Midterm

Noah Estrada-Rand

10/29/2019

Question 1

a)

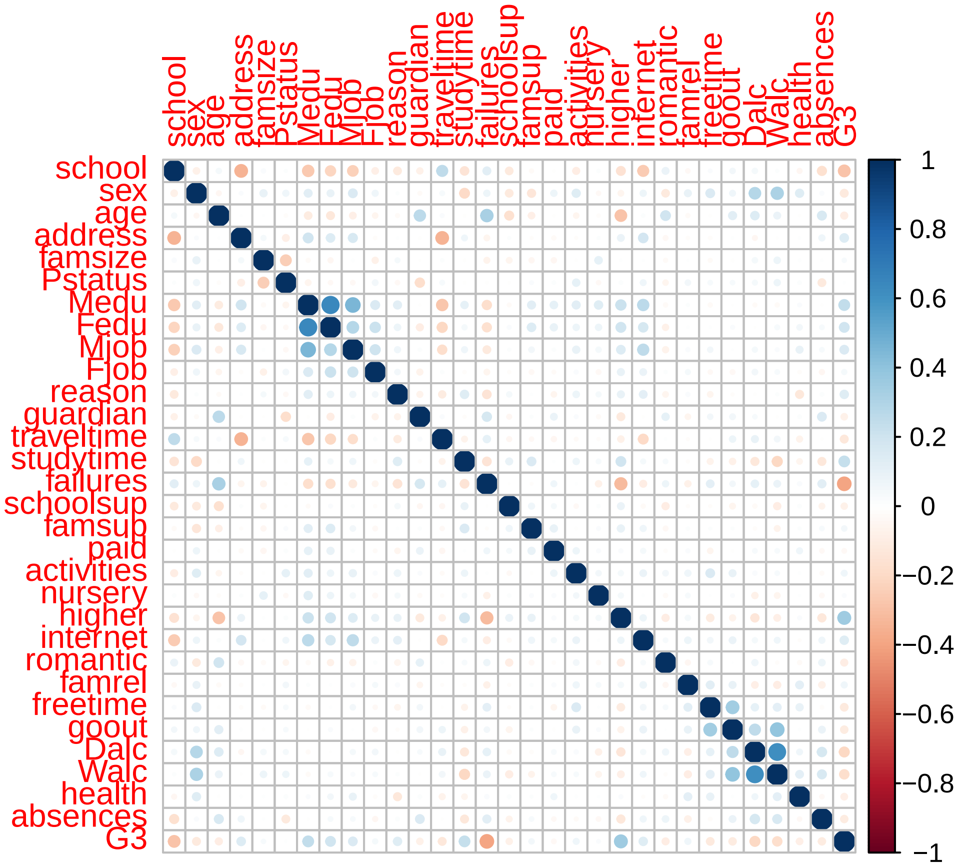
```
str(school)
```

```
## 'data.frame':    628 obs. of  31 variables:
##  $ school    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : num  1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : num  1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : num  1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : num  1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : num  5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : num  1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : num  2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : num  2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : num  1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ activities: num  1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : num  2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : num  1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : num  1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
cormat <- cor(school)
corrplot(cormat)
```



b)

```
library(doBy)
##famsize == 1 is GT3 else LE3
summaryBy(G3~factor(famsize) + factor(Fedu),data = school,FUN = "mean")
```

```
##    famsize Fedu G3."mean"
## 1        1    0  12.16667
## 2        1    1  10.99099
## 3        1    2  11.66906
## 4        1    3  12.29167
## 5        1    4  12.87640
## 6        2    0  12.00000
## 7        2    1  11.23636
## 8        2    2  12.21538
## 9        2    3  12.54839
## 10       2    4  12.80000
```

c)

```
cors <-cor(school$G3,school)

print(cors)
```

```
##          school        sex         age   address    famsize     Pstatus
## [1,] -0.2896278 -0.1198804 -0.09510058 0.1453809 0.03018238 0.004073012
##           Medu      Fedu      Mjob      Fjob    reason    guardian
## [1,] 0.2403547 0.1945171 0.1501966 0.0447195 0.1374548 -0.07802731
##      traveltime studytime   failures   schoolsup     famsup       paid
## [1,] -0.1201439  0.236049 -0.3958347 -0.07421017 0.05086374 -0.0361577
##      activities    nursery    higher  internet    romantic     famrel
## [1,]  0.0626254 0.02108581 0.3535042 0.1386421 -0.09295971 0.06658356
##        freetime      goout       Dalc       Walc      health   absences G3
## [1,] -0.1129029 -0.1056375 -0.2019167 -0.1775532 -0.08259722 -0.1029501  1
```

```
###use size because the last one will be correlated as 1 with g3
tail(sort(abs(cors)),6)
```

```
## [1] 0.2360490 0.2403547 0.2896278 0.3535042 0.3958347 1.0000000
```

