Assignment 3

NAME: Nicholas Tolley
DUE DATE: March 21th, 6:00pm

Problem 1 (100 pts)

In the folder Assignment 3, you will find the data set called FF_wave6_2020v2.dta. This data set is from the Fragile Family Data Set, and it includes many different variables (socio-demographic, economics, and health status) of teenagers (15 years old) and their parents. The codebook (ff_wave6_codebook.txt) associated with the data set is on Canvas (folder Assignment 3).

Loading the dataset:

```
rm(list=ls())
library(plyr)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-6
```

```
library(haven)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following objects are masked from 'package:Matrix':
##
##     expand, pack, unpack
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(tidyr)
library(viridis)
```

```
## Loading required package: viridisLite
```

(a) (20 points) Consider the variable *doctor diagnosed youth with depression/anxiety*. In the data set, the name of this variable is *p6b5*. Then consider in the data set these variables: *p6b10*, *p6b35*, *p6b55*, *p6b60*, *p6c21*, *p6f32*, *p6f35*, *p6h74*, *p6h102*, *p6i7*, *p6i8*, *p6i11*, *p6j37*, *k6b21a*, *k6b22a*, *k6c1*, *k6c4e*, *k6c28*, *k6d37*, *k6f63*, *ck6cbmi*, *k6d10*. Now, you have a data set with 4898 subjects and 23 variables. Clean the data in these three steps. 1- Each variable has a value with a number and a text (for example, a value for the variable *p6b5* is *2 No*). Remove the text from all the variables in the data set (hint: use the function sub for each column). 2- Transform each variable in numeric (hint: use the function as.numeric for each column). 3- Transform all the values less than 0 in NA and then remove all your NA values from the data set. Show the dimensions of the cleaned data and print the first 6 rows.

The code below subsets the columns indicated, transforms the columns into type numeric, replaces negative numbers with NA, and then drops NA rows from the dataframe.

```
subset_cols <- c('p6b5', 'p6b10', 'p6b35', 'p6b55', 'p6b60', 'p6c21', 'p6f32', 'p6f35', 'p6h74', 'p6h10

df<-read_dta(file='FF_wave6_2020v2.dta')
df <- subset(df, select=subset_cols)

for (col in subset_cols){
  df[col] <- as.numeric(unlist(df[col]))
}

df[df < 0] <- NA
df <- drop_na(df)
```

This produces a dataframe with dimensions:

```
dim(df)
```

```
## [1] 488  23
```

And the following first 6 rows:

```
df[1:6,]
```

```
## # A tibble: 6 x 23
```

```
##    p6b5 p6b10 p6b35 p6b55 p6b60 p6c21 p6f32 p6f35 p6h74 p6h102  p6i7  p6i8 p6i11
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1     1     2     2     1     1     2     2     1     1      1     2     2     3
## 2     2     2     1     1     1     2     2     2     2      2     2     2     4
## 3     1     2     1     2     1     2     2     2     2      2     2     2     4
## 4     2     2     1     2     1     2     2     2     1      2     1     1     4
## 5     2     2     1     1     1     2     2     2     1      2     2     2     4
## 6     2     2     1     1     1     2     2     2     2      2     2     3     4
## # ... with 10 more variables: p6j37 <dbl>, k6b21a <dbl>, k6b22a <dbl>,
## #   k6c1 <dbl>, k6c4e <dbl>, k6c28 <dbl>, k6d37 <dbl>, k6f63 <dbl>,
## #   ck6cbmi <dbl>, k6d10 <dbl>
```

(a) (20 points) Now call the variables with an appropriate name (for example *p6b5* can become *Depression*). Perform a logistic regression using the variable *Depression* as the outcome and the remaining variables as the covariates. Be careful: the variable *Depression* has value 1 and 2, you should transform in 0,1 before running the logistic regression in R (1 for Yes, 0 for No). What are the important and significant covariates for the depression? For these, what can you say about the standard error? Perform the binned residual plot by using the library ggplot2 in R. Then write a function in R that gives the odds ratio for each beta and its upper and lower confidence intervals (CI). Use this function to produce the beta coefficient related to the covariate (ADD or *p6b10*), and its CI in term of ODDS RATIO. What can you say about that? Is still significant?

The code below renames the columns to more human readable names, as well as transforms the outcome column to values of 0 or 1

```
col_names <- c('depression',    'ADD',  'mean', 'sleep_trouble',    'runaway',  'suspended',    'drug_p

df_pred <- df
colnames(df_pred) <- col_names
df_pred['depression'] <- df_pred['depression'] - 1
```

Next we fit a logistic regression model to the data and summarize the results. As shown, the significant covariates include: ADD, sleep_trouble, suspended, and school_attention_problem

