# Final Project

London Wagner, Erik Lovece, Carmen Canedo

2021-04-24

# 1 Introduction

Our final project analyzes the student performance dataset from the UCI Machine Learning Repository, originally gathered by Paulo Cortez from the University of Minho. This dataset measures the final student grade in a Portuguese class based on a variety of predictors. These predictors cover numerous aspects of not only students' academic lives, but also family life predictors such as parental employment, and personal predictors like whether or not they have home internet access and whether or not they are in a romantic relationship.

We seek to answer the question of what predictors have the greatest influence in how a student does in class. Conventional wisdom seems to dictate that high-achieving students have come from particularly favorable academic, filial, and personal environments, and previous studies have confirmed this. Our model, if properly constructed along the best machine learning practices, should corroborate this, although unexpected conclusions may also lie in store.

Our workflow for finding a sufficient model from which we will draw our conclusions is as follows:

1) Run a linear regression model with final grade as the response and all other variables as predictors
2) Use best subset, forward step, and backwards step to select variables for a reduced model
3) Use ridge and lasso to conduct further dimension reduction
4) Use cross validation methods to determine which model predicts the final grade with the greatest accuracy
5) Make more definitive determinations based on the chosen model.

# 2 Loading & Cleaning Data

```
student_por <- read_csv2("data/student-por.csv")

student_por
```

```
## # A tibble: 649 x 33
##    school sex     age address famsize Pstatus  Medu  Fedu Mjob    Fjob   reason
##    <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>   <chr>  <chr>
##  1 GP     F        18 U       GT3     A           4     4 at_home teach~ course
##  2 GP     F        17 U       GT3     T           1     1 at_home other  course
##  3 GP     F        15 U       LE3     T           1     1 at_home other  other
##  4 GP     F        15 U       GT3     T           4     2 health  servi~ home
##  5 GP     F        16 U       GT3     T           3     3 other   other  home
```

```
##  6 GP     M        16 U       LE3     T           4     3 servic~ other  reputa~
##  7 GP     M        16 U       LE3     T           2     2 other   other  home
##  8 GP     F        17 U       GT3     A           4     4 other   teach~ home
##  9 GP     M        15 U       LE3     A           3     2 servic~ other  home
## 10 GP     M        15 U       GT3     T           3     4 other   other  home
## # ... with 639 more rows, and 22 more variables: guardian <chr>,
## #   traveltime <dbl>, studytime <dbl>, failures <dbl>, schoolsup <chr>,
## #   famsup <chr>, paid <chr>, activities <chr>, nursery <chr>, higher <chr>,
## #   internet <chr>, romantic <chr>, famrel <dbl>, freetime <dbl>, goout <dbl>,
## #   Dalc <dbl>, Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>,
## #   G3 <dbl>
```

The student attributes and grades forming the predictors and response, quoted verbatim from a text file provided with the dataset, are as follows:

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2 sex - student's sex (binary: "F" - female or "M" - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: "U" - urban or "R" - rural)

5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

12 guardian - student's guardian (nominal: "mother", "father" or "other")

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

```r
student_por <-
  student_por %>%
  mutate(school = factor(school),
         sex = factor(sex),
         address = factor(address),
         famsize = factor(famsize),
         Pstatus = factor(Pstatus),
         schoolsup = factor(schoolsup),
         famsup = factor(famsup),
         paid = factor(paid),
         activities = factor(activities),
         nursery = factor(nursery),
         higher = factor(higher),
         internet = factor(internet),
         romantic = factor(romantic),
         reason = factor(reason))

student_por
```

```
## # A tibble: 649 x 33
##    school sex     age address famsize Pstatus  Medu  Fedu Mjob    Fjob   reason
##    <fct>  <fct> <dbl> <fct>   <fct>   <fct>   <dbl> <dbl> <chr>   <chr>  <fct>
##  1 GP     F        18 U       GT3     A           4     4 at_home teach~ course
##  2 GP     F        17 U       GT3     T           1     1 at_home other  course
##  3 GP     F        15 U       LE3     T           1     1 at_home other  other
##  4 GP     F        15 U       GT3     T           4     2 health  servi~ home
##  5 GP     F        16 U       GT3     T           3     3 other   other  home
##  6 GP     M        16 U       LE3     T           4     3 servic~ other  reputa~
##  7 GP     M        16 U       LE3     T           2     2 other   other  home
##  8 GP     F        17 U       GT3     A           4     4 other   teach~ home
##  9 GP     M        15 U       LE3     A           3     2 servic~ other  home
## 10 GP     M        15 U       GT3     T           3     4 other   other  home
## # ... with 639 more rows, and 22 more variables: guardian <chr>,
## #   traveltime <dbl>, studytime <dbl>, failures <dbl>, schoolsup <fct>,
## #   famsup <fct>, paid <fct>, activities <fct>, nursery <fct>, higher <fct>,
## #   internet <fct>, romantic <fct>, famrel <dbl>, freetime <dbl>, goout <dbl>,
## #   Dalc <dbl>, Walc <dbl>, health <dbl>, absences <dbl>, G1 <dbl>, G2 <dbl>,
## #   G3 <dbl>
```

# 3 EDA & Checking Assumptions

Before we begin our analysis, we wish to explore the distribution of the data and confirm it follows the typical assumptions of linear regresion.