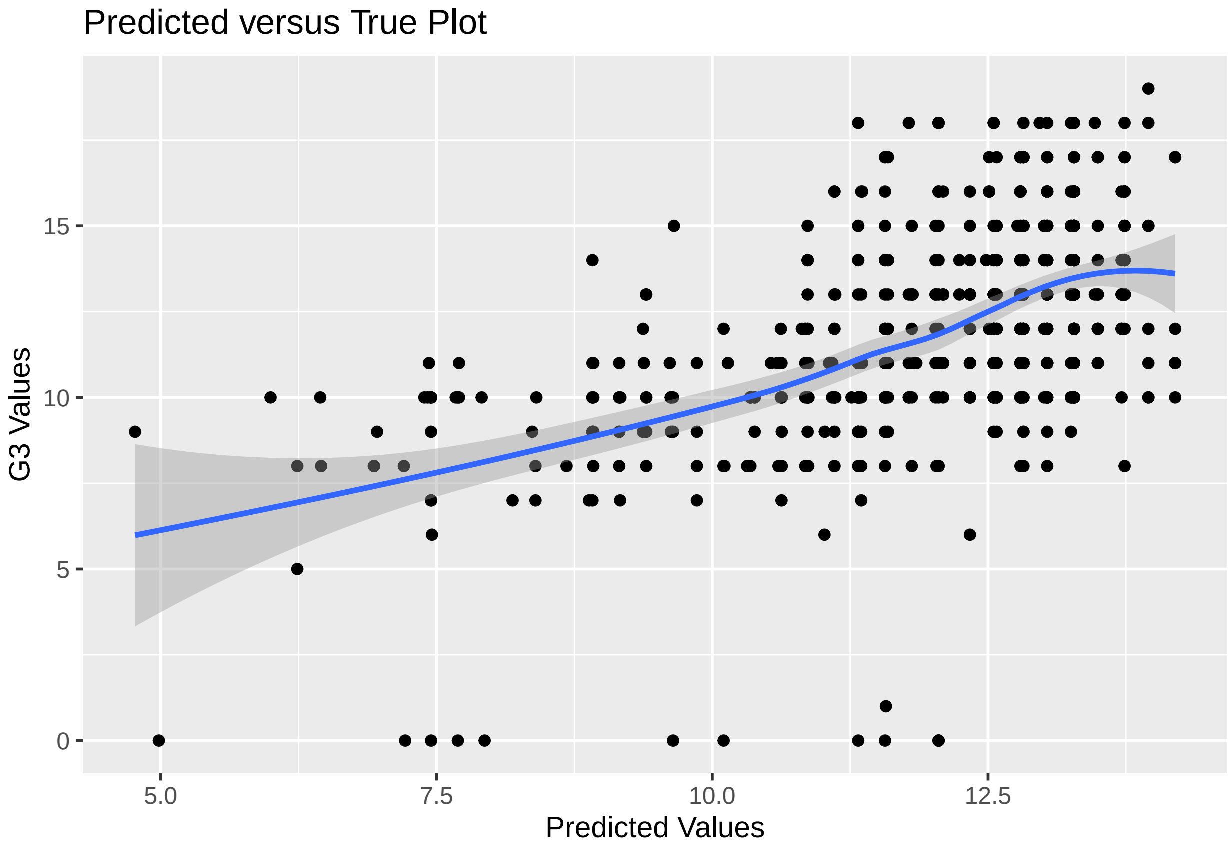
```
### top 5 correlation: failures, higher, school, Medu, studytime
mod1 <- lm(G3~factor(higher)+Medu+studytime + failures +factor(school),
           data = school)
summary(mod1)
```

```
##
## Call:
## lm(formula = G3 ~ factor(higher) + Medu + studytime + failures +
##     factor(school), data = school)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0516  -1.4954  -0.0443   1.7202   6.6768
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.44075    0.48102  19.626  < 2e-16 ***
## factor(higher)2    1.95102    0.36794   5.302 1.59e-07 ***
## Medu               0.24278    0.09955   2.439 0.015013 *
## studytime          0.45843    0.13168   3.482 0.000533 ***
## failures          -1.46301    0.18901  -7.741 4.02e-14 ***
## factor(school)2   -1.22820    0.23816  -5.157 3.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.663 on 622 degrees of freedom
## Multiple R-squared:  0.2828, Adjusted R-squared:  0.2771
## F-statistic: 49.06 on 5 and 622 DF,  p-value: < 2.2e-16
```

f)

```
preds <- predict(mod1)
school$preds <- preds
library(ggplot2)
ggplot(school,aes(y = G3, x = preds)) + geom_point() +geom_smooth() +
  labs(x = "Predicted Values", y="G3 Values",title = "Predicted versus True Plot")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
library(caret)
```

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

```
(RMSE(school$G3,school$preds))^2
```

```
## [1] 7.023489
```

#g)

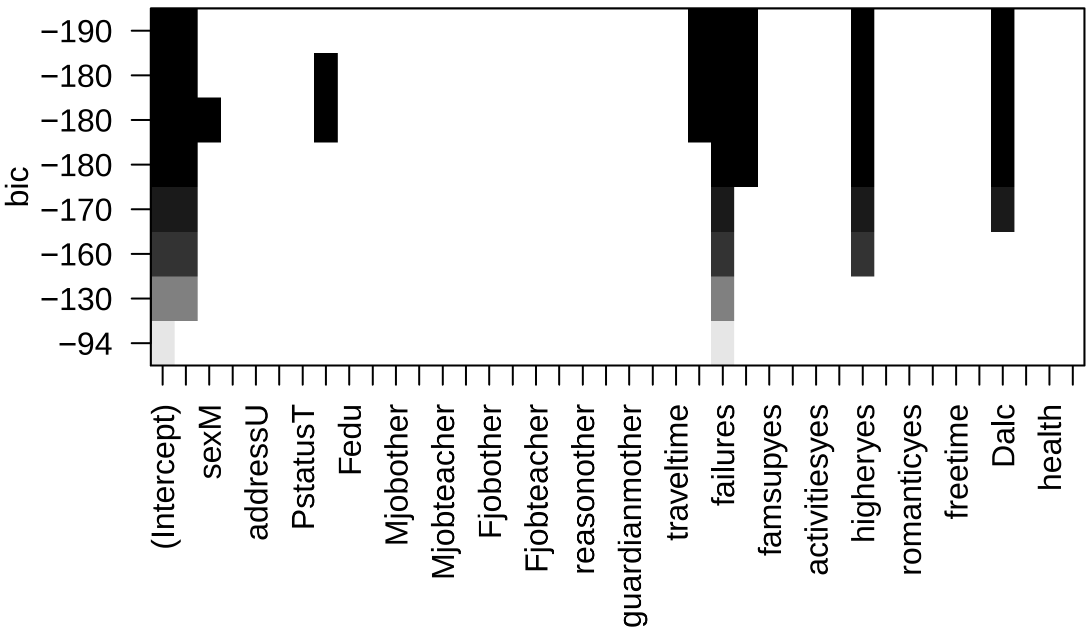
```
school <- read.csv("perf_at_school.csv")
school <- subset(school, select = -c(G1,G2))
str(school)
```

```
## 'data.frame':    628 obs. of  31 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