The reason they are significant is because the values have a very low probability of being equal to zero (the null hypothesis), therefore the standard errors for the significant coefficients are relatively small compared to the magnitude of the coefficient's estimate.

```
fit_logistic <- glm(depression ~ ., data=df_pred, family=binomial(link="logit"))
summary_logistic <- summary(fit_logistic)
summary_logistic
```

```
##
## Call:
## glm(formula = depression ~ ., family = binomial(link = "logit"),
##     data = df_pred)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1486   0.1772   0.2570   0.3921   2.0580
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.603900   3.291053   1.095 0.273490
## ADD                        1.677897   0.471019   3.562 0.000368 ***
## mean                      -0.295209   0.337778  -0.874 0.382132
## sleep_trouble             -1.453627   0.270194  -5.380 7.45e-08 ***
```

```
## runaway                  -0.678764   0.740884  -0.916 0.359586
## suspended                -1.041740   0.458242  -2.273 0.023006 *
## drug_problem              0.891316   0.481588   1.851 0.064200 .
## parent_jailed            -0.587508   0.474254  -1.239 0.215418
## smoker                    0.424382   0.357364   1.188 0.235016
## jailed                    0.362720   0.651642   0.557 0.577785
## neighbor_help             0.164381   0.275061   0.598 0.550097
## close_neighborhood       -0.066191   0.234927  -0.282 0.778133
## gang_problem             -0.008095   0.185011  -0.044 0.965101
## free_food                 0.863758   0.447409   1.931 0.053535 .
## school_attention_problem -0.731812   0.239586  -3.054 0.002254 **
## sports_team              -0.014254   0.111795  -0.128 0.898542
## parent_relationship      -0.015920   0.161038  -0.099 0.921249
## calm_home                -0.599335   0.309494  -1.936 0.052807 .
## father_close              0.070795   0.173388   0.408 0.683049
## physically_active         0.100177   0.091585   1.094 0.274038
## marijuana                 0.360426   0.385711   0.934 0.350074
## BMI                      -0.039232   0.026002  -1.509 0.131354
## menstruation_age         -0.010842   0.130825  -0.083 0.933952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 347.79  on 487  degrees of freedom
## Residual deviance: 245.05  on 465  degrees of freedom
## AIC: 291.05
##
## Number of Fisher Scoring iterations: 6
```

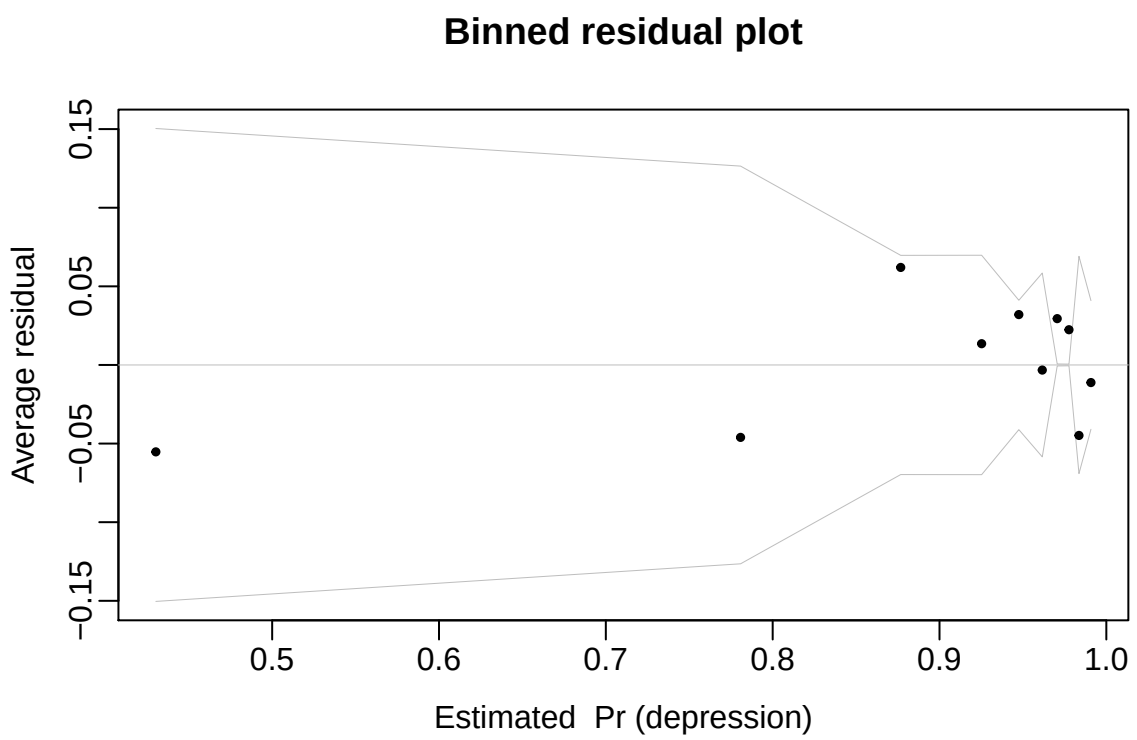
Next we can visualize the errors of this model using a binned residual plot

```
pred_logistic <- fit_logistic$fitted.values

binned.resids <- function (x, y, nclass=sqrt(length(x))){
  breaks.index <- floor(length(x)*(1:(nclass-1))/nclass)
  breaks <- c (-Inf, sort(x)[breaks.index], Inf)
  output <- NULL
  xbreaks <- NULL
  x.binned <- as.numeric (cut (x, breaks))
  for (i in 1:nclass){
    items <- (1:length(x))[x.binned==i]
    x.range <- range(x[items])
    xbar <- mean(x[items])
    ybar <- mean(y[items])
    n <- length(items)
    sdev <- sd(y[items])
    output <- rbind (output, c(xbar, ybar, n, x.range, 2*sdev/sqrt(n)))
  }
  colnames (output) <- c ("xbar", "ybar", "n", "x.lo", "x.hi", "2se")
  return (list (binned=output, xbreaks=xbreaks))
}

br <- binned.resids(pred_logistic, df_pred$depression-pred_logistic, nclass=10)$binned
plot(range(br[,1]), range(br[,2],br[,6],-br[,6]), xlab="Estimated  Pr (depression)", ylab="Average resi
```

```
abline (0,0, col="gray", lwd=.5)
lines (br[,1], br[,6], col="gray", lwd=.5)
lines (br[,1], -br[,6], col="gray", lwd=.5)
points (br[,1], br[,2], pch=19, cex=.5)
```