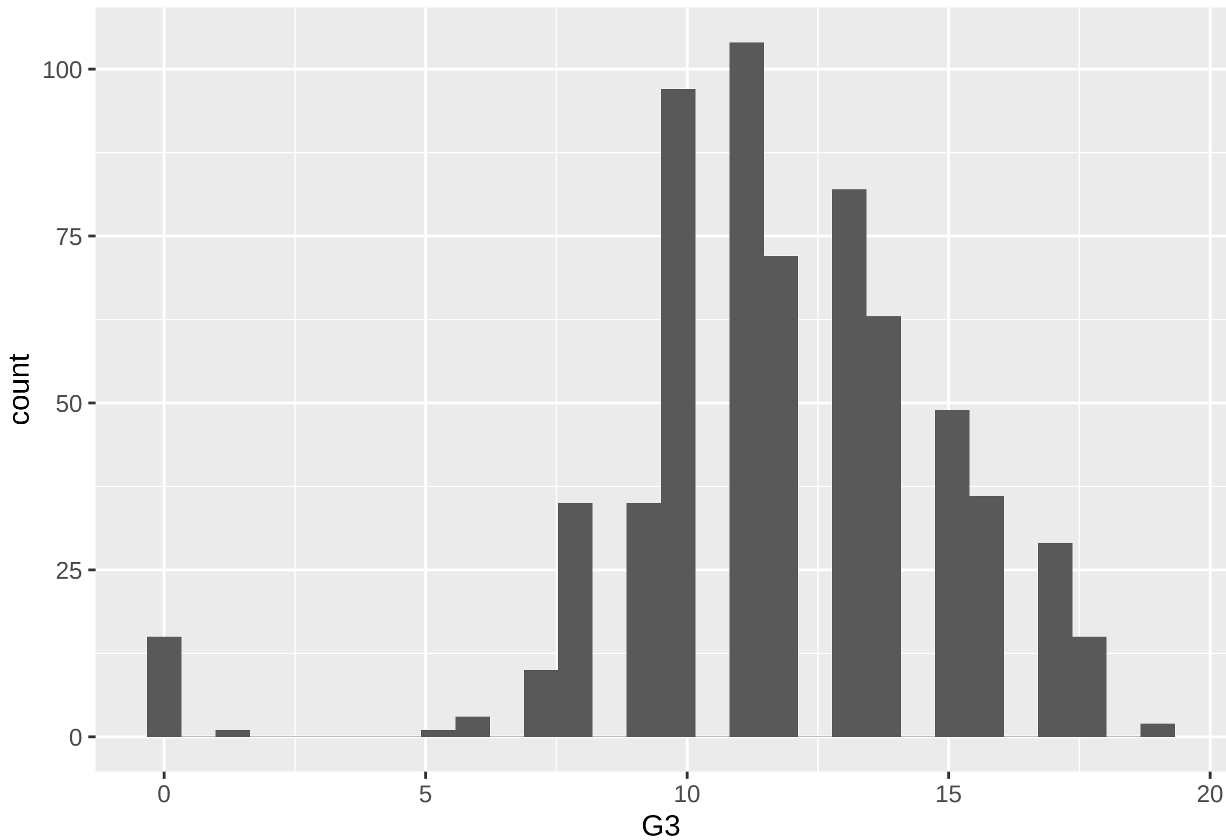
```
summary(student_por)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000
##                   Median :17.00                             Median :2.000
##                   Mean   :16.74                             Mean   :2.515
##                   3rd Qu.:18.00                             3rd Qu.:4.000
##                   Max.   :22.00                             Max.   :4.000
##      Fedu           Mjob               Fjob               reason
##  Min.   :0.000   Length:649         Length:649         course    :285
##  1st Qu.:1.000   Class :character   Class :character   home      :149
##  Median :2.000   Mode  :character   Mode  :character   other     : 72
##  Mean   :2.307                                         reputation:143
##  3rd Qu.:3.000
##  Max.   :4.000
##    guardian           traveltime      studytime       failures      schoolsup
##  Length:649         Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581
##  Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68
##  Mode  :character   Median :1.000   Median :2.000   Median :0.0000
##                     Mean   :1.569   Mean   :1.931   Mean   :0.2219
##                     3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##                     Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  famsup     paid     activities nursery   higher    internet  romantic
##  no :251   no :610   no :334    no :128   no : 69   no :151   no :410
##  yes:398   yes: 39   yes:315    yes:521   yes:580   yes:498   yes:239
##
##
##
##
##      famrel         freetime        goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:4.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00
##  Median :4.000   Median :3.00   Median :3.000   Median :1.000   Median :2.00
##  Mean   :3.931   Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28
##  3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00
##  Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00
##      health         absences            G1             G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 0.0   Min.   : 0.00
##  1st Qu.:2.000   1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00
##  Median :4.000   Median : 2.000   Median :11.0   Median :11.00
##  Mean   :3.536   Mean   : 3.659   Mean   :11.4   Mean   :11.57
##  3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00
##  Max.   :5.000   Max.   :32.000   Max.   :19.0   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.:10.00
##  Median :12.00
```

```
##  Mean   :11.91
##  3rd Qu.:14.00
##  Max.   :19.00
```

```
student_por %>%
  ggplot(aes(x = G3)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




The histogram above is skewed to the left when run our initial linear regression, we will check the residuals and QQ plots.

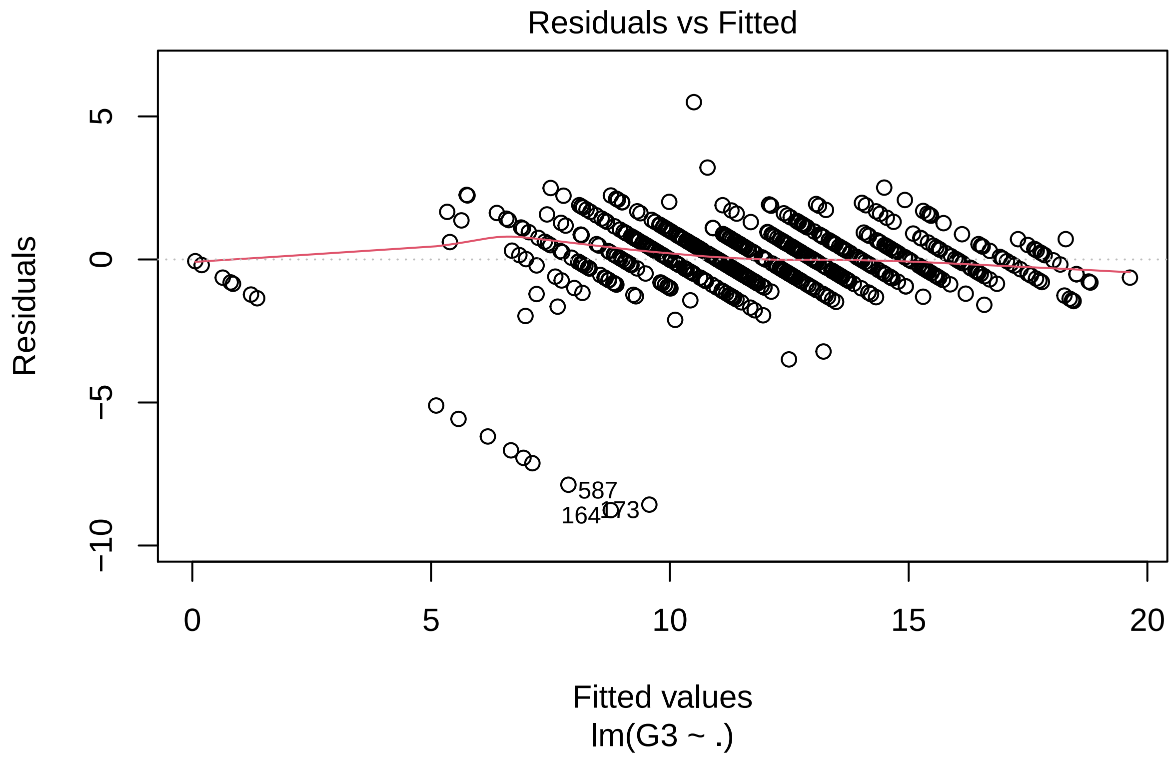
## 3.1 Running a simple linear regression

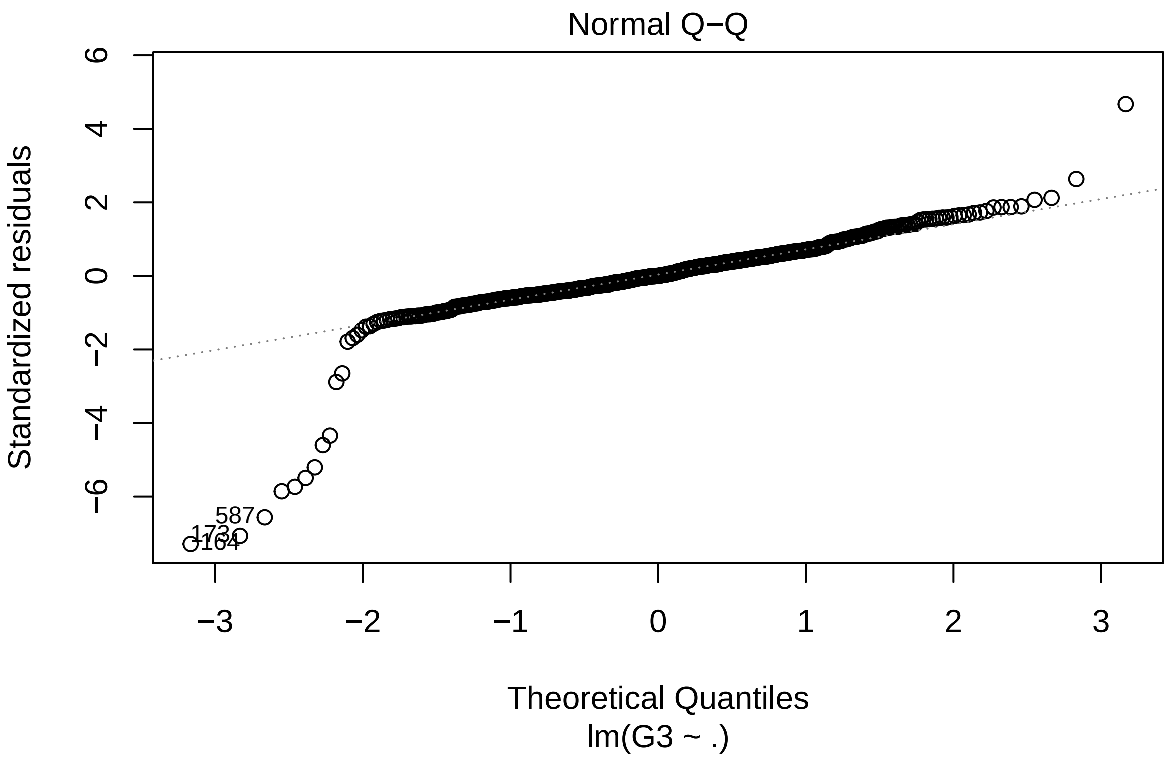
```
por_reg <- lm(G3 ~ ., data = student_por)
summary(por_reg)
```