```r
library(leaps)
fwd_step <- regsubsets(G3~.,data = school,
                       nvmax = 8,
                       method = "forward")
summary(fwd_step)
```

```
## Subset selection object
## Call: regsubsets.formula(G3 ~ ., data = school, nvmax = 8, method = "forward")
## 39 Variables  (and intercept)
##                   Forced in Forced out
## schoolMS             FALSE      FALSE
## sexM                 FALSE      FALSE
## age                  FALSE      FALSE
## addressU             FALSE      FALSE
## famsizeLE3           FALSE      FALSE
## PstatusT             FALSE      FALSE
## Medu                 FALSE      FALSE
## Fedu                 FALSE      FALSE
## Mjobhealth           FALSE      FALSE
## Mjobother            FALSE      FALSE
## Mjobservices         FALSE      FALSE
## Mjobteacher          FALSE      FALSE
## Fjobhealth           FALSE      FALSE
## Fjobother            FALSE      FALSE
## Fjobservices         FALSE      FALSE
## Fjobteacher          FALSE      FALSE
## reasonhome           FALSE      FALSE
## reasonother          FALSE      FALSE
## reasonreputation     FALSE      FALSE
## guardianmother       FALSE      FALSE
## guardianother        FALSE      FALSE
## traveltime           FALSE      FALSE
## studytime            FALSE      FALSE
## failures             FALSE      FALSE
## schoolsupyes         FALSE      FALSE
## famsupyes            FALSE      FALSE
## paidyes              FALSE      FALSE
## activitiesyes        FALSE      FALSE
## nurseryyes           FALSE      FALSE
## higheryes            FALSE      FALSE
## internetyes          FALSE      FALSE
## romanticyes          FALSE      FALSE
## famrel               FALSE      FALSE
## freetime             FALSE      FALSE
## goout                FALSE      FALSE
## Dalc                 FALSE      FALSE
## Walc                 FALSE      FALSE
## health               FALSE      FALSE
## absences             FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          schoolMS sexM age addressU famsizeLE3 PstatusT Medu Fedu
## 1  ( 1 ) " "      " "  " " " "      " "        " "      " "  " " 
## 2  ( 1 ) "*"      " "  " " " "      " "        " "      " "  " " 
## 3  ( 1 ) "*"      " "  " " " "      " "        " "      " "  " " 
## 4  ( 1 ) "*"      " "  " " " "      " "        " "      " "  " " 
## 5  ( 1 ) "*"      " "  " " " "      " "        " "      " "  " " 
## 6  ( 1 ) "*"      " "  " " " "      " "        " "      " "  " " 
## 7  ( 1 ) "*"      " "  " " " "      " "        " "      "*"  " " 
## 8  ( 1 ) "*"      "*"  " " " "      " "        " "      "*"  " " 
##          Mjobhealth Mjobother Mjobservices Mjobteacher Fjobhealth
## 1  ( 1 ) " "        " "       " "          " "         " "       
## 2  ( 1 ) " "        " "       " "          " "         " "       
## 3  ( 1 ) " "        " "       " "          " "         " "       
## 4  ( 1 ) " "        " "       " "          " "         " "       
## 5  ( 1 ) " "        " "       " "          " "         " "       
## 6  ( 1 ) " "        " "       " "          " "         " "       
## 7  ( 1 ) " "        " "       " "          " "         " "       
## 8  ( 1 ) " "        " "       " "          " "         " "       
##          Fjobother Fjobservices Fjobteacher reasonhome reasonother
## 1  ( 1 ) " "       " "          " "         " "        " "        
## 2  ( 1 ) " "       " "          " "         " "        " "        
## 3  ( 1 ) " "       " "          " "         " "        " "        
## 4  ( 1 ) " "       " "          " "         " "        " "        
## 5  ( 1 ) " "       " "          " "         " "        " "        
## 6  ( 1 ) " "       " "          " "         " "        " "        
## 7  ( 1 ) " "       " "          " "         " "        " "        
## 8  ( 1 ) " "       " "          " "         " "        " "        
##          reasonreputation guardianmother guardianother traveltime
## 1  ( 1 ) " "              " "            " "           " "       
## 2  ( 1 ) " "              " "            " "           " "       
## 3  ( 1 ) " "              " "            " "           " "       
## 4  ( 1 ) " "              " "            " "           " "       
## 5  ( 1 ) " "              " "            " "           " "       
## 6  ( 1 ) " "              " "            " "           " "       
## 7  ( 1 ) " "              " "            " "           " "       
## 8  ( 1 ) " "              " "            " "           " "       
##          studytime failures schoolsupyes famsupyes paidyes activitiesyes
## 1  ( 1 ) " "       "*"      " "          " "       " "     " "          
## 2  ( 1 ) " "       "*"      " "          " "       " "     " "          
## 3  ( 1 ) " "       "*"      " "          " "       " "     " "
## 4  ( 1 ) " "       "*"      " "          " "       " "     " "
## 5  ( 1 ) " "       "*"      "*"          " "       " "     " "
## 6  ( 1 ) "*"       "*"      "*"          " "       " "     " "
## 7  ( 1 ) "*"       "*"      "*"          " "       " "     " "
## 8  ( 1 ) "*"       "*"      "*"          " "       " "     " "
##          nurseryyes higheryes internetyes romanticyes famrel freetime
## 1  ( 1 ) " "        " "       " "         " "         " "    " "
## 2  ( 1 ) " "        " "       " "         " "         " "    " "
## 3  ( 1 ) " "        "*"       " "         " "         " "    " "
## 4  ( 1 ) " "        "*"       " "         " "         " "    " "
## 5  ( 1 ) " "        "*"       " "         " "         " "    " "
## 6  ( 1 ) " "        "*"       " "         " "         " "    " "
## 7  ( 1 ) " "        "*"       " "         " "         " "    " "
## 8  ( 1 ) " "        "*"       " "         " "         " "    " "
##          goout Dalc Walc health absences
## 1  ( 1 ) " "   " "  " "  " "    " "
## 2  ( 1 ) " "   " "  " "  " "    " "
## 3  ( 1 ) " "   " "  " "  " "    " "
## 4  ( 1 ) " "   "*"  " "  " "    " "
## 5  ( 1 ) " "   "*"  " "  " "    " "
## 6  ( 1 ) " "   "*"  " "  " "    " "
## 7  ( 1 ) " "   "*"  " "  " "    " "
## 8  ( 1 ) " "   "*"  " "  " "    " "
```