The code below expresses the beta coefficients in terms of odds ratios. First we create a function that takes in the estimate and its standard error, and returns a 95% confidence interval. Given the beta coefficient for ADD, the 95% CI for the odds ration is [2.12, 13.48], which we can interpret as: given the presence of ADD (value=2), the ratio of the probabilities of p(depression=1) / p(depression=0) is within that range.

```
beta_to_odds <- function(estimate, standard_error){
  beta_lower <- estimate - (1.96) * standard_error
  beta_upper <- estimate + (1.96) * standard_error

  odds_CI <- c(exp(beta_lower), exp(beta_upper))
  return (odds_CI)
}

ADD_est <- summary_logistic$coefficients[2,1]
ADD_se <- summary_logistic$coefficients[2,2]

print(beta_to_odds(ADD_est, ADD_se))
```

```
## [1]  2.126975 13.478477
```

(a) (20 points) Use the forward step procedure to detect the important covariates. Then, only for estimates that are greater than 0, draw with ggplot a plot similar to Figure 1. So in the x-axis, you should

have each beta (beta1, beta2, etc.). In the y-axis, the estimate greater than 0 with the correspondent standard error. Be careful this plot is taken from another data set, so do not expect similar results. Take special care of the legend and the label. What can you say about this plot?

```
fit_start <- glm(depression ~ 1, data=df_pred, family=binomial(link="logit"))
fit_end <- glm(depression ~ ., data=df_pred, family=binomial(link="logit"))
forward_logistic <- stepAIC(fit_start, direction="forward", scope=list(upper=fit_end, lower=fit_start))
```

```
## Start:  AIC=349.79
## depression ~ 1
##
##                             Df Deviance    AIC
## + sleep_trouble              1   291.83 295.83
## + ADD                        1   317.20 321.20
## + school_attention_problem   1   330.97 334.97
## + smoker                     1   339.03 343.03
## + free_food                  1   340.11 344.11
## + drug_problem               1   340.58 344.58
## + sports_team                1   341.07 345.07
## + mean                       1   342.98 346.98
## + runaway                    1   343.61 347.61
## + BMI                        1   344.85 348.85
## + marijuana                  1   345.10 349.10
## + jailed                     1   345.45 349.45
## <none>                           347.79 349.79
## + physically_active          1   346.49 350.49
## + father_close               1   346.65 350.65
## + parent_relationship        1   346.95 350.95
## + calm_home                  1   347.06 351.06
## + close_neighborhood         1   347.41 351.41
## + neighbor_help              1   347.45 351.45
## + menstruation_age           1   347.47 351.47
## + parent_jailed              1   347.58 351.58
## + gang_problem               1   347.78 351.78
## + suspended                  1   347.78 351.78
##
## Step:  AIC=295.83
## depression ~ sleep_trouble
##
##                             Df Deviance    AIC
## + ADD                        1   277.47 283.47
## + school_attention_problem   1   277.60 283.60
## + drug_problem               1   287.92 293.92
## + free_food                  1   288.82 294.82
## + smoker                     1   288.95 294.95
## + marijuana                  1   289.48 295.48
## + BMI                        1   289.82 295.82
## <none>                           291.83 295.83
## + sports_team                1   289.98 295.98
## + mean                       1   290.01 296.01
## + runaway                    1   290.10 296.10
## + physically_active          1   290.32 296.32
## + jailed                     1   290.95 296.95
## + suspended                  1   291.27 297.27
## + parent_relationship        1   291.45 297.45
```

```
## + calm_home                 1   291.49 297.49
## + menstruation_age          1   291.67 297.67
## + father_close              1   291.70 297.70
## + parent_jailed             1   291.76 297.76
## + neighbor_help             1   291.76 297.76
## + close_neighborhood        1   291.79 297.79
## + gang_problem              1   291.82 297.82
##
## Step:  AIC=283.47
## depression ~ sleep_trouble + ADD
##
##                            Df Deviance    AIC
## + school_attention_problem  1   267.71 275.71
## + free_food                 1   273.37 281.37
## + drug_problem              1   273.50 281.50
## + smoker                    1   274.74 282.74
## + suspended                 1   275.40 283.40
## <none>                          277.47 283.47
## + BMI                       1   275.90 283.90
## + calm_home                 1   276.06 284.06
## + runaway                   1   276.50 284.50
## + physically_active         1   276.52 284.52
## + marijuana                 1   276.69 284.69
## + jailed                    1   276.71 284.71
## + sports_team               1   276.75 284.75
## + mean                      1   277.07 285.07
## + neighbor_help             1   277.19 285.19
## + parent_relationship       1   277.31 285.31
## + gang_problem              1   277.38 285.38
## + father_close              1   277.39 285.39
## + menstruation_age          1   277.45 285.45
## + close_neighborhood        1   277.47 285.47
## + parent_jailed             1   277.47 285.47
##
## Step:  AIC=275.71
## depression ~ sleep_trouble + ADD + school_attention_problem
##
##                       Df Deviance    AIC
## + drug_problem         1   264.66 274.66
## + free_food            1   264.73 274.73
## + suspended            1   265.21 275.21
## + smoker               1   265.39 275.39
## + calm_home            1   265.42 275.42
## <none>                     267.71 275.71
## + BMI                  1   265.76 275.76
## + physically_active    1   266.87 276.87
## + marijuana            1   266.94 276.94
## + sports_team          1   267.20 277.20
## + runaway              1   267.22 277.22
## + jailed               1   267.39 277.39
## + neighbor_help        1   267.42 277.42
## + parent_jailed        1   267.57 277.57
## + menstruation_age     1   267.63 277.63
## + mean                 1   267.64 277.64
```