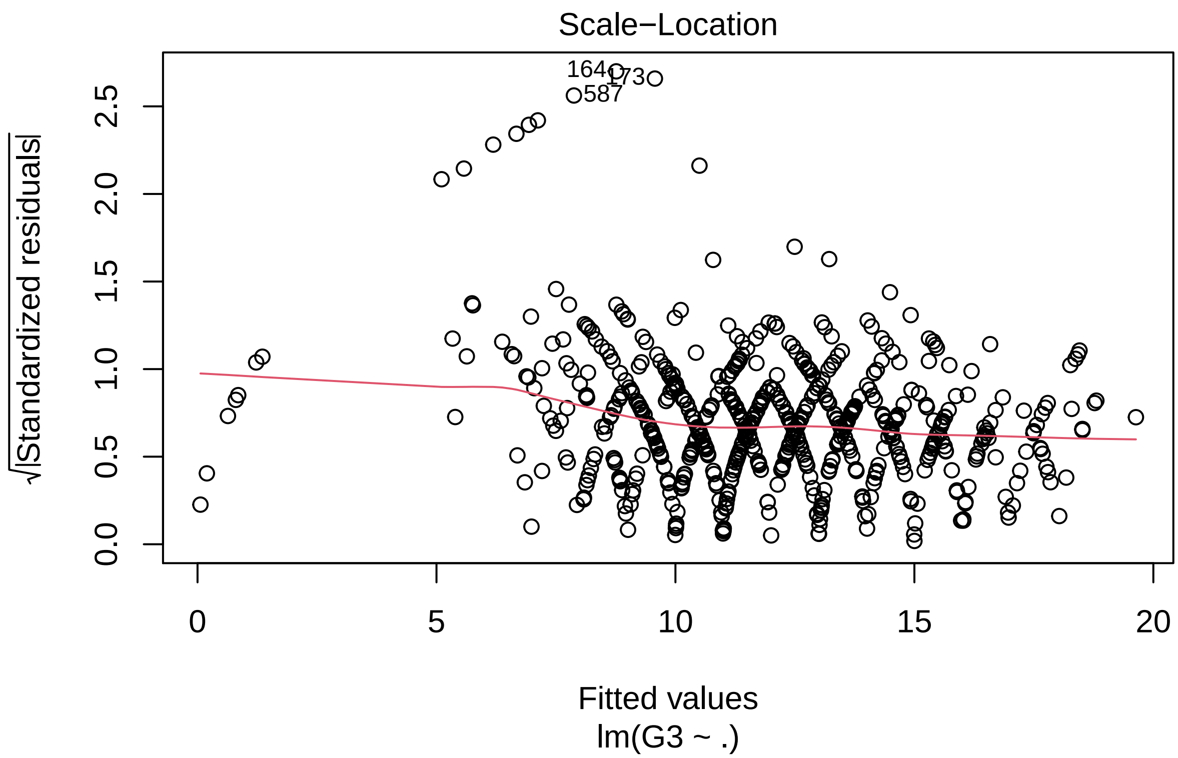
```
##
## Call:
## lm(formula = G3 ~ ., data = student_por)
##
## Residuals:
```

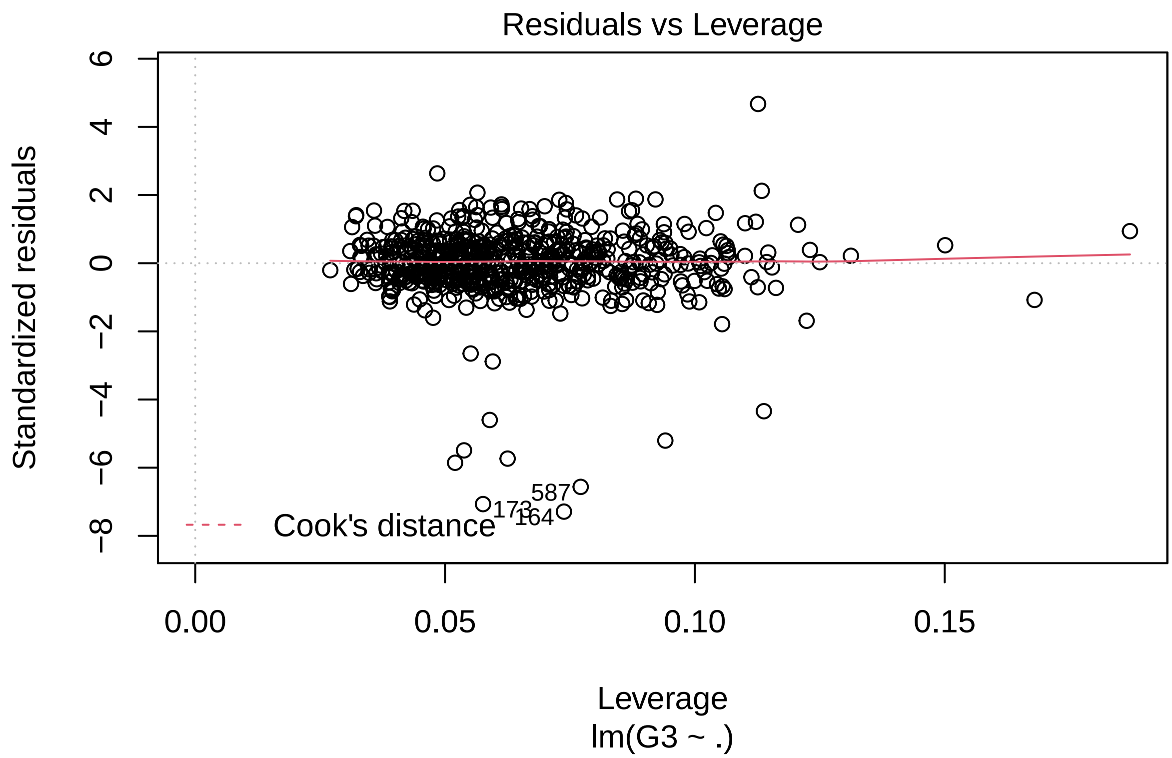
```
##     Min      1Q  Median      3Q     Max 
## -8.7618 -0.5148  0.0038  0.6047  5.4973 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       0.63823    0.96361   0.662 0.508011    
## schoolMS         -0.19797    0.12783  -1.549 0.121992    
## sexM             -0.12258    0.11778  -1.041 0.298423    
## age               0.02869    0.04835   0.593 0.553208    
## addressU          0.11446    0.12277   0.932 0.351565    
## famsizeLE3        0.01560    0.11505   0.136 0.892197    
## PstatusT         -0.09746    0.16256  -0.600 0.549055    
## Medu             -0.09170    0.07097  -1.292 0.196799    
## Fedu              0.04962    0.06461   0.768 0.442773    
## Mjobhealth        0.26583    0.25225   1.054 0.292379    
## Mjobother        -0.09351    0.14208  -0.658 0.510720    
## Mjobservices      0.17255    0.17510   0.985 0.324808    
## Mjobteacher       0.22115    0.23558   0.939 0.348232    
## Fjobhealth       -0.44420    0.35256  -1.260 0.208189    
## Fjobother        -0.33805    0.21391  -1.580 0.114544    
## Fjobservices     -0.47121    0.22477  -2.096 0.036457 *  
## Fjobteacher      -0.54368    0.31611  -1.720 0.085958 .  
## reasonhome       -0.07885    0.13366  -0.590 0.555479    
## reasonother      -0.36174    0.17236  -2.099 0.036251 *  
## reasonreputation -0.16934    0.13990  -1.210 0.226584    
## guardianmother   -0.02513    0.12461  -0.202 0.840252    
## guardianother     0.21732    0.24922   0.872 0.383539    
## traveltime        0.13859    0.07459   1.858 0.063667 .  
## studytime         0.04965    0.06620   0.750 0.453569    
## failures         -0.25495    0.09900  -2.575 0.010254 *  
## schoolsupyes     -0.18419    0.17319  -1.064 0.287969    
## famsupyes         0.09456    0.10701   0.884 0.377230    
## paidyes          -0.19166    0.21664  -0.885 0.376663    
## activitiesyes     0.01208    0.10482   0.115 0.908275    
## nurseryyes       -0.09562    0.12722  -0.752 0.452553    
## higheryes         0.20749    0.18261   1.136 0.256285    
## internetyes       0.08517    0.12955   0.657 0.511152    
## romanticyes      -0.04209    0.10786  -0.390 0.696483    
## famrel           -0.01597    0.05471  -0.292 0.770469    
## freetime         -0.05002    0.05267  -0.950 0.342694    
## goout            -0.01889    0.05041  -0.375 0.708033    
## Dalc             -0.05194    0.07185  -0.723 0.469977    
## Walc             -0.01693    0.05553  -0.305 0.760521    
## health           -0.05522    0.03633  -1.520 0.129064    
## absences          0.01359    0.01173   1.158 0.247198    
## G1                0.12933    0.03762   3.438 0.000626 ***
## G2                0.87037    0.03495  24.906  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.249 on 607 degrees of freedom
## Multiple R-squared:   0.86,  Adjusted R-squared:  0.8506 
## F-statistic: 90.95 on 41 and 607 DF,  p-value: < 2.2e-16
```