```r
plot(fwd_step)
```



#h)

```r
set.seed(2019)
train_index <- sample(1:nrow(school),size = .75*nrow(school),replace = FALSE)
train_school <- school[train_index,]
test_school <- school[-train_index,]
####estimate model of c and g on train
#from c
train_lm <- lm(G3~factor(higher)+Medu+studytime + failures +factor(school),
               data = train_school)
#from g
train_fwd_lm <- lm(G3~factor(school) + sex + Medu + studytime +
                     failures + factor(higher) + schoolsup + Dalc,
                   data = train_school)

train_lm_preds <-  predict(train_lm)
train_fwd_lm_preds <- predict(train_fwd_lm)
#testpreds simple linear
preds_test_lm <- predict(train_lm,newdata = test_school)
preds_test_fwd_lm <- predict(train_fwd_lm,newdata = test_school)

####residuals
reg_lm_trainMSE <- (RMSE(train_school$G3,train_lm_preds))^2
fwd_lm_trainMSE <- (RMSE(train_school$G3,train_fwd_lm_preds))^2

reg_lm_testMSE <- (RMSE(test_school$G3,preds_test_lm))^2
fwd_lm_testMSE <- (RMSE(test_school$G3,preds_test_fwd_lm))^2
```

Regular LM MSE for train and test

```r
print(reg_lm_trainMSE)
```

```
## [1] 6.41402
```

```r
print(reg_lm_testMSE)
```

```
## [1] 9.076875
```

Forward Stepwise MSE for train and test

```r
print(fwd_lm_trainMSE)
```

```
## [1] 6.231283
```

```r
print(fwd_lm_testMSE)
```

```
## [1] 8.399581
```

#j)

```r
library(glmnet)
```

```
## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-18
```

```r
library(glmnetUtils)
```

```
##
```

```
## Attaching package: 'glmnetUtils'
```

```
## The following objects are masked from 'package:glmnet':
##
##     cv.glmnet, glmnet
```