```
## + gang_problem         1   267.70 277.70
## + father_close         1   267.70 277.70
## + close_neighborhood   1   267.71 277.71
## + parent_relationship  1   267.71 277.71
##
## Step:  AIC=274.66
## depression ~ sleep_trouble + ADD + school_attention_problem +
##     drug_problem
##
##                       Df Deviance    AIC
## + suspended            1   261.82 273.82
## + calm_home            1   262.32 274.32
## + parent_jailed        1   262.56 274.56
## + smoker               1   262.56 274.56
## <none>                     264.66 274.66
## + free_food            1   262.70 274.70
## + BMI                  1   262.70 274.70
## + physically_active    1   263.86 275.86
## + marijuana            1   263.86 275.86
## + sports_team          1   264.11 276.11
## + runaway              1   264.34 276.34
## + neighbor_help        1   264.40 276.40
## + jailed               1   264.43 276.43
## + father_close         1   264.55 276.55
## + menstruation_age     1   264.55 276.55
## + parent_relationship  1   264.57 276.57
## + mean                 1   264.60 276.60
## + gang_problem         1   264.66 276.66
## + close_neighborhood   1   264.66 276.66
##
## Step:  AIC=273.82
## depression ~ sleep_trouble + ADD + school_attention_problem +
##     drug_problem + suspended
##
##                       Df Deviance    AIC
## + free_food            1   259.16 273.16
## + BMI                  1   259.29 273.29
## + calm_home            1   259.38 273.38
## <none>                     261.82 273.82
## + smoker               1   259.86 273.86
## + parent_jailed        1   260.06 274.06
## + marijuana            1   260.43 274.43
## + physically_active    1   260.67 274.67
## + runaway              1   261.01 275.01
## + sports_team          1   261.13 275.13
## + mean                 1   261.50 275.50
## + jailed               1   261.55 275.55
## + neighbor_help        1   261.64 275.64
## + father_close         1   261.67 275.67
## + parent_relationship  1   261.67 275.67
## + menstruation_age     1   261.69 275.69
## + gang_problem         1   261.75 275.75
## + close_neighborhood   1   261.81 275.81
##
```

```
## Step:  AIC=273.16
## depression ~ sleep_trouble + ADD + school_attention_problem +
##     drug_problem + suspended + free_food
##
##                       Df Deviance    AIC
## + calm_home            1   256.24 272.24
## + BMI                  1   256.73 272.73
## + parent_jailed        1   257.05 273.05
## <none>                     259.16 273.16
## + smoker               1   257.64 273.64
## + marijuana            1   257.71 273.71
## + physically_active    1   257.76 273.76
## + sports_team          1   258.31 274.31
## + runaway              1   258.34 274.34
## + mean                 1   258.66 274.66
## + jailed               1   258.82 274.82
## + neighbor_help        1   258.86 274.86
## + parent_relationship  1   258.99 274.99
## + father_close         1   259.10 275.10
## + menstruation_age     1   259.10 275.10
## + gang_problem         1   259.16 275.16
## + close_neighborhood   1   259.16 275.16
##
## Step:  AIC=272.24
## depression ~ sleep_trouble + ADD + school_attention_problem +
##     drug_problem + suspended + free_food + calm_home
##
##                       Df Deviance    AIC
## + BMI                  1   253.42 271.42
## <none>                     256.24 272.24
## + parent_jailed        1   254.24 272.24
## + marijuana            1   254.46 272.46
## + smoker               1   254.47 272.47
## + physically_active    1   254.51 272.51
## + runaway              1   255.33 273.33
## + sports_team          1   255.54 273.55
## + mean                 1   255.63 273.63
## + jailed               1   255.79 273.79
## + neighbor_help        1   255.97 273.97
## + parent_relationship  1   256.11 274.11
## + menstruation_age     1   256.20 274.20
## + father_close         1   256.21 274.21
## + close_neighborhood   1   256.23 274.23
## + gang_problem         1   256.24 274.24
##
## Step:  AIC=271.42
## depression ~ sleep_trouble + ADD + school_attention_problem +
##     drug_problem + suspended + free_food + calm_home + BMI
##
##                       Df Deviance    AIC
## <none>                     253.42 271.42
## + parent_jailed        1   251.55 271.55
## + marijuana            1   251.72 271.72
## + smoker               1   251.72 271.72
```

```
## + physically_active    1   252.36 272.36
## + mean                 1   252.53 272.52
## + runaway              1   252.53 272.53
## + sports_team          1   252.89 272.89
## + jailed               1   253.00 273.00
## + neighbor_help        1   253.09 273.09
## + father_close         1   253.37 273.37
## + parent_relationship  1   253.38 273.38
## + close_neighborhood   1   253.40 273.40
## + gang_problem         1   253.42 273.42
## + menstruation_age     1   253.42 273.42
```