```r
#par(mfrow = c(2, 2))
plot(por_reg)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(G3 ~ .)
587
173
164
−6
−4
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
164
173
587
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20
Fitted values
lm(G3 ~ .)




From our results, we can conclude that a simple linear regression might not be the appropriate form of analsis for this data set. We will continue to explore cross validation techniques to compare its prediction accuracy to other models.

**We should write out the model here**

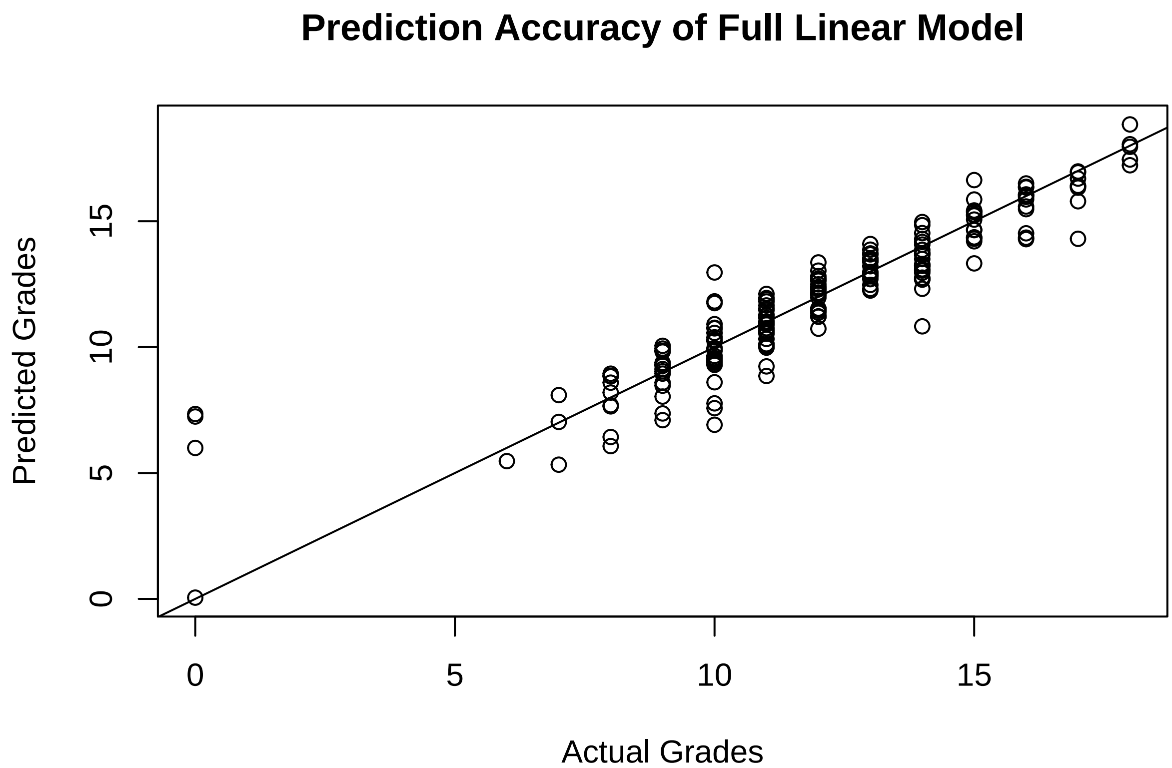
### 3.2 Validation Set Technique on Initial Regression

```
set.seed(1)
n <- nrow(student_por)
Z <- sample(n, .7*n)

reg.fit <- lm(G3 ~ ., data = student_por, subset = Z)
```

```
g3_predicted <- predict(reg.fit, student_por)
```

```
plot(student_por$G3[-Z], g3_predicted[-Z], xlab = "Actual Grades", ylab = "Predicted Grades", main = "P
abline(0,1)
```




```r
mse_lm <- mean((student_por$G3 - g3_predicted)[-Z]^2)
```

A lot of the predictors do not appear to be significant, so we are going to use some variable selection methods to simplify the model while still maintaining accuracy. We hope to generate new models that have lower prediction MSE than this linear model.