```
lasso <- cv.glmnet(G3~.,data = train_school,
                   alpha = 1)
coef(lasso)
```

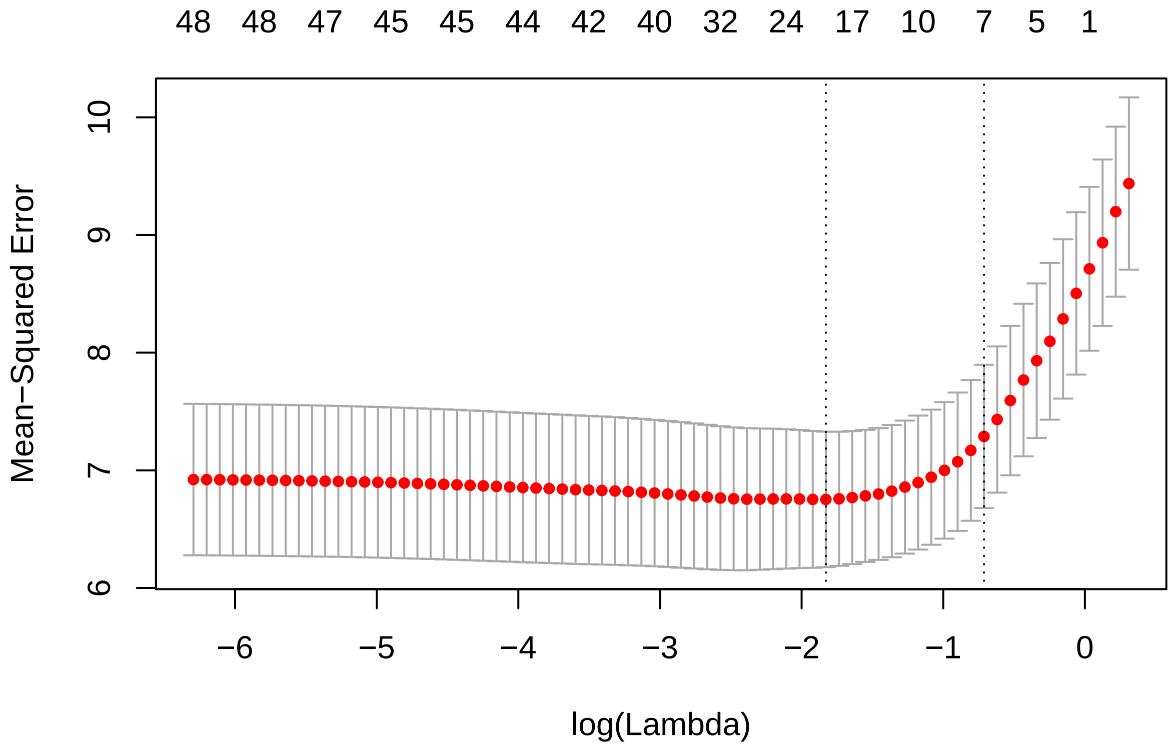
```
## 57 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)       1.176045e+01
## schoolGP          3.202541e-01
## schoolMS         -6.055528e-14
## sexF              .
## sexM              .
## age               .
## addressR          .
## addressU          .
## famsizeGT3        .
## famsizeLE3        .
## PstatusA          .
## PstatusT          .
## Medu              1.010242e-01
## Fedu              .
## Mjobat_home       .
## Mjobhealth        .
## Mjobother         .
## Mjobservices      .
## Mjobteacher       .
## Fjobat_home       .
## Fjobhealth        .
## Fjobother         .
## Fjobservices      .
## Fjobteacher       .
## reasoncourse      .
## reasonhome        .
## reasonother       .
## reasonreputation  .
## guardianfather    .
## guardianmother    .
## guardianother     .
## traveltime        .
## studytime         1.265080e-01
## failures         -1.138555e+00
## schoolsupno       .
## schoolsupyes      .
## famsupno          .
## famsupyes         .
## paidno            .
## paidyes           .
## activitiesno      .
## activitiesyes     .
## nurseryno         .
## nurseryyes        .
## higherno         -1.019480e+00
## higheryes         2.016043e-13
## internetno        .
## internetyes       .
## romanticno        .
## romanticyes       .
## famrel            .
## freetime          .
## goout             .
## Dalc              .
## Walc              .
## health            .
## absences          .
```

```
coef_min_matrix <- as.matrix(coef(lasso,s = lasso$lambda.min))
coef_1se_matrix <- as.matrix(coef(lasso,s = lasso$lambda.1se))
coef_df <- data.frame(Lambda_min_coefs = coef_min_matrix,
                      Lambda_1se_coefs = coef_1se_matrix)
colnames(coef_df) <- c("Lambda_Min_coefs","Lambda_1se_coefs")
print(coef_df)
```

```
##                    Lambda_Min_coefs Lambda_1se_coefs
## (Intercept)             1.077792e+01     1.176045e+01
## schoolGP                7.096478e-01     3.202541e-01
## schoolMS               -4.840306e-13    -6.055528e-14
## sexF                    2.722645e-01     0.000000e+00
## sexM                    0.000000e+00     0.000000e+00
## age                     0.000000e+00     0.000000e+00
## addressR               -1.797094e-01     0.000000e+00
## addressU                8.906097e-15     0.000000e+00
## famsizeGT3              0.000000e+00     0.000000e+00
## famsizeLE3              0.000000e+00     0.000000e+00
## PstatusA               -2.723748e-02     0.000000e+00
## PstatusT                0.000000e+00     0.000000e+00
## Medu                    2.362895e-01     1.010242e-01
## Fedu                    0.000000e+00     0.000000e+00
## Mjobat_home             0.000000e+00     0.000000e+00
## Mjobhealth              2.657887e-03     0.000000e+00
## Mjobother               0.000000e+00     0.000000e+00
## Mjobservices            0.000000e+00     0.000000e+00
## Mjobteacher             0.000000e+00     0.000000e+00
## Fjobat_home             0.000000e+00     0.000000e+00
## Fjobhealth              0.000000e+00     0.000000e+00
## Fjobother               0.000000e+00     0.000000e+00
## Fjobservices            0.000000e+00     0.000000e+00
## Fjobteacher             5.756096e-01     0.000000e+00
## reasoncourse            0.000000e+00     0.000000e+00
## reasonhome              0.000000e+00     0.000000e+00
## reasonother             0.000000e+00     0.000000e+00
## reasonreputation        5.279789e-02     0.000000e+00
## guardianfather          0.000000e+00     0.000000e+00
## guardianmother          0.000000e+00     0.000000e+00
## guardianother           0.000000e+00     0.000000e+00
## traveltime              0.000000e+00     0.000000e+00
## studytime            3.354026e-01     1.265080e-01
## failures            -1.391685e+00    -1.138555e+00
## schoolsupno          3.290068e-01     0.000000e+00
## schoolsupyes        -3.244346e-14     0.000000e+00
## famsupno             0.000000e+00     0.000000e+00
## famsupyes            0.000000e+00     0.000000e+00
## paidno               0.000000e+00     0.000000e+00
## paidyes              0.000000e+00     0.000000e+00
## activitiesno         0.000000e+00     0.000000e+00
## activitiesyes        0.000000e+00     0.000000e+00
## nurseryno            0.000000e+00     0.000000e+00
## nurseryyes           0.000000e+00     0.000000e+00
## higherno            -1.485225e+00    -1.019480e+00
## higheryes            9.031735e-13     2.016043e-13
## internetno          -8.574115e-02     0.000000e+00
## internetyes          0.000000e+00     0.000000e+00
## romanticno           3.141639e-03     0.000000e+00
## romanticyes          0.000000e+00     0.000000e+00
## famrel               0.000000e+00     0.000000e+00
## freetime             0.000000e+00     0.000000e+00
## goout               -5.544507e-02     0.000000e+00
## Dalc                 0.000000e+00     0.000000e+00
## Walc                -6.929914e-02     0.000000e+00
## health               0.000000e+00     0.000000e+00
## absences             0.000000e+00     0.000000e+00
```

k)