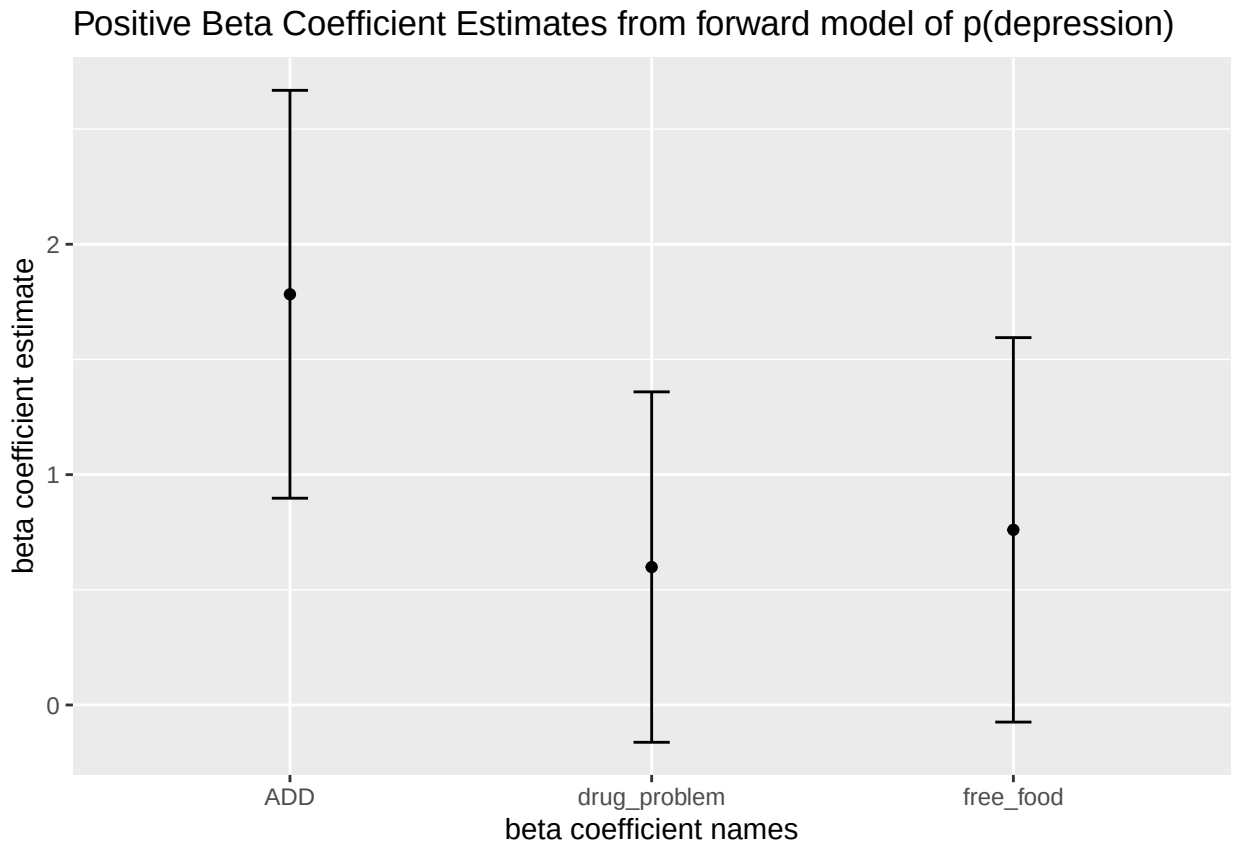
```
summary_forward_logistic <- summary(forward_logistic)
```

```
forward_coefficients <- summary_forward_logistic$coefficients
forward_mean <- forward_coefficients[2:dim(forward_coefficients)[1],1]
forward_errors <- forward_coefficients[2:dim(forward_coefficients)[1],2]
forward_names <- rownames(forward_coefficients)[2:dim(forward_coefficients)[1]]

df_forward <- data.frame(forward_names, forward_mean, forward_errors)
colnames(df_forward) <- c('names', 'mean', 'error')

df_forward <- filter(df_forward, df_forward$mean > 0)

ggplot(df_forward, aes(x=names, y=mean)) +
  geom_point() +
  geom_errorbar(width=.1, aes(ymin=mean - (1.96*error), ymax=mean + (1.96*error))) +
  labs(y='beta coefficient estimate', x='beta coefficient names',
  title='Positive Beta Coefficient Estimates from forward model of p(depression)')
```