### 3.3 Best Subset

```r
# Takes a while to run

subsets <- regsubsets(G3 ~ ., data = student_por, nvmax = 15)
```

```r
summary(subsets)
```

```
## Subset selection object
## Call: regsubsets.formula(G3 ~ ., data = student_por, nvmax = 15)
## 41 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
## G1                   FALSE      FALSE
## G2                   FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##            schoolMS sexM age addressU famsizeLE3 PstatusT Medu Fedu Mjobhealth
## 1  ( 1 )  " "      " "  " " " "      " "        " "      " "  " "  " "
## 2  ( 1 )  " "      " "  " " " "      " "        " "      " "  " "  " "
## 3  ( 1 )  " "      " "  " " " "      " "        " "      " "  " "  " "
## 4  ( 1 )  " "      " "  " " " "      " "        " "      " "  " "  " "
## 5  ( 1 )  " "      "*"  " " " "      " "        " "      " "  " "  " "
## 6  ( 1 )  " "      "*"  " " " "      " "        " "      " "  " "  " "
## 7  ( 1 )  "*"      "*"  " " " "      " "        " "      " "  " "  " "
## 8  ( 1 )  "*"      "*"  " " " "      " "        " "      " "  " "  " "
## 9  ( 1 )  " "      "*"  " " "*"      " "        " "      " "  " "  " "
## 10  ( 1 ) " "      "*"  " " "*"      " "        " "      " "  " "  " "
## 11  ( 1 ) " "      "*"  " " "*"      " "        " "      " "  " "  " "
## 12  ( 1 ) " "      "*"  " " "*"      " "        " "      " "  " "  " "
## 13  ( 1 ) "*"      "*"  " " "*"      " "        " "      " "  " "  " "
## 14  ( 1 ) "*"      "*"  " " "*"      " "        " "      " "  " "  " "
## 15  ( 1 ) "*"      "*"  " " "*"      " "        " "      " "  " "  " "
```

```
##           Mjobother Mjobservices Mjobteacher Fjobhealth Fjobother Fjobservices
## 1  ( 1 )  " "       " "          " "         " "        " "       " "
## 2  ( 1 )  " "       " "          " "         " "        " "       " "
## 3  ( 1 )  " "       " "          " "         " "        " "       " "
## 4  ( 1 )  " "       " "          " "         " "        " "       " "
## 5  ( 1 )  " "       " "          " "         " "        " "       " "
## 6  ( 1 )  " "       " "          " "         " "        " "       " "
## 7  ( 1 )  " "       " "          " "         " "        " "       " "
## 8  ( 1 )  "*"       " "          " "         " "        " "       " "
## 9  ( 1 )  "*"       " "          " "         " "        " "       " "
## 10  ( 1 ) "*"       " "          " "         " "        " "       " "
## 11  ( 1 ) "*"       " "          " "         " "        " "       " "
## 12  ( 1 ) "*"       " "          " "         " "        " "       " "
## 13  ( 1 ) "*"       " "          " "         " "        " "       " "
## 14  ( 1 ) "*"       " "          " "         " "        " "       " "
## 15  ( 1 ) "*"       " "          " "         " "        " "       " "
##           Fjobteacher reasonhome reasonother reasonreputation guardianmother
## 1  ( 1 )  " "         " "        " "         " "              " "
## 2  ( 1 )  " "         " "        " "         " "              " "
## 3  ( 1 )  " "         " "        "*"         " "              " "
## 4  ( 1 )  " "         " "        "*"         " "              " "
## 5  ( 1 )  " "         " "        "*"         " "              " "
## 6  ( 1 )  " "         " "        "*"         " "              " "
## 7  ( 1 )  " "         " "        "*"         " "              " "
## 8  ( 1 )  " "         " "        "*"         " "              " "
## 9  ( 1 )  " "         " "        "*"         " "              " "
## 10  ( 1 ) " "         " "        "*"         " "              " "
## 11  ( 1 ) " "         " "        "*"         " "              " "
## 12  ( 1 ) " "         " "        "*"         " "              " "
## 13  ( 1 ) " "         " "        "*"         " "              " "
## 14  ( 1 ) " "         " "        "*"         " "              " "
## 15  ( 1 ) " "         " "        "*"         " "              " "
##           guardianother traveltime studytime failures schoolsupyes famsupyes
## 1  ( 1 )  " "           " "        " "       " "      " "          " "
## 2  ( 1 )  " "           " "        " "       " "      " "          " "
## 3  ( 1 )  " "           " "        " "       " "      " "          " "
## 4  ( 1 )  " "           " "        " "       "*"      " "          " "
## 5  ( 1 )  " "           " "        " "       "*"      " "          " "
## 6  ( 1 )  " "           " "        " "       "*"      " "          " "
## 7  ( 1 )  " "           "*"        " "       "*"      " "          " "
## 8  ( 1 )  " "           "*"        " "       "*"      " "          " "
## 9  ( 1 )  " "           "*"        " "       "*"      " "          " "
## 10  ( 1 ) "*"           "*"        " "       "*"      " "          " "
## 11  ( 1 ) "*"           "*"        " "       "*"      " "          " "
## 12  ( 1 ) "*"           "*"        " "       "*"      " "          " "
## 13  ( 1 ) "*"           "*"        " "       "*"      " "          " "
## 14  ( 1 ) "*"           "*"        " "       "*"      " "          " "
## 15  ( 1 ) "*"           "*"        " "       "*"      "*"          " "
##           paidyes activitiesyes nurseryyes higheryes internetyes romanticyes
## 1  ( 1 )  " "     " "           " "        " "       " "         " "
## 2  ( 1 )  " "     " "           " "        " "       " "         " "
## 3  ( 1 )  " "     " "           " "        " "       " "         " "
## 4  ( 1 )  " "     " "           " "        " "       " "         " "
## 5  ( 1 )  " "     " "           " "        " "       " "         " "
```

```
## 6  ( 1 )  " "     " "           " "        " "       " "         " "
## 7  ( 1 )  " "     " "           " "        " "       " "         " "
## 8  ( 1 )  " "     " "           " "        " "       " "         " "
## 9  ( 1 )  " "     " "           " "        " "       " "         " "
## 10  ( 1 ) " "     " "           " "        " "       " "         " "
## 11  ( 1 ) " "     " "           " "        " "       " "         " "
## 12  ( 1 ) " "     " "           " "        " "       " "         " "
## 13  ( 1 ) " "     " "           " "        " "       " "         " "
## 14  ( 1 ) " "     " "           " "        "*"       " "         " "
## 15  ( 1 ) " "     " "           " "        "*"       " "         " "
##           famrel freetime goout Dalc Walc health absences G1  G2
## 1  ( 1 )  " "    " "      " "   " "  " "  " "    " "      " " "*"
## 2  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 3  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 4  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 5  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 6  ( 1 )  " "    " "      " "   " "  " "  " "    "*"      "*" "*"
## 7  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 8  ( 1 )  " "    " "      " "   " "  " "  " "    " "      "*" "*"
## 9  ( 1 )  " "    " "      " "   " "  " "  " "    "*"      "*" "*"
## 10  ( 1 ) " "    " "      " "   " "  " "  " "    "*"      "*" "*"
## 11  ( 1 ) " "    " "      " "   "*"  " "  " "    "*"      "*" "*"
## 12  ( 1 ) " "    " "      " "   "*"  " "  "*"    "*"      "*" "*"
## 13  ( 1 ) " "    " "      " "   "*"  " "  "*"    "*"      "*" "*"
## 14  ( 1 ) " "    " "      " "   "*"  " "  "*"    "*"      "*" "*"
## 15  ( 1 ) " "    " "      " "   "*"  " "  "*"    "*"      "*" "*"
```

```
summary(subsets)$adjr2
```

```
##  [1] 0.8434889 0.8472902 0.8489562 0.8501271 0.8507762 0.8513288 0.8518147
##  [8] 0.8522527 0.8526228 0.8528667 0.8531441 0.8533129 0.8534618 0.8534808
## [15] 0.8535407
```

```
summary(subsets)$cp
```

```
##  [1] 32.5841845 17.1053566 10.8932613  6.8369620  5.0413803  3.6686368
##  [7]  2.5898385  1.7227267  1.1515489  1.1240570  0.9573861  1.2563559
## [13]  1.6420508  2.5809380  3.3465888
```

```
summary(subsets)$bic
```

```
##  [1] -1191.705 -1202.191 -1203.840 -1203.422 -1200.772 -1197.715 -1194.376
##  [8] -1190.835 -1187.002 -1182.618 -1178.385 -1173.676 -1168.881 -1163.512
## [15] -1158.327
```

## 3.4 Set validation for Best Subset

## 3.5 Best Subset

```
which.max(summary(subsets)$adjr2)
```

```
## [1] 15
```

```
which.min(abs(summary(subsets)$cp - 1:15))
```

```
## [1] 5
```

```
which.min(summary(subsets)$bic)
```

```
## [1] 3
```

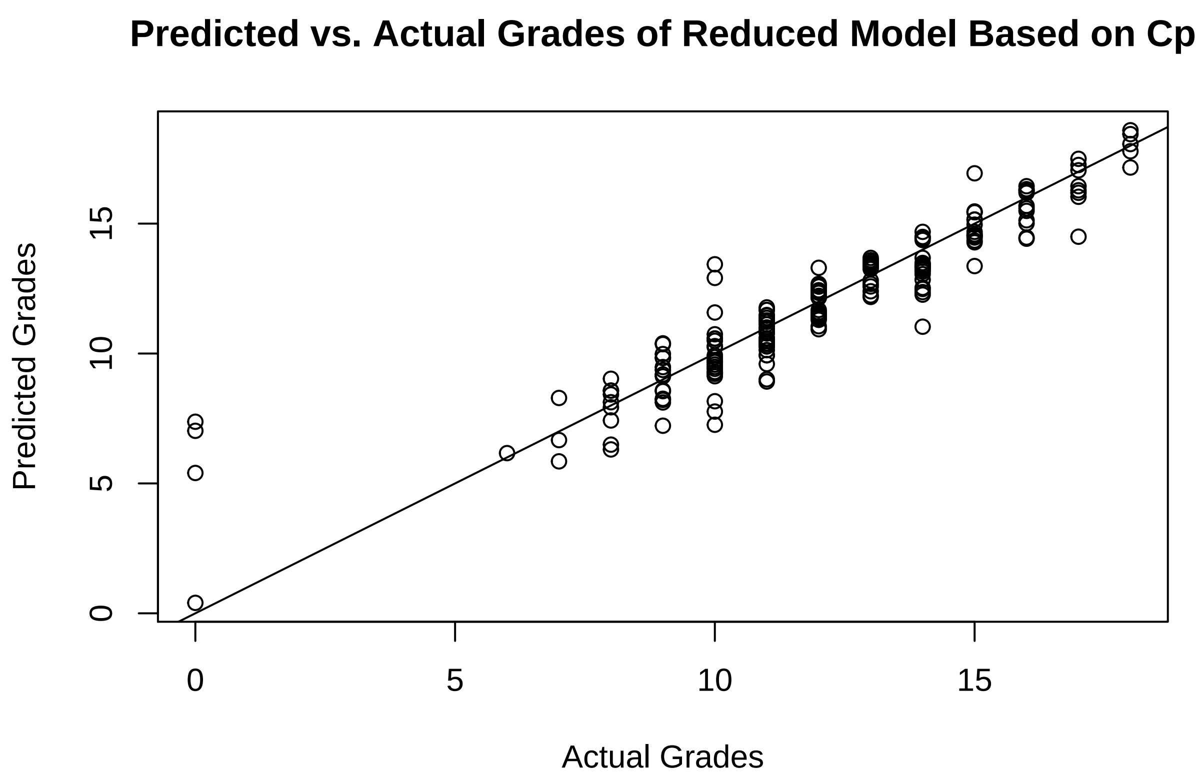
Didn't use adjr2 model bc they all make little difference.