```
plot(lasso)
```



```
print(lasso$lambda.1se)
```

```
## [1] 0.4905485
```

```
print(lasso$lambda.min)
```

```
## [1] 0.1606325
```

#Question 2

```
credit <- read.csv("gmsc_cs-training.csv")


dim(credit)
```

```
## [1] 99998    12
```

```
sum(complete.cases(credit))
```

```
## [1] 80186
```

```
credit <- credit[complete.cases(credit),]
nrow(credit)
```

```
## [1] 80186
```

```
sum(is.na(credit))
```

```
## [1] 0
```

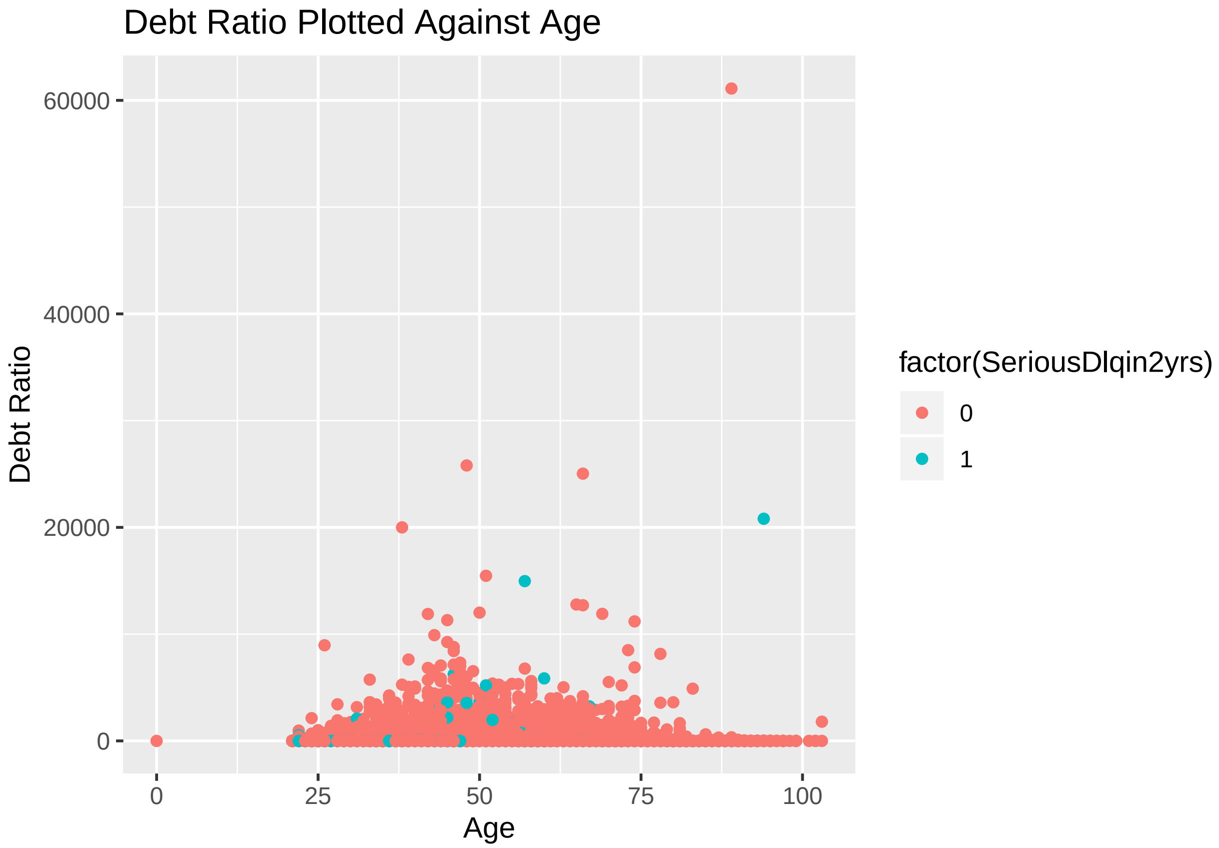
```
names(credit)
```

```
##  [1] "X"
##  [2] "SeriousDlqin2yrs"
##  [3] "RevolvingUtilizationOfUnsecuredLines"
##  [4] "age"
##  [5] "NumberOfTime30.59DaysPastDueNotWorse"
##  [6] "DebtRatio"
##  [7] "MonthlyIncome"
##  [8] "NumberOfOpenCreditLinesAndLoans"
##  [9] "NumberOfTimes90DaysLate"
## [10] "NumberRealEstateLoansOrLines"
## [11] "NumberOfTime60.89DaysPastDueNotWorse"
## [12] "NumberOfDependents"
```

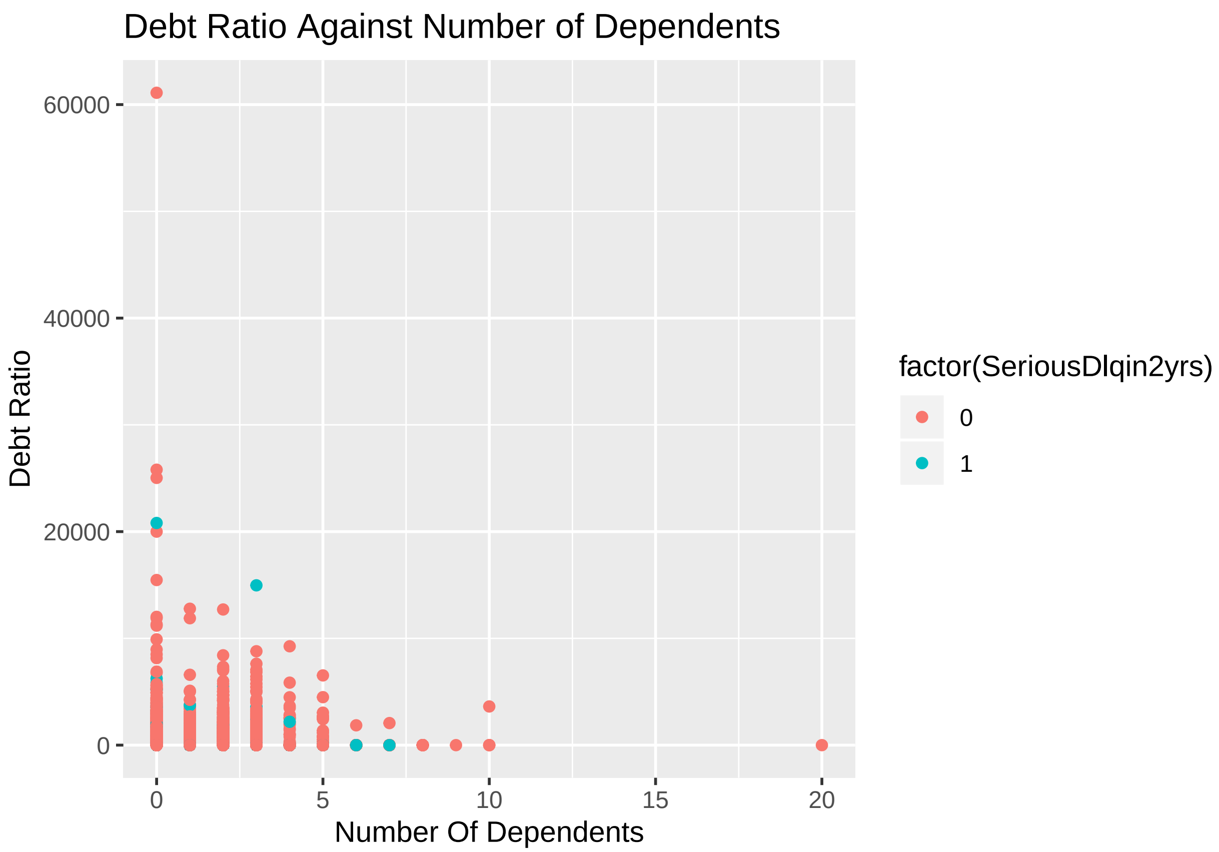
#a)

```
trainindex <- sample(1:nrow(credit),.75*nrow(credit), replace = FALSE)
credit_train <- credit[trainindex,]
credit_test <- credit[-trainindex,]


ggplot(credit,aes(x = age, y = DebtRatio)) + geom_point(aes(color = factor(SeriousDlqin2yrs)))+
  labs(x = "Age",y = "Debt Ratio",title = "Debt Ratio Plotted Against Age")
```



```
ggplot(credit,aes(x = NumberOfDependents,y = DebtRatio)) +
  geom_point(aes(color = factor(SeriousDlqin2yrs))) +
  labs(x = "Number Of Dependents",y = "Debt Ratio",
       title = "Debt Ratio Against Number of Dependents")
```



#b)

```
cors <- cor(credit$SeriousDlqin2yrs,credit)
##to get top 4 correlated