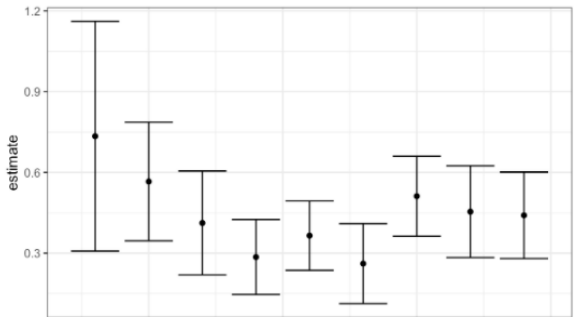


Figure 1: Estimate

(d) (20 points) Perform a bootstrap of 1000 samples for beta 1 (ADD or *p6b10*), beta 2 (sleep or *p6b55*), and beta 3 (attention at school or *k6b21a*) with a model that contains all the coefficients obtained in the forward procedure in point c. Plot these three bootstrapped beta coefficients that you have obtained with a boxplot in the ggplot (similar to Figure 2). (make sure not to use the default colors but rather choose your own). What can you say about these three distributions obtained?

The code below performs a 1000 sample bootstrap using the forward model solution from part c.

```
fc <- function(df, i){
  df_boot <- df[i,]
  boot_glm <- glm(forward_logistic[["formula"]], data=df_boot, family=binomial(link="logit"))
  return(boot_glm[['coefficients']])
}
```

```
depression_boot <- boot(df_pred, fc, R=1000)
```

Making the box plot for the bootstrap estimates for the beta coefficients indicated, we can see that all 3 seem to be symmetric with the first and 3rd quantiles being roughly equidistant from the median. Additionally, all 3 exhibit outliers at the maximum and minimum extents.

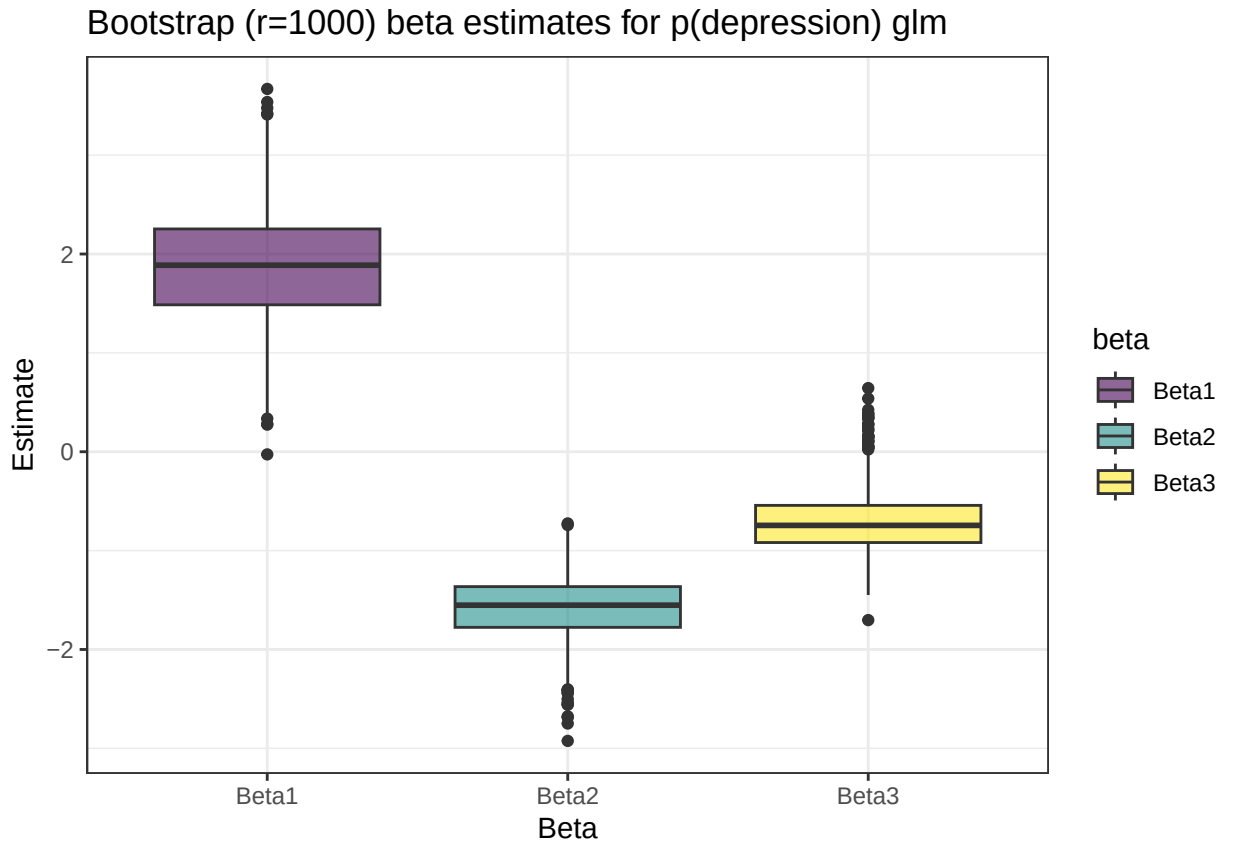
Interestingly the sign of the coefficients indicate that ADD (beta1) is associate with increased probability of depression, whereas sleep trouble (beta2) and school attention problems (beta3) are both associated to increased probability of depression.

```
df_boot_beta <- data.frame(depression_boot[["t"]][,1:length(forward_names) + 1])
colnames(df_boot_beta) <- forward_names

boot_cols <- c('ADD', 'sleep_trouble', 'school_attention_problem')
df_boot_beta <-df_boot_beta[, boot_cols]
colnames(df_boot_beta) <- c('Beta1', 'Beta2', 'Beta3')

df_boot_beta <- pivot_longer(df_boot_beta, cols=colnames(df_boot_beta), names_to = "beta", values_to = '

ggplot(df_boot_beta, aes(x=beta, y=coef, fill=beta)) + geom_boxplot() +
  scale_fill_viridis(discrete = TRUE, alpha=0.6) +
  labs(x='Beta', y='Estimate', title='Bootstrap (r=1000) beta estimates for p(depression) glm') + theme_
```