### 3.6 Model Based on Mallow's Cp

```
reg.bestsubCP <- lm(G3 ~ sex + reason + failures + G1 + G2, data = student_por, subset = Z)

g3_pred_bestsubCP <- predict(reg.bestsubCP, student_por)
```

```
plot(student_por$G3[-Z], g3_pred_bestsubCP[-Z], xlab = "Actual Grades", ylab = "Predicted Grades", main
abline(0,1)
```



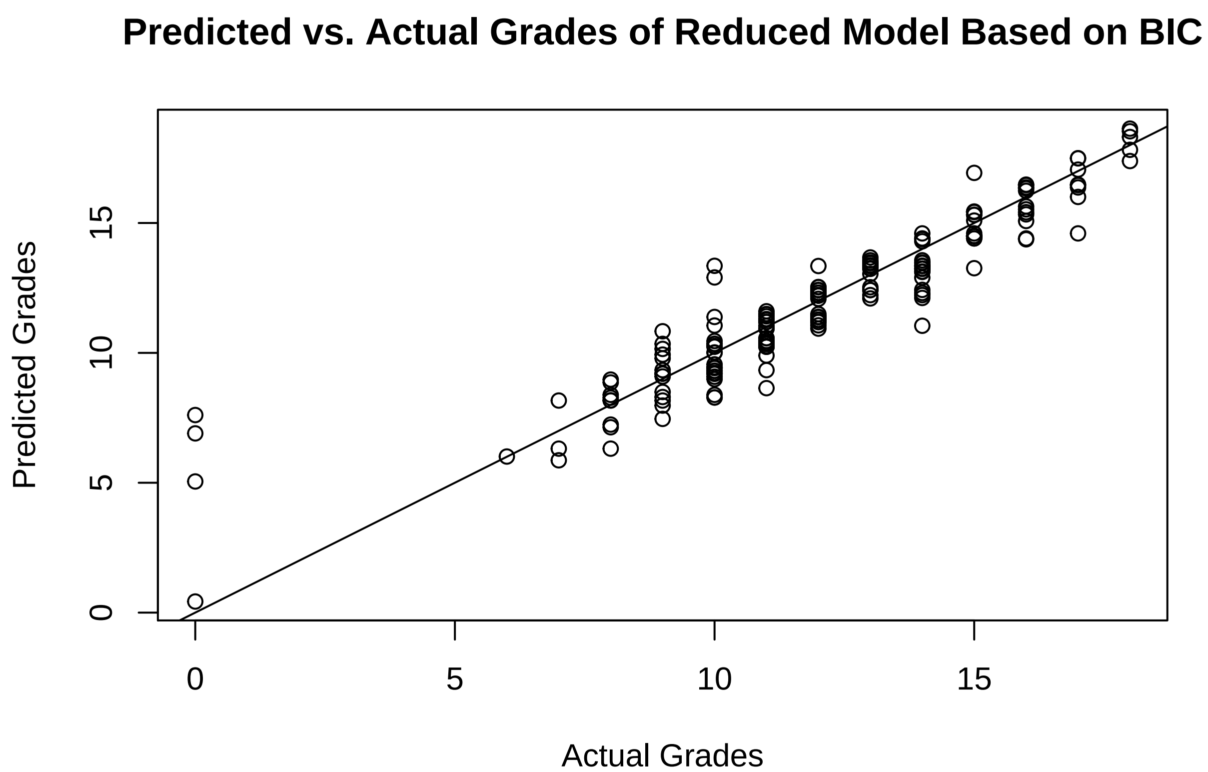
```
mse_cp <- mean((student_por$G3 - g3_pred_bestsubCP)[-Z] ^ 2)
```

### 3.7 Model Based on BIC

```
reg.bestsubBIC <- lm(G3 ~ reason + G1 + G2, data = student_por, subset = Z)

g3_pred_bestsubBIC <- predict(reg.bestsubBIC, student_por)
```

```
plot(student_por$G3[-Z], g3_pred_bestsubBIC[-Z], xlab = "Actual Grades", ylab = "Predicted Grades", mai
abline(0,1)
```




```
mse_bic <- mean((student_por$G3 - g3_pred_bestsubBIC)[-Z] ^ 2)
```

### 3.8 Step Functions

```
summary(forward)
```

```
##
## Call:
## lm(formula = G3 ~ G2 + G1 + failures + reason + absences + sex +
```

```
##     school + traveltime + health, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0833 -0.5178 -0.0053  0.6398  5.2097
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.44063    0.34169   1.290 0.197678
## G2                0.87996    0.03379  26.042  < 2e-16 ***
## G1                0.13706    0.03615   3.792 0.000164 ***
## failures         -0.24049    0.09074  -2.650 0.008244 **
## reasonhome       -0.09222    0.13010  -0.709 0.478659
## reasonother      -0.44994    0.16627  -2.706 0.006990 **
## reasonreputation -0.16537    0.13290  -1.244 0.213816
## absences          0.01623    0.01100   1.476 0.140522
## sexM             -0.20022    0.10191  -1.965 0.049894 *
## schoolMS         -0.22981    0.11621  -1.977 0.048419 *
## traveltime        0.11228    0.06839   1.642 0.101138
## health           -0.05394    0.03469  -1.555 0.120451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.242 on 637 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8522
## F-statistic: 340.7 on 11 and 637 DF,  p-value: < 2.2e-16
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = G3 ~ school + sex + reason + traveltime + failures +
##     health + absences + G1 + G2, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0833 -0.5178 -0.0053  0.6398  5.2097
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.44063    0.34169   1.290 0.197678
## schoolMS         -0.22981    0.11621  -1.977 0.048419 *
## sexM             -0.20022    0.10191  -1.965 0.049894 *
## reasonhome       -0.09222    0.13010  -0.709 0.478659
## reasonother      -0.44994    0.16627  -2.706 0.006990 **
## reasonreputation -0.16537    0.13290  -1.244 0.213816
## traveltime        0.11228    0.06839   1.642 0.101138
## failures         -0.24049    0.09074  -2.650 0.008244 **
## health           -0.05394    0.03469  -1.555 0.120451
## absences          0.01623    0.01100   1.476 0.140522
## G1                0.13706    0.03615   3.792 0.000164 ***
## G2                0.87996    0.03379  26.042  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.242 on 637 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8522
## F-statistic: 340.7 on 11 and 637 DF,  p-value: < 2.2e-16
```

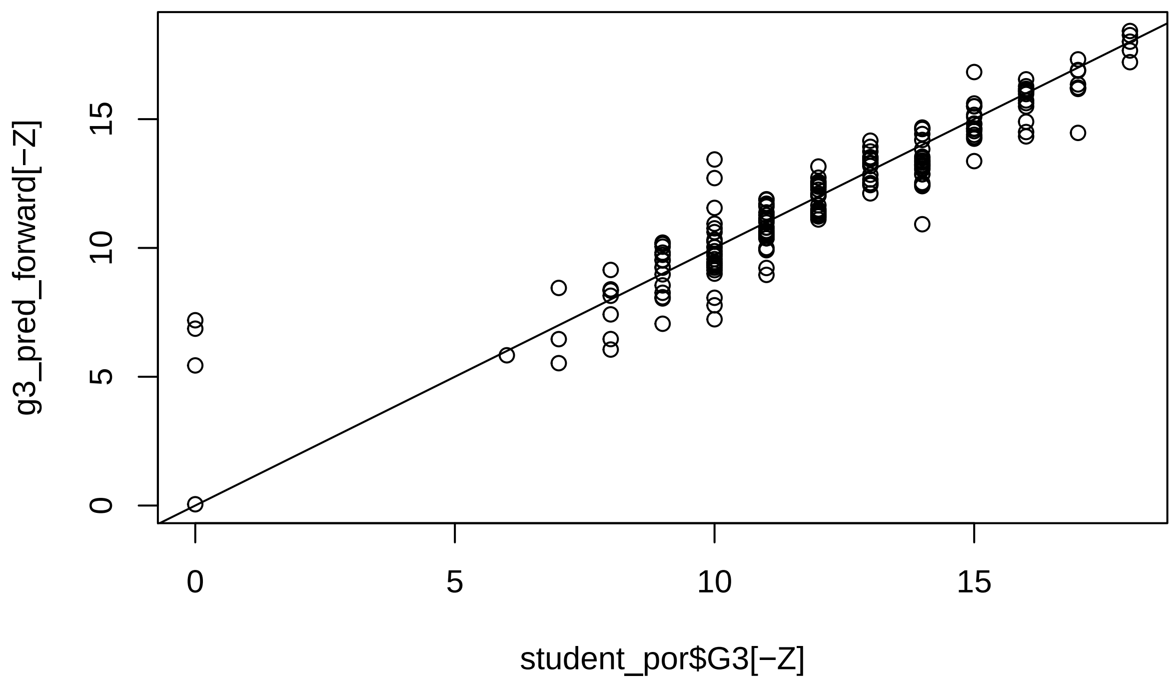
Forward and backward step functions yield the exact same model; proceeding with forward step-generated model.

