SimpleStatisticalAnalysis

Xiaosong Zhang

9/18/2016

1.What's your proudest achievement and why?

It can be a personal project or something you've worked on professionally. Just a short paragraph is fine.

This is a great question, and it really let me calm down and drive me to spend some time to think what are things I am truly caring and desire to do.

From my understanding, there are three types of things that I think could be considered as proud achievement:

- Self-improvement: Be able to complete difficult tasks through learning and practice. For example, build a computer from pieces of hardware for the first time; install and run my personal server or improve the performance of the code I wrote, etc.
- Standing out in competition: Prepare and give a presentation that impresses the audiences better than my counterpart; Solve a given problem in a more efficient and elegant way, etc.
- Helping others overcome their obstacles: Tutor students and friends to not only getting understand the questions they cannot figure out but also try to locate the knowledge points they didn't know well which preventing them from figuring out the problem, etc.

So I think the proudest achievement of me should be able to fit in all those three categories. One of my recent achievement, which I believe can satisfy those should be that I independently instructed an Intro to Statistics class for 40 undergraduate students.

Volunteer to be an instructor is a tough decision to make, it requires much more enthusiasm, preparation time and stamina compare to teach regular recitation sessions. Even I know the last semester will extremely busy and pressure, I still decided to choose the hard way. I truly enjoying share my knowledge with others, many of my friends and sometimes my friends' friends come to me to seeking help with but not limited to academic problems. Sometimes I even wonder if the characteristics like this are inheritable because my grandparents from both sides are teachers and my father is a professor in Environmental Sciences, I must be somehow affected.

I carefully prepared for and gave a trail lecture in front of professors and graduate students and been selected as one of top two candidates to teach the Intro to Stats course.

During the semester I taught the course, I Carefully prepared for each lecture, provide and walk them through worksheets for every chapter, encourage them forming study groups and actively keeping touch with students' success coaches upon request.

I closely keep tracking the performance of all my student, tried my best to give reasonable accommodations to help them keep on the right track. At the end of the semester, there is zero student withdraw from my session, and only one student failed – who didn't show up and take the final exam.

2. Tell us about a book or article related to data analysis you read recently, why you liked it, and why we should read it.

Cortez, Paulo, and Alice Maria GonÃ§alves Silva. "Using data mining to predict secondary school student performance." (2008).

I think the paper related to data analysis and also related to student performance prediction should be the **Using data mining to predict secondary school student performance** by Dr.Paulo Cortez is inspirational.

In this paper, the author analyzed the 33 attributes thought to be related to the secondary students' academic performance. The author uses not only commonly used regression methods but also use multiple different classification approaches(Might be inspired by Dr. Behrouz Minaei-Bidgoli from MSU).

By transform the numerical final-grade variable into a categorical variable has only limited levels, Dr. Cortez successfully trained models based on Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines(SVM) which can predict the student performance with a relatively high Percentage of Correct Classifications (PCC) value, especially when the levels of the response variable had been reduced to binary.

I think if the purpose of DIG is to build a system that could make a prediction whether a student need encourage, alert or may be intervention, those classification tools should be considered.

3. Tell us about one aspect of the Digital Innovation Greenhouse you really like, and why.

From the materials that I read and watch, the DIG is trying very hard to help the student to became successful in their college life. In my opinion, the success in the undergraduate study is more depends on motivation, persistence other than intelligence. I believe everyone stepped on the campus were planning to graduate, but since we are human our ability to keep overcome unpredictable obstacles are limited continuously. The learning curve is not smooth for any student, and I would like to say it's constantly changing even in each and every course. I think the ECoach system along with the other systems makes by DIG are dedicated to detecting when a student is facing a steep segment of his/her learning curve and deliver the personalized help to help them defeat the peak and move on.

I would love to have some help like this when I was in college, even a fraction of it.

Besides the lofty goal of DIG, I also strongly attracted by the working environment(University) and colleagues of DIG. It's a great opportunity for me to work with and learn from those masterminds.

4.STATS250 KEY FACTOR ANALYSIS

Please answer the following question: Other than GPAO, which variable(s) best predict the variable for GRD_PTS_PER_UNIT for the course STATS 250 (SUBJECT="STATS" and CATALOG_NBR=250)? In other words, what is the best predictor of a student's performance in STATS 250 other than the student's own GPA?

Conclusion

Due to the limited time I have, I only tried the most simple linear models to test the data.

lm(formula = GRD_PTS_PER_UNIT ~ HSGPA + LAST_ACT_COMP_SCORE + LAST_ACT_MATH_SCORE + SEX, data = S.A.c)

The result shows that the p-value of HSGPA, LAST_ACT_COMP_SCORE, LAST_ACT_MATH_SCORE and SEXM are all significant, but from the residual plot and Q-Q plot we can easily tell that the relationship is not linear.