tail(sort(abs(cors)),5)
```

```
## [1] 0.08538574 0.10226590 0.10227154 0.11437236 1.00000000
```

```
logitMod_train <- glm(SeriousDlqin2yrs ~ age+
                        NumberOfTime60.89DaysPastDueNotWorse +
                        NumberOfTimes90DaysLate +
                        NumberOfTime30.59DaysPastDueNotWorse,
                      data = credit_train,
                      family = "binomial")
summary(logitMod_train)
```

```
##
## Call:
## glm(formula = SeriousDlqin2yrs ~ age + NumberOfTime60.89DaysPastDueNotWorse +
##     NumberOfTimes90DaysLate + NumberOfTime30.59DaysPastDueNotWorse,
##     family = "binomial", data = credit_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2497  -0.3934  -0.3291  -0.2747   4.2312
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -1.441049   0.060637  -23.77   <2e-16
## age                                 -0.028392   0.001244  -22.83   <2e-16
## NumberOfTime60.89DaysPastDueNotWorse -0.951048   0.027647  -34.40   <2e-16
## NumberOfTimes90DaysLate              0.474323   0.023779   19.95   <2e-16
## NumberOfTime30.59DaysPastDueNotWorse  0.509910   0.016668   30.59   <2e-16
##
## (Intercept)                         ***
## age                                 ***
## NumberOfTime60.89DaysPastDueNotWorse ***
## NumberOfTimes90DaysLate             ***
## NumberOfTime30.59DaysPastDueNotWorse ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30301  on 60138  degrees of freedom
## Residual deviance: 28107  on 60134  degrees of freedom
## AIC: 28117
##
## Number of Fisher Scoring iterations: 6
```

#c)