### 3.9 Set validation

```
reg.forward <- lm(G3 ~ G2 + G1 + failures + reason + absences + sex + school + traveltime + health, data

g3_pred_forward <- predict(reg.forward, student_por)
```

```
plot(student_por$G3[-Z], g3_pred_forward[-Z])
abline(0, 1)
```




```
mse_valSet <- mean((student_por$G3 - g3_pred_forward)[-Z] ^ 2)
```

### 3.10 Ridge Regression & LASSO Preparation

```r
# Training/test split
# set.seed(1)
# train <- sample(1:n, n/2)
G3_test <- student_por$G3[-Z]

# Creating model matrix for rr and lasso calculations
x_col <- model.matrix(G3 ~ ., student_por)[, -1]
```

### 3.11 Ridge Regression

```r
set.seed(1)
cv.out1 <- cv.glmnet(x_col, student_por$G3, alpha = 0) # alpha = 0 ---> Ridge regression
predict(cv.out1, s = cv.out1$lambda.min, type = "coefficients")
```

```
## 42 x 1 sparse Matrix of class "dgCMatrix"
##                             1
## (Intercept)       0.647281475
## schoolMS         -0.203906036
## sexM             -0.140157325
## age               0.052338494
## addressU          0.134471281
## famsizeLE3        0.031162924
## PstatusT         -0.065759455
## Medu             -0.052585350
## Fedu              0.042315840
## Mjobhealth        0.214144623
## Mjobother        -0.125812519
## Mjobservices      0.107348800
## Mjobteacher       0.126800272
## Fjobhealth       -0.238572263
## Fjobother        -0.156282858
## Fjobservices     -0.300920468
## Fjobteacher      -0.249881588
## reasonhome       -0.066205753
## reasonother      -0.362017744
## reasonreputation -0.112276025
## guardianmother   -0.032801219
## guardianother     0.212481561
## traveltime        0.111686852
## studytime         0.060781281
## failures         -0.316944745
## schoolsupyes     -0.201688220
## famsupyes         0.087520158
## paidyes          -0.153298259
## activitiesyes     0.018965743
## nurseryyes       -0.084879671
## higheryes         0.280580244
## internetyes       0.099020055
## romanticyes      -0.087129209
## famrel            0.008539498
```

```
## freetime         -0.053552694
## goout            -0.025522281
## Dalc             -0.056693604
## Walc             -0.025883700
## health           -0.065912944
## absences          0.011871609
## G1                0.258399918
## G2                0.683417827
```

```
rr.mod <- glmnet(x_col[Z, ], student_por$G3[Z], alpha = 0, lambda = cv.out1$lambda.min)
rr.pred <- predict(rr.mod, s = cv.out1$lambda.min, newx = x_col[-Z, ])

mse_rr <- mean((rr.pred - student_por$G3[-Z])^2)
```

$\lambda = .30$

### 3.12 LASSO

```
set.seed(1)
cv.out2 <- cv.glmnet(x_col, student_por$G3, alpha = 1)
predict(cv.out2, s = cv.out2$lambda.min, type = "coefficients")
```

```
## 42 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)       0.46985582
## schoolMS         -0.03190401
## sexM             -0.01841156
## age               .
## addressU          .
## famsizeLE3        .
## PstatusT          .
## Medu              .
## Fedu              .
## Mjobhealth        .
## Mjobother         .
## Mjobservices      .
## Mjobteacher       .
## Fjobhealth        .
## Fjobother         .
## Fjobservices      .
## Fjobteacher       .
## reasonhome        .
## reasonother      -0.14557639
## reasonreputation  .
## guardianmother    .
## guardianother     .
## traveltime        .
## studytime         .
## failures         -0.09120067
## schoolsupyes      .
## famsupyes         .
```

```
## paidyes            .
## activitiesyes      .
## nurseryyes         .
## higheryes          .
## internetyes        .
## romanticyes        .
## famrel             .
## freetime           .
## goout              .
## Dalc               .
## Walc               .
## health             .
## absences           .
## G1                 0.12252007
## G2                 0.87247067
```

$\lambda = .10$

```
lasso.mod <- glmnet(x_col[Z, ], student_por$G3[Z], alpha = 1, lambda = cv.out2$lambda.min)
lasso.pred <- predict(lasso.mod, s = cv.out2$lambda.min, newx = x_col[-Z, ])

mse_lasso <- mean((lasso.pred - student_por$G3[-Z])^2)
```

```
student_por.dimred <- lm(G3 ~ school + sex + reason + failures + G1 + G2, student_por)
summary(student_por.dimred)
```

```
##
## Call:
## lm(formula = G3 ~ school + sex + reason + failures + G1 + G2,
##     data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2349 -0.4970  0.0057  0.6422  5.3485
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.53906    0.27382   1.969 0.049420 *
## schoolMS         -0.20979    0.11165  -1.879 0.060691 .
## sexM             -0.21513    0.10121  -2.126 0.033927 *
## reasonhome       -0.08152    0.12873  -0.633 0.526772
## reasonother      -0.45325    0.16677  -2.718 0.006748 **
## reasonreputation -0.14147    0.13181  -1.073 0.283529
## failures         -0.22420    0.09075  -2.471 0.013751 *
## G1                0.12912    0.03608   3.579 0.000371 ***
## G2                0.88212    0.03382  26.081  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.247 on 640 degrees of freedom
## Multiple R-squared:  0.8529, Adjusted R-squared:  0.8511
## F-statistic: 463.9 on 8 and 640 DF,  p-value: < 2.2e-16
```

### 3.13 Comparing MSEs

```r
tibble("method" = c("BIC-Minimized", "Cp-Minimized", "LASSO", "Linear Regression", "Ridge Regression",
       "MSE" = c(mse_bic, mse_cp, mse_lasso, mse_lm, mse_rr, mse_valSet)) %>%
  arrange(MSE)
```

```
## # A tibble: 6 x 2
##   method              MSE
##   <chr>             <dbl>
## 1 BIC-Minimized      1.43
## 2 AIC-Minimized      1.46
## 3 Cp-Minimized       1.47
## 4 LASSO              1.53
## 5 Linear Regression  1.55
## 6 Ridge Regression   1.60
```

```r
reg.bestsubBIC
```

```
##
## Call:
## lm(formula = G3 ~ reason + G1 + G2, data = student_por, subset = Z)
##
## Coefficients:
##      (Intercept)        reasonhome       reasonother  reasonreputation
##         -0.01247          -0.08366          -0.33815          -0.10231
##               G1                G2
##          0.10995           0.92593
```

```r
reg.forward # Picking this one
```

```
##
## Call:
## lm(formula = G3 ~ G2 + G1 + failures + reason + absences + sex +
##     school + traveltime + health, data = student_por, subset = Z)
##
## Coefficients:
##      (Intercept)                G2                G1          failures
##          0.50898           0.90144           0.09950          -0.37556
##       reasonhome       reasonother  reasonreputation          absences
##         -0.13167          -0.33802          -0.21118           0.01524
##             sexM          schoolMS        traveltime            health
##         -0.20738          -0.22059           0.16919          -0.04444
```

```r
reg.bestsubCP
```

```
##
## Call:
## lm(formula = G3 ~ sex + reason + failures + G1 + G2, data = student_por,
##     subset = Z)
##
```

```
## Coefficients:
##      (Intercept)              sexM        reasonhome       reasonother
##          0.55807          -0.18604          -0.11971          -0.36133
## reasonreputation          failures                G1                G2
##         -0.14844          -0.35971           0.09759           0.90438
```