(e) (20 points) Perform the Lasso method for the full model. Choose λ with the cross-validation. Then perform the lasso with the best λ obtained. Plot the results in ggplot. Describe the results you obtained. Are the coefficients obtained with the lasso procedure similar to the coefficients obtained with the forward procedure? Explain!

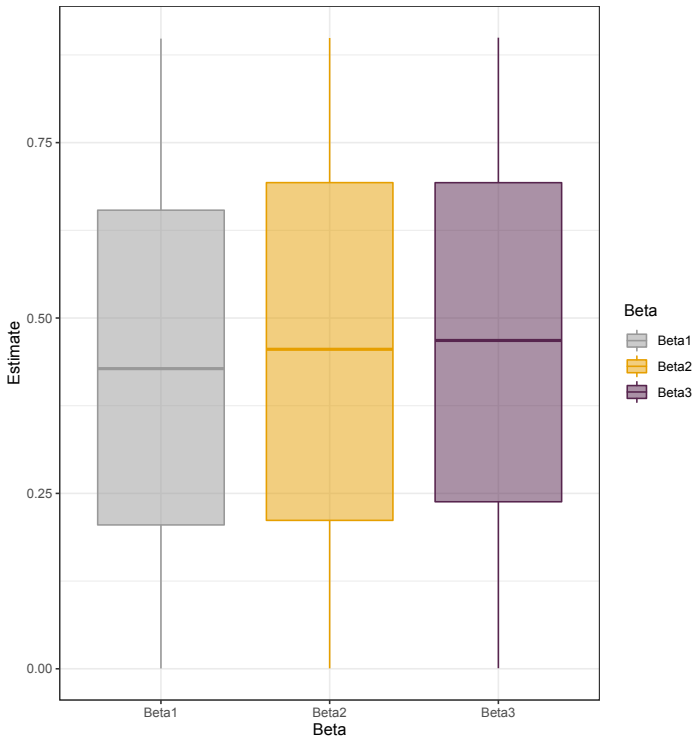


Figure 2: Boxplot

The code below performs cross validation using a 50/50 train/test split to find the best lambda value. With a logarithmically spaced grid between [10e-3, 1] with 100 elements, the best lambda value was identified to be 0.0376

```
set.seed(1)
grid <- 10^seq(0, -3, length = 100)

# Prepare train and test set for full_model
train <- sample_frac(df_pred, 0.5)
test <- setdiff(df_pred, train)

x_train <- model.matrix(depression~., train)[,-1]
x_test <- model.matrix(depression~., test)[,-1]

y_train <- train$depression
y_test <- test$depression

cv.out = cv.glmnet(x_train, y_train, alpha = 1, lambda=grid)

# Select lambda that minimizes training MSE
bestlam = cv.out$lambda.min
fit_lasso = glmnet(x_train, y_train, alpha = 1, lambda=bestlam)
print(bestlam)
```

```
## [1] 0.03764936
```

Next we can inspect the coefficients arrived at using the "best lambda" lasso model. Similar to the forward procedure, we see that ADD, school attention problems, and sleep trouble are all identified as good predictors

with relatively large beta values compared to the remaining covariates. Unlike the forward model however, we see that the magnitude of all the beta coefficients are comparatively smaller (ADD with beta=1.783 vs. 0.116 when comparing forward vs. lasso). Additionally, these are the only 3 active coefficients in the final model. This is because lasso regularization penalizes the L1 norm of the coefficients, which enforces models that are simultaneously sparse, as well as possessing low magnitude beta coefficients.

