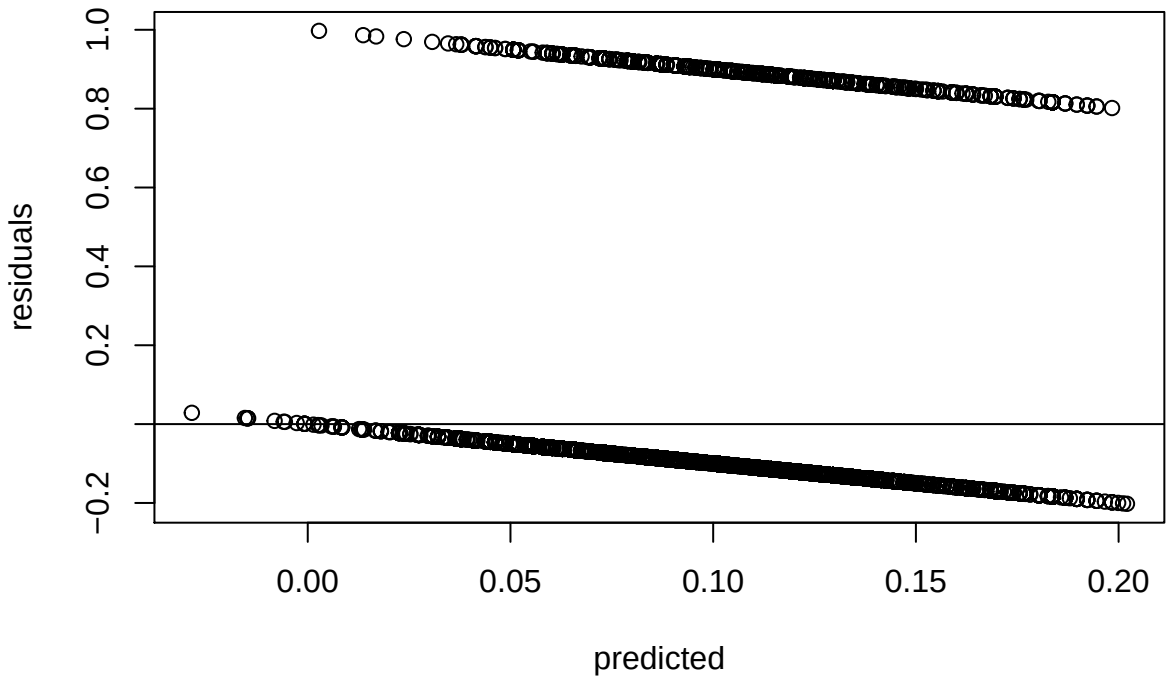
The p-value from the reset value test is below .05, therefor there is non-linearity in the model. Let's try dropping the ethanol and ethanolsq covariates and check the linearity again.

```
x4_new <- lapply(paste(outcome_variable, " ~", var1, "+", var2, "+", var3,"+",
                       var4,"+",var7), as.formula)

attach(alcohol)
lm4_new <- lm(x4_new[[1]])
detach(alcohol)

resettest(lm4_new)
```

```
##
##  RESET test
##
## data:  lm4_new
## RESET = 3.1639, df1 = 2, df2 = 9814, p-value = 0.0423
```

The Pvalue is still below .05, signifying non-linearity. There is still non linearity, but the its closer to the .05 mark. Lets check the residuals and outliers for non linearity as well.

```
predicted <- lm4_new$fitted.values
residuals <- lm4_new$residuals
plot(predicted,residuals)+ abline(h = 0)
```