Picking forward-selected candidate model b/c best balance of number of predictors while sacrificing only a little accuracy.

```r
summary(reg.forward)
```

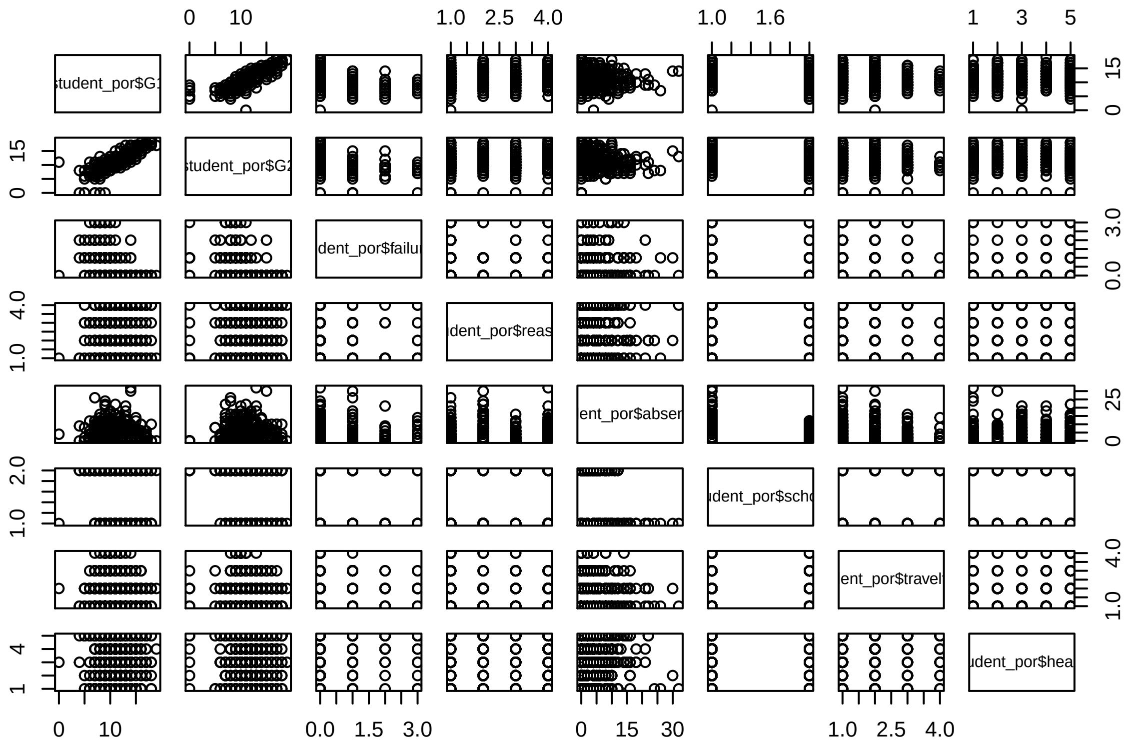
```
##
## Call:
## lm(formula = G3 ~ G2 + G1 + failures + reason + absences + sex +
##     school + traveltime + health, data = student_por, subset = Z)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8824 -0.4672 -0.0923  0.6427  5.0271
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.50898    0.42147   1.208 0.227839
## G2                0.90144    0.03983  22.634  < 2e-16 ***
## G1                0.09950    0.04334   2.296 0.022162 *
## failures         -0.37556    0.10959  -3.427 0.000667 ***
## reasonhome       -0.13167    0.15640  -0.842 0.400327
## reasonother      -0.33802    0.20922  -1.616 0.106879
## reasonreputation -0.21118    0.16202  -1.303 0.193108
## absences          0.01524    0.01385   1.100 0.271973
## sexM             -0.20738    0.12313  -1.684 0.092851 .
## schoolMS         -0.22059    0.13985  -1.577 0.115432
## traveltime        0.16919    0.08337   2.029 0.043026 *
## health           -0.04444    0.04150  -1.071 0.284835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.264 on 442 degrees of freedom
## Multiple R-squared:  0.854,  Adjusted R-squared:  0.8504
## F-statistic: 235.1 on 11 and 442 DF,  p-value: < 2.2e-16
```

```r
summary(
  lm(formula = G3 ~ G2 + G1 + failures + reason + absences + sex +
    school + traveltime + health, data = student_por)
)
```

```
##
## Call:
## lm(formula = G3 ~ G2 + G1 + failures + reason + absences + sex +
##     school + traveltime + health, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -9.0833 -0.5178 -0.0053  0.6398  5.2097
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.44063    0.34169   1.290 0.197678
## G2                0.87996    0.03379  26.042  < 2e-16 ***
## G1                0.13706    0.03615   3.792 0.000164 ***
## failures         -0.24049    0.09074  -2.650 0.008244 **
## reasonhome       -0.09222    0.13010  -0.709 0.478659
## reasonother      -0.44994    0.16627  -2.706 0.006990 **
## reasonreputation -0.16537    0.13290  -1.244 0.213816
## absences          0.01623    0.01100   1.476 0.140522
## sexM             -0.20022    0.10191  -1.965 0.049894 *
## schoolMS         -0.22981    0.11621  -1.977 0.048419 *
## traveltime        0.11228    0.06839   1.642 0.101138
## health           -0.05394    0.03469  -1.555 0.120451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.242 on 637 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8522
## F-statistic: 340.7 on 11 and 637 DF,  p-value: < 2.2e-16
```

```
pairs(tibble(student_por$G1,
             student_por$G2,
             student_por$failures,
             student_por$reason,
             student_por$absences,
             student_por$school,
             student_por$traveltime,
             student_por$health))
```




```
cor(data.frame(student_por$G1, student_por$G2))
```

```
##                student_por.G1 student_por.G2
## student_por.G1      1.0000000      0.8649816
## student_por.G2      0.8649816      1.0000000
```

```
car::vif(reg.forward)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## G2         3.947302  1        1.986782
## G1         3.931017  1        1.982679
## failures   1.233508  1        1.110634
## reason     1.175917  3        1.027376
## absences   1.082722  1        1.040539
## sex        1.044820  1        1.022164
## school     1.264176  1        1.124356
## traveltime 1.101912  1        1.049720
## health     1.066354  1        1.032644
```

```
reg.forward_mod <- lm(G3 ~ G2 + failures + reason + absences + sex + school + traveltime + health, stud

summary(reg.forward_mod)
```

```
##
```

```
## Call:
## lm(formula = G3 ~ G2 + failures + reason + absences + sex + school +
##     traveltime + health, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0375 -0.4999 -0.0428  0.6332  5.1354
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.81202    0.33077   2.455  0.01436 *
## G2                0.98521    0.01947  50.606  < 2e-16 ***
## failures         -0.27267    0.09129  -2.987  0.00293 **
## reasonhome       -0.07680    0.13139  -0.584  0.55910
## reasonother      -0.44246    0.16799  -2.634  0.00865 **
## reasonreputation -0.15001    0.13422  -1.118  0.26415
## absences          0.01204    0.01105   1.089  0.27641
## sexM             -0.21827    0.10286  -2.122  0.03422 *
## schoolMS         -0.28698    0.11643  -2.465  0.01397 *
## traveltime        0.11205    0.06911   1.621  0.10543
## health           -0.04955    0.03503  -1.414  0.15775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 638 degrees of freedom
## Multiple R-squared:  0.8514, Adjusted R-squared:  0.8491
## F-statistic: 365.7 on 10 and 638 DF,  p-value: < 2.2e-16
```

```
car::vif(reg.forward_mod)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## G2         1.324014  1        1.150658
## failures   1.206829  1        1.098558
## reason     1.190013  3        1.029418
## absences   1.083037  1        1.040691
## sex        1.054733  1        1.027002
## school     1.268032  1        1.126069
## traveltime 1.101477  1        1.049513
## health     1.056308  1        1.027768
```