If I have choose only the 'best' predictor, I would choose LAST_ACT_COMP_SCORE.

And with the result of Lilliefors (Kolmogorov-Smirnov) test, none of those explanatory and response variables are normally distributed.

To train a better model to make a more accurate prediction, we might need to try GLM, GAM or maybe classification methods. Well, after cleaning the data properly.

I didn't look into those major, department and group data since there are too many missing data, I think I need more information to learn how to handle them correctly.

Data Import

```
require(nortest)
require("dplyr")
require("ggplot2")
require("rstudioapi")
require("ggthemes")
#set working directory to current .r file path
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

#sourceDir is use to source all .r file under same Dir all in once
sourceDir <- function(path, trace = TRUE, ...) {
  for (nm in list.files(path, pattern = "[.][RrSsQq]$")) {
    if(trace) cat(nm,":")
    source(file.path(path, nm), ...)
    if(trace) cat("\n")
  }
}
#Source all functions uder PLA-MOOC into R-Environment for future use.
sourceDir('PLA-MOOC')
```

```
## course.impact.R :
## course.pathways.barplots.R :
## course.pathways.treemaps.R :
## course.persistence.module.R :
## grade.penalty.module.R :
```

```r
#loading data into Environment
# Start the clock!
ptm <- proc.time()
student.course <- read.csv("PLA-MOOC/student.course.csv")
proc.time() - ptm
```

```
##    user  system elapsed
##    4.28    0.16    4.43
```

```r
ptm <- proc.time()
student.record <- read.csv("PLA-MOOC/student.record.csv")
proc.time() - ptm
```

```
##    user  system elapsed
##    0.78    0.03    0.82
```

Data Exploration

Operations which creat to much unimportant output will not be run, but the code can still be find in the *STATS250 KEY FACTOR ANALYSIS.R* file

```r
names(student.record)
```

```
##  [1] "MAJOR3_DESCR"          "MAJOR2_DESCR"
##  [3] "MAJOR1_DESCR"          "HSGPA"
##  [5] "LAST_ACT_ENGL_SCORE"   "LAST_ACT_MATH_SCORE"
##  [7] "LAST_ACT_READ_SCORE"   "LAST_ACT_SCIRE_SCORE"
##  [9] "LAST_ACT_COMP_SCORE"   "LAST_SATI_VERB_SCORE"
## [11] "LAST_SATI_MATH_SCORE"  "LAST_SATI_TOTAL_SCORE"
## [13] "SEX"                   "STDNT_GROUP1"
## [15] "STDNT_GROUP2"          "MAJOR1_DEPT"
## [17] "MAJOR2_DEPT"           "MAJOR3_DEPT"
## [19] "ANONID"                "ADMIT_TERM"
## [21] "MAJOR1_TERM"           "MAJOR2_TERM"
## [23] "MAJOR3_TERM"
```

```r
names(student.course)
```

```
## [1] "ANONID"           "SUBJECT"          "CATALOG_NBR"
## [4] "GRD_PTS_PER_UNIT" "GPAO"             "DIV"
## [7] "ANON_INSTR_ID"    "TERM"
```

```r
# use summary(student.record) ;summary(student.course) to get a big picture of the data
summary(student.record$HSGPA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   3.300   3.700   3.127   3.900  36.000   13666
```

HSGPA= 36? it most likely to be a error. UM admit more than two thousands students with HSGPA 0?? or it should be NAs?

```
# HSGPA= 36? it most likely to be a error
# UM admit students with HSGPA 0?? or it should be NAs?
filter(student.record, HSGPA>4.0)[,1:6]
```

```
##   MAJOR3_DESCR        MAJOR2_DESCR               MAJOR1_DESCR HSGPA
## 1         <NA> Movement Science BS Physical Education T.E. BS    32
## 2         <NA>                <NA>                       <NA>    36
##   LAST_ACT_ENGL_SCORE LAST_ACT_MATH_SCORE
## 1                  NA                  NA
## 2                  NA                  NA
```

```
# many of those who have HSGPA = 0 have LAST_ACT_MATH_SCORE close to Max
# Cloud be a evidence their HSGPA should be NA instead of 0.
count(filter(student.record, HSGPA< 1 & LAST_ACT_MATH_SCORE >35 ))
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1   118
```