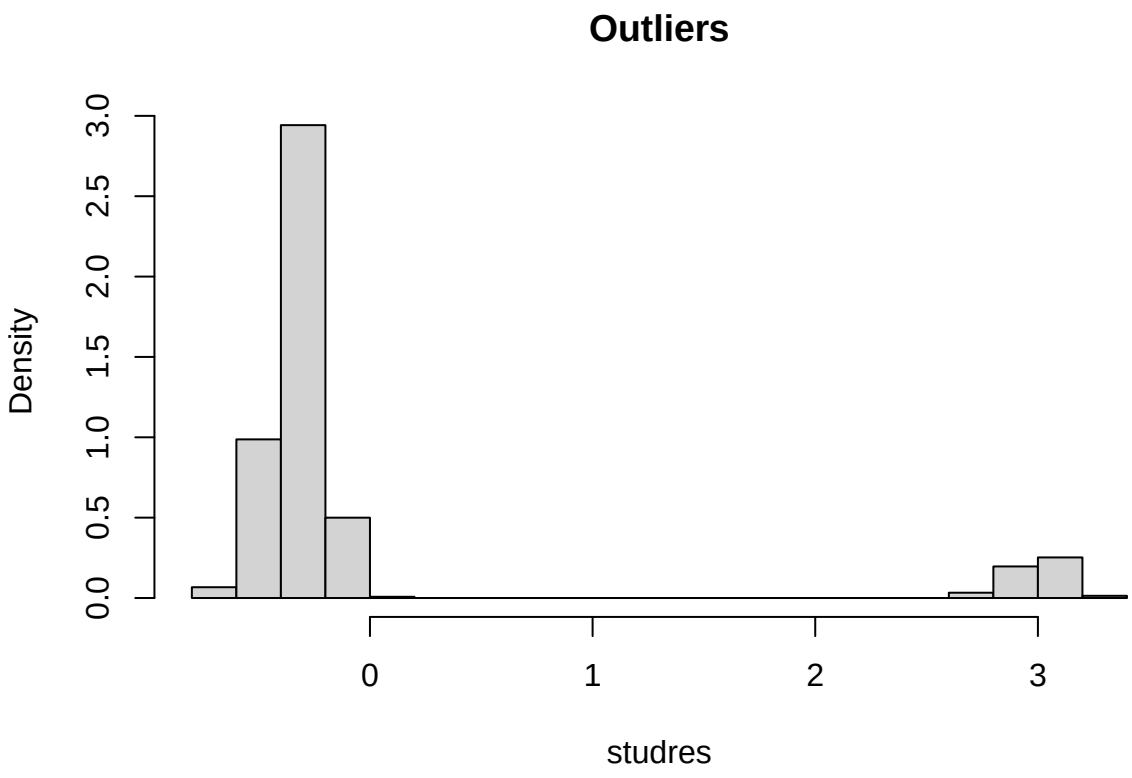



```
## integer(0)
```

This looks like nonlinearity to me. Let's check into the outliers using the rstudent function.

Table 2: Outlier Distribution Min and Max

max	min
-0.68	3.36



When looking at the histogram, there are outliers slightly to the right of 3. The table also confirms slight outliers with a max standard deviation of 3.36. The outlier is not too large, so I am not going to delete/omit any data.

Now lets check the Homoskedasticity assumption using the BP test function.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm4_new
## BP = 115.15, df = 5, p-value < 2.2e-16
```

The Model is significant with a p-value $< .05$ and the model does violate the homoscedasticity assumption. We have heteroskedasticity, and we need to estimate the robust errors. Below are the errors using the coeftest function.

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.16323181  0.01031484 15.8249 < 2.2e-16 ***
## famsize     -0.01408535  0.00197431 -7.1343 1.042e-12 ***
## fathalc      0.04453690  0.01309825  3.4002 0.0006760 ***
## livealc      0.00959970  0.01212474  0.7917 0.4285287
## exhealth    -0.02133789  0.00621045 -3.4358 0.0005932 ***
## educsq      -0.00013612  0.00004134 -3.2928 0.0009955 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We still have the current issue of non-linearity in our data. Through trial and error, I have identified that one of our variables needed to be transformed. To figure this out, I changed each variable, first squaring then logging them. With that variable transformed, I then used the reset test to check for linearity. After multiple trials I was able to identity that edusq needed to be logged.

```
lm4_linear <- lm(abuse~ famsize + fathalc + livealc+ exhealth + I(log(educsq+1)), data = alcohol)

resettest(lm4_linear)
```

```
##
##  RESET test
##
## data:  lm4_linear
## RESET = 2.473, df1 = 2, df2 = 9814, p-value = 0.08438
```

After running the reset test using the logged variable, we know have a p-value $>.05$, which means our model has linearity. Please see the results above.

Next we will identify the Logit and Probit Models

Now that we have identified the best fitting LPM Model, we know need to determine what is the best fitting model for the logit and probit models. In order to perform this task we will use the likelihood ratio tests.