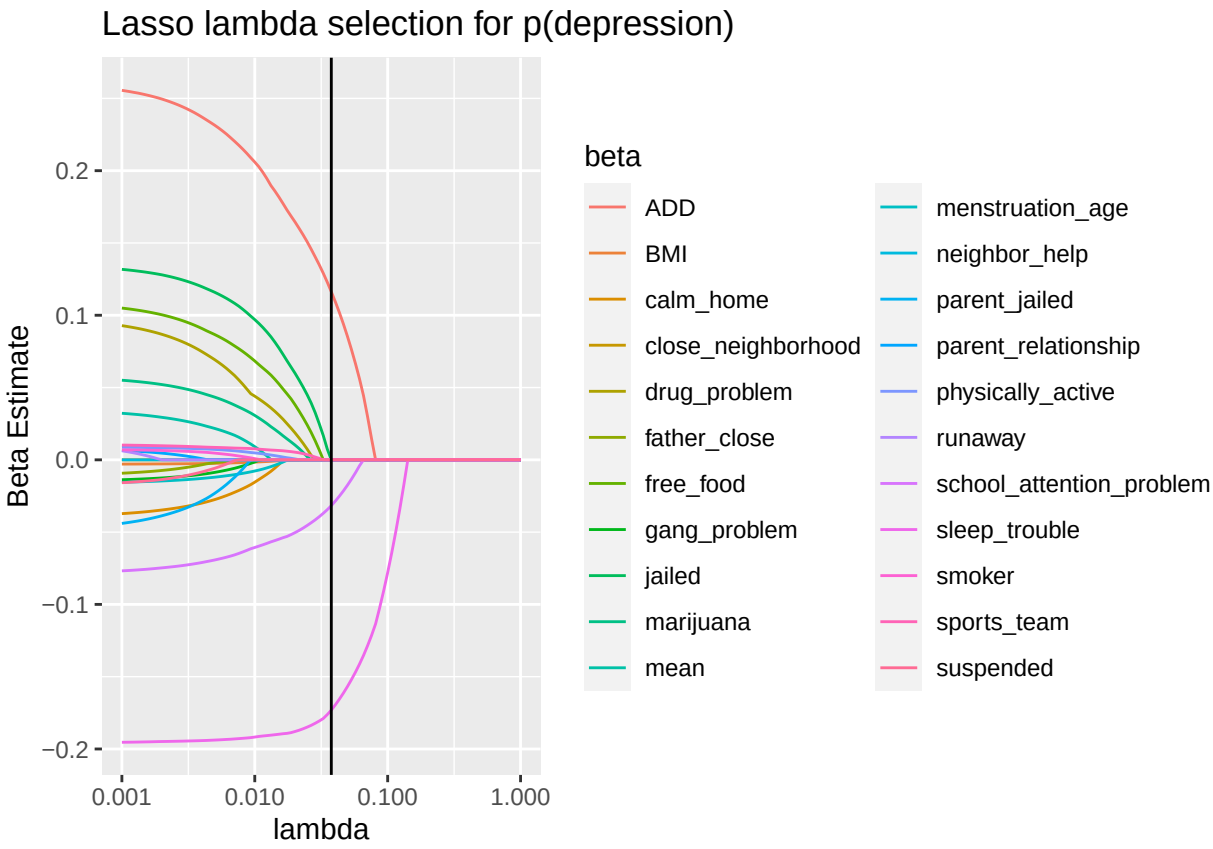
```
print(fit_lasso[["beta"]])
```

```
## 22 x 1 sparse Matrix of class "dgCMatrix"
##                                    s0
## ADD                         0.1162409
## mean                        .
## sleep_trouble              -0.1731253
## runaway                     .
## suspended                   .
## drug_problem                .
## parent_jailed               .
## smoker                      .
## jailed                      .
## neighbor_help               .
## close_neighborhood          .
## gang_problem                .
## free_food                   .
## school_attention_problem -0.0317013
## sports_team                 .
## parent_relationship         .
## calm_home                   .
## father_close                .
## physically_active           .
## marijuana                   .
## BMI                         .
## menstruation_age            .
```

We can view the relationship between lambda and the regression coefficients by plotting the magnitude of the fit beta coefficients as a function of lambda. As we can see, a larger lambda produces a large penalty of regression coefficients such that they approach zero. The best lambda is plotted in black which intersects with the 3 non-zero beta coefficients in the fit model.

```
beta_matrix <- t(as.matrix(coef(cv.out[["glmnet.fit"]])))
lasso_df <- data.frame(beta_matrix[, 2:dim(beta_matrix)[2]])
lasso_beta_cols <- colnames(lasso_df)
lasso_df['lambda'] <- grid
lasso_df <- pivot_longer(lasso_df, cols=all_of(lasso_beta_cols),
                         names_to="beta", values_to = "value")

ggplot(lasso_df, aes(x=lambda, y=value, color=beta)) +
  geom_line() + scale_x_log10() +
  geom_vline(xintercept=bestlam, color="black") +
  labs(title='Lasso lambda selection for p(depression)', x='lambda', y='Beta Estimate')
```


Lasso lambda selection for p(depression)
Beta Estimate
0.2
0.1
0.0
−0.1
−0.2
0.001
0.010
0.100
1.000
lambda
beta
ADD
BMI
calm_home
close_neighborhood
drug_problem
father_close
free_food
gang_problem
jailed
marijuana
mean
menstruation_age
neighbor_help
parent_jailed
parent_relationship
physically_active
runaway
school_attention_problem
sleep_trouble
smoker
sports_team
suspended