Many of those who have HSGPA = 0 have LAST_ACT_MATH_SCORE close to Max. This cloud be a evidence their HSGPA should be NA instead of 0.

```
#prepare data for dplyr
Course <-tbl_df(student.course)
Record <-tbl_df(student.record)

#getting to know the dataset's big picture
glimpse(Course)
```

```
## Observations: 1,327,065
## Variables: 8
## $ ANONID           <int> 26, 114, 121, 125, 180, 207, 224, 249, 356, 4...
## $ SUBJECT          <fctr> ACC, ACC, ACC, ACC, ACC, ACC, ACC, ACC, ACC,...
## $ CATALOG_NBR      <int> 272, 272, 272, 272, 272, 272, 272, 272, 272, ...
## $ GRD_PTS_PER_UNIT <dbl> 2.0, 2.0, 4.0, 1.3, 3.0, 2.0, 3.0, 3.0, 4.0, ...
## $ GPAO             <dbl> 3.343636, 2.817857, 4.000000, 3.639063, 3.846...
## $ DIV              <fctr> P, P, P, P, P, P, P, P, P, P, P, P, P, P, P,...
## $ ANON_INSTR_ID    <int> 2920, 2920, 201, 3360, 2920, 2920, 3604, 1914...
## $ TERM             <int> 79, 83, 111, 84, 107, 93, 69, 123, 107, 84, 7...
```

```
glimpse(Record)
```

```
## Observations: 138,888
## Variables: 23
## $ MAJOR3_DESCR          <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ MAJOR2_DESCR          <fctr> NA, NA, NA, NA, NA, NA, NA, NA, Archite...
## $ MAJOR1_DESCR          <fctr> NA, Asian Studies BA, Psychology BA, Ci...
## $ HSGPA                 <dbl> NA, 3.8, 3.3, 3.9, 3.8, 3.9, 4.0, 3.4, 3...
## $ LAST_ACT_ENGL_SCORE   <int> NA, 25, NA, 33, 24, 31, 29, 28, NA, 24, ...
## $ LAST_ACT_MATH_SCORE   <int> NA, 22, NA, 32, 22, 22, 34, 30, NA, 31, ...
```

```
## $ LAST_ACT_READ_SCORE   <int> NA, 29, NA, 28, 18, 34, 28, 27, NA, 24, ...
## $ LAST_ACT_SCIRE_SCORE  <int> NA, 28, NA, 30, 18, 22, 36, 27, NA, 27, ...
## $ LAST_ACT_COMP_SCORE   <int> NA, 26, NA, 31, 21, 27, 32, 28, NA, 27, ...
## $ LAST_SATI_VERB_SCORE  <int> NA, NA, NA, NA, 610, NA, NA, NA, NA, 450...
## $ LAST_SATI_MATH_SCORE  <int> NA, NA, NA, NA, 540, NA, NA, NA, NA, 640...
## $ LAST_SATI_TOTAL_SCORE <int> NA, NA, NA, NA, 1150, NA, NA, NA, NA, 10...
## $ SEX                   <fctr> F, F, M, F, F, F, M, F, M, F, F, M, F, ...
## $ STDNT_GROUP1          <fctr> NA, NA, E, NA, C, NA, NA, NA, NA, B, NA...
## $ STDNT_GROUP2          <fctr> NA, NA, NA, NA, NA, NA, G, NA, NA, NA, ...
## $ MAJOR1_DEPT           <fctr> NA, Asian Languages And Cultures, Psych...
## $ MAJOR2_DEPT           <fctr> NA, NA, NA, NA, NA, NA, NA, NA, College...
## $ MAJOR3_DEPT           <fctr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ ANONID                <int> 1, 2, 4, 6, 7, 8, 9, 10, 12, 13, 14, 17,...
## $ ADMIT_TERM            <int> NA, 110, 63, 106, 83, 126, 126, 110, 63,...
## $ MAJOR1_TERM           <int> NA, 123, 79, 125, NA, NA, NA, 123, 77, 1...
## $ MAJOR2_TERM           <int> NA, NA, NA, NA, NA, NA, NA, NA, 77, NA, ...
## $ MAJOR3_TERM           <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
```

```
#names(Course)
#names(Record)
levels(Course$SUBJECT)
```

```
##  [1] "ACC"      "AMCULT"   "ANTHRBIO" "ANTHRCUL" "ARTDES"   "ASIAN"
##  [7] "ASTRO"    "BE"       "BIOLOGY"  "BIT"      "BUDDHST"  "CHEM"
## [13] "CICS"     "CLCIV"    "CMPTRSC"  "COMM"     "DANCE"    "ECON"
## [19] "EECS"     "ENGLISH"  "ENGR"     "FIN"      "FRENCH"   "GEOSCI"
## [25] "GTBOOKS"  "HISTORY"  "LHC"      "LING"     "MATH"     "MCDB"
## [31] "MECHENG"  "MKT"      "MO"       "NURS"     "OB"       "OM"
## [37] "OMS"      