```
lrtest(logit1.1,logit1.2)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   5 -3121.9
## 2   6 -3114.5  1 14.94   0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(logit1.2,logit1.3)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -3114.5
## 2   7 -3113.9  1 1.0528     0.3049
```

```
lrtest(logit1.3, logit1.4)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -3113.9
## 2   8 -3107.4  1 12.976  0.0003156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(logit1.4,logit1.5)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -3107.4
## 2   9 -3107.0  1 0.8124     0.3674
```

For The logit models, 2 is a better fit than 3 with (chisq = 1.0528, pr(>chisq >.05)). Model 4 is a better fit than model 5 with (chisq =.8124, pr(>chisq >.05)). I will choose logit1.4 as the logit model.

```
lrtest(probit1.1,probit1.2)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -3122.2
## 2   6 -3114.6  1 15.299  9.177e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(probit1.2,probit1.3)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -3114.6
## 2   7 -3114.1  1 0.9922     0.3192
```

```
lrtest(probit1.3,probit1.4)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -3114.1
## 2   8 -3107.6  1 13.091  0.0002966 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(probit1.4,probit1.5)
```

```
## Likelihood ratio test
##
## Model 1: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq
## Model 2: abuse ~ famsize + fathalc + livealc + exhealth + ethanol + ethanolsq +
##     educsq + status
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -3107.6
## 2   9 -3107.0  1 1.0317     0.3098
```

The probit models have similar results when compared to the logit models. For The probit models, 2 is a better fit than 3 (chisq = .992, pr(>chisq >.05)). Model 4 is a better fit than (chisq =1.0317, pr(>chisq >.05)). I will choose probit1.4 as the probit model.

Fixing up logit models

Now that we have our logit and Probit models, logit1.4 and Probit1.4. We also need to check the vifs like did in the LPM models. Her are the results below:

We are seeing similar results for what we identified in the LPM model. We need to get rid of the ethanol and ethanolsq covariates from the logit and probit models. I will also include the logged variable from the LPM from our earlier research.

Table 3: Logit Vifs

famsize	1.009437
fathalc	2.776656
livealc	2.782741
exhealth	1.038261
ethanol	24.005089
ethanolsq	23.982810
educsq	1.050736

Table 4: Probit Vifs

famsize	1.009902
fathalc	2.669483
livealc	2.676435
exhealth	1.039080
ethanol	24.036773
ethanolsq	24.012263
educsq	1.052182

Creating new logit and Probit Models

```
logit1.4_new <- glm(abuse~ famsize + fathalc+ livealc+exhealth+ I(log(educsq+1))
                    ,family = binomial(link = logit), data = alcohol)
probit1.4_new <- glm(abuse~ famsize + fathalc+ livealc+exhealth+ I(log(educsq+1)
                                                                   ),family =
                       binomial(link = probit), data = alcohol)

  resettest(logit1.4_new)
```

```
##
##  RESET test
##
## data:  logit1.4_new
## RESET = 2.473, df1 = 2, df2 = 9814, p-value = 0.08438
```

```
  resettest(probit1.4_new)
```

```
##
##  RESET test
##
## data:  probit1.4_new
## RESET = 2.473, df1 = 2, df2 = 9814, p-value = 0.08438
```

The new logit and probit models have now been created. I also did a quick reset test on each one. Both p-values are greater than .05, we have linearity in each model. Before we start speaking to the LPM, Logit and probit models, let's look into the difference in R squared values from the LPM models and the AIC from the new logit and probit models to the old models.

Table 5: R squared value by LPM

lpm1	0.00584
lpm2	0.00664
lpm3	0.00672
lpm4	0.00827
lpm5	0.00829
Newest LPM	0.01142

LPM Comparisons

```
a1 <- summary(lpm1)
a2 <- summary(lpm2)
a3 <- summary(lpm3)
a4 <- summary(lpm4)
a5 <- summary(lpm5)
a6 <-summary(lm4_linear)
a1 <-a1$r.squared
a2 <-a2$r.squared
a3 <-a3$r.squared
a4 <-a4$r.squared
a5 <-a5$r.squared
a6 <-a6$r.squared

string <- c("lpm1","lpm2","lpm3","lpm4","lpm5","Newest LPM")
string2 <- round(c(a1,a2,a3,a4,a5,a6),5)
table <- cbind(string,string2)
kable(table, caption = "R squared value by LPM", col.names = c("",""))
```

Results

We can see from the table above that with our current LPM model, we were able to increase the variance explained from .00829 (lpm5, which had the highest R squared value) all the way to .01142. This may seem small, but when looking at it from a percentage change, we were able to increase our R squared value by 37%. The automated method was able to increase the variance explained and increased our R squared values.

#logit and Probit Comparisons

```
L_AIC1 <- logit1$aic
L_AIC2 <- logit2$aic
L_AIC3 <- logit3$aic
L_AIC4 <- logit4$aic
L_AIC5 <- logit5$aic
P_AIC1 <- probit1$aic
P_AIC2 <- probit2$aic
P_AIC3 <- probit3$aic
P_AIC4 <- probit4$aic
P_AIC5 <- probit5$aic
```

Table 6: Logit and Probit AICS

logit1	6306.83
Logit2	6302.3
Logit3	6303.56
logit4	6290.77
logit5	6292.76
Probit1	6306.43
Probit2	6301.83
Probit3	6303.1
Probit4	6290.39
Probit5	6292.39

Table 7: New Logit and Probit AICS

Logit Model	6251.15
Probit Model	6251.58

```
x <- round(c(L_AIC1, L_AIC2 ,L_AIC3 ,L_AIC4, L_AIC5 , P_AIC1 ,P_AIC2 ,P_AIC3 , P_AIC4 ,P_AIC5 ),2)
z <- c("logit1","Logit2","Logit3","logit4","logit5","Probit1","Probit2","Probit3","Probit4","Probit5")
y <- cbind(z,x)
kable(y, caption = "Logit and Probit AICS", col.names = c("",""))
```

```
L_new <- logit1.4_new$aic
P_new <- probit1.4_new$aic

a <- round(c(L_new,P_new),2)
b <- c("Logit Model"," Probit Model")
c <- cbind(b,a)
kable(c, caption = "New Logit and Probit AICS", col.names = c("",""))
```

Results

The lowest AIC for the first logit and probit models were both in models 4. The logit model had an AIC of 6292.76 and the probit model had an AIC of 6290.39. The new models had lower AICs, 6251.15 for the logit model, and 6251.58 for the probit model.

Looking at the Models

```
##
## ============================================================================
##                                       LPM Models
##                  ---------------------------------------------------------
##                  lm1          lm2          lm3          lm4          Lm4_linear
## ----------------------------------------------------------------------------
## (Intercept)        0.14 ***     0.07 ***     0.05         0.07         0.20 ***
##                   (0.01)       (0.02)       (0.04)       (0.04)       (0.03)
## famsize           -0.01 ***    -0.01 ***    -0.01 ***    -0.01 ***    -0.01 ***
##                   (0.00)       (0.00)       (0.00)       (0.00)       (0.00)
## fathalc            0.04 ***     0.05 ***     0.05 ***     0.05 ***     0.04 ***
##                   (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## livealc            0.01         0.01         0.01         0.01         0.01
##                   (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## exhealth          -0.03 ***    -0.03 ***    -0.03 ***    -0.02 ***    -0.02 ***
##                   (0.01)       (0.01)       (0.01)       (0.01)       (0.01)
## ethanol                         0.03 ***     0.05         0.06
##                                 (0.01)       (0.04)       (0.04)
## ethanolsq                                    -0.00        -0.01
##                                              (0.01)       (0.01)
## educsq                                                    -0.00 ***
##                                                           (0.00)
## log(educsq + 1)                                                        -0.01 *
##                                                                        (0.01)
## ----------------------------------------------------------------------------
## R^2                0.01         0.01         0.01         0.01         0.01
## Adj. R^2           0.01         0.01         0.01         0.01         0.01
## Num. obs.       9822         9822         9822         9822         9822
## ============================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
##
## Ordered Logit of Alcohol Abuse (Odds Ratio)
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                abuse
## ------------------------------------------------
## famsize                       0.845***
##                               p = 0.000
##
## fathalc                       1.539***
##                               p = 0.002
##
## livealc                         1.117
##                               p = 0.402
##
## exhealth                      0.767***
##                              p = 0.0003
##
## I(log(educsq + 1))             0.876**
##                               p = 0.026
##
## Constant                      0.335***
##                              p = 0.0005
##
## ------------------------------------------------
## Observations                    9,822
## Log Likelihood                -3,119.576
## Akaike Inf. Crit.              6,251.151
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Probit models do not have odd's ratios, so I will speak to the LPM and Logit models. When looking at the moving from Lm1 to lm2, the covariate fathalc increases from .04 to .05 when we include the covariate ethanol. When going from model 4 to 5, educsq increase from very small <-.00 to -.01, but it went from down in signficance to p<.05 from p<.001. When looking at lm4 compared to Lm4_linear, we saw a decrease in fathalc from .05 to .04, but the significance stayed the same. When looking at Lm4_linear, if your father is an alcoholic, that is associated with a 4% (p<.001) increase in abusing alcohol. Continuing to look at Lmr_linear, a one unit increase in famsize is associated with a 1% (p<.001) decrease in abusing alcohol.

When looking at the Logit model. If one's father is an alcholic, the likelihood of abusing alcohol versus not abusing alcohol is 1.539 times higher with p = .002. If one lives with an alcoholic, they are 1.117 more times likely to abuse alcohol. For A one-unit increase in famsize, the likelihood of abusing alcohol is .845 times as likely.

Finally I will create synthetic individuals and make predictions of the likelihood of an individual to abuse alcohol based on the preset conditions. These can be seen below.

Table 8: Alcohol Abuse Probabilites

	LPM	Logit	Probit
1	24.43	32.72	30.91
2	0.24	3.24	2.99

```
x_values = list(famsize = c(1,8), fathalc = c(1,0), livealc = c(1,0),
                exhealth = c(0,1), educsq = c(0, mean(alcohol$educsq)))

lm_predict <- predict(lm4_linear,x_values)
logit_predict <- predict(logit1.4_new, x_values, type= "response")


probit_predict <- predict(probit1.4_new, x_values, type = "response")

predictions <- round(cbind(lm_predict, logit_predict, probit_predict),4)*100
kable(predictions, caption = "Alcohol Abuse Probabilites",
      col.names = c("LPM","Logit","Probit"),
      row.names = TRUE)
```

Results

I made predictions for 2 different individuals. Individual 1 has a family size of 1, their father is an alcoholic, they live with an alcoholic, they are not in perfect health, and I used 0 as the educsq variable. For individual 2, I pretty much did the opposite from individual 1 except for the educsq covariate, in which I used the mean. From looking at the LPM and Logit model above, I would expect the probability for individual 1 when compared to individual 2 to be much higher.

As expected, individual 1 has a much higher probability of abusing alcohol, 24%, 33% and 30% for the LPM, Logit, and Probit models respectively. Individual 2 has a much lower probability of abusing alcohol, .24%, 3.2% and 3% for the LPM, Logit, and Probit models respectively.

Conclusion

Overall, I found this project very useful. If I were to do this project over again, I would probably choose a different dataset, or use a different indicator as my outcome variable. When I saw the very low R squared value for my initial models, I thought I could increase that number by choosing the most the right covariates. I was able to increase the R squared value by 37% but it was still quite low.

Automating the linear models was quite difficult. The models would not run smoothly in my loops, there were many different issues that were quite time consuming. For example, to loop through a Linear Model and print the results, the variables in the LM function needed to be in formula state before running the loop, or you would have to do that in the loop. They needed to be pasted as a formula before looped through. I found the loops with regression, not very user friendly. I also found that it was not very easy to loop and print summaries from models, I believe this was due to the functionality of LM, Glm functions.

Even though the model did not explain much variation in the data, I do believe I was able to pick the best variables to predict whether an individual would abuse alcohol. I was not able to fully automate the process of picking the best fit model. I did not have enough time to focus on this part, I was able to automate part of the process. I found it a bit difficult to extract the linearHypothesis data to use for automation, but I was able to make this process a bit easier when applying to other data sets.

I have a lot to learn, and I look forward to continuing my study into linear regression.