```r
exp(logitMod_train$coefficients)
```

```
##                          (Intercept)                                  age
##                            0.2366793                            0.9720072
## NumberOfTime60.89DaysPastDueNotWorse              NumberOfTimes90DaysLate
##                            0.3863360                            1.6069266
## NumberOfTime30.59DaysPastDueNotWorse
##                            1.6651412
```

#d)

```r
train_preds_df <- predict(logitMod_train, type = "response")
test_preds_df <- predict(logitMod_train,newdata =  credit_test,type = "response")
credit_train$delinqScores <- train_preds_df
credit_test$delinqScores <- test_preds_df

credit_train$preds05 <- ifelse(credit_train$delinqScores >.5,1,0)
credit_train$preds07 <- ifelse(credit_train$delinqScores > .7,1,0)
credit_test$preds05 <- ifelse(credit_test$delinqScores >.5,1,0)
credit_test$preds07 <- ifelse(credit_test$delinqScores >.7,1,0)

library(gmodels)
###cutoff 50 train
CrossTable(credit_train$SeriousDlqin2yrs,credit_train$preds05,
           prop.r = FALSE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
## 
## 
## Total Observations in Table:  60139 
## 
## 
##                               | credit_train$preds05 
## credit_train$SeriousDlqin2yrs |         0 |         1 | Row Total | 
## ------------------------------|-----------|-----------|-----------|
##                             0 |     55853 |       116 |     55969 | 
## ------------------------------|-----------|-----------|-----------|
##                             1 |      4024 |       146 |      4170 | 
## ------------------------------|-----------|-----------|-----------|
##                  Column Total |     59877 |       262 |     60139 | 
## ------------------------------|-----------|-----------|-----------|
## 
## 
```

```
#cutoff 70 train
CrossTable(credit_train$SeriousDlqin2yrs,credit_train$preds07,
           prop.r = FALSE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
## 
## 
## Total Observations in Table:  60139 
## 
## 
##                               | credit_train$preds07 
## credit_train$SeriousDlqin2yrs |         0 |         1 | Row Total | 
## ------------------------------|-----------|-----------|-----------|
##                             0 |     55920 |        49 |     55969 | 
## ------------------------------|-----------|-----------|-----------|
##                             1 |      4118 |        52 |      4170 | 
## ------------------------------|-----------|-----------|-----------|
##                  Column Total |     60038 |       101 |     60139 | 
## ------------------------------|-----------|-----------|-----------|
## 
## 
```

```
###cutoff 50 test
CrossTable(credit_test$SeriousDlqin2yrs,credit_test$preds05,
           prop.r = FALSE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
## 
## 
## Total Observations in Table:  20047 
## 
## 
##                              | credit_test$preds05 
## credit_test$SeriousDlqin2yrs |         0 |         1 | Row Total | 
## -----------------------------|-----------|-----------|-----------|
##                            0 |     18647 |        49 |     18696 | 
## -----------------------------|-----------|-----------|-----------|
##                            1 |      1292 |        59 |      1351 | 
## -----------------------------|-----------|-----------|-----------|
##                 Column Total |     19939 |       108 |     20047 | 
## -----------------------------|-----------|-----------|-----------|
## 
## 
```

```
#cutoff 70 test
CrossTable(credit_test$SeriousDlqin2yrs,credit_test$preds07,
           prop.r = FALSE,
           prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
## 
## 
## Total Observations in Table:  20047 
## 
## 
##                              | credit_test$preds07 
## credit_test$SeriousDlqin2yrs |         0 |         1 | Row Total | 
## -----------------------------|-----------|-----------|-----------|
##                            0 |     18673 |        23 |     18696 | 
## -----------------------------|-----------|-----------|-----------|
##                            1 |      1327 |        24 |      1351 | 
## -----------------------------|-----------|-----------|-----------|
##                 Column Total |     20000 |        47 |     20047 | 
## -----------------------------|-----------|-----------|-----------|
##
##
```

f)

```
library(plotROC)
roc_plot_delinq <- ggplot(credit_test,aes(m = delinqScores,
                                          d = SeriousDlqin2yrs)) +
  geom_roc(labelsize = 3.5,
           cutoffs.at = c(.9,.7,.6,.5,.4,.3,.2,.1)) +
  labs(title = "ROC Curve for Test Data Logit Model",x = "False Positive Fraction",
       y= "True Positive Fraction")
roc_plot_delinq_train <- ggplot(credit_train,aes(m = delinqScores,
                                                 d = SeriousDlqin2yrs)) +
  geom_roc(labelsize = 3.5,
           cutoffs.at = c(.9,.7,.6,.5,.4,.3,.2,.1)) +
  labs(title = "ROC Curve for Test Data Logit Model",x = "False Positive Fraction",
       y= "True Positive Fraction")
```

g)

```
calc_auc(roc_plot_delinq)
```

```
##   PANEL group       AUC
## 1     1    -1 0.6742182
```

```
calc_auc(roc_plot_delinq_train)
```

```
##   PANEL group       AUC
## 1     1    -1 0.6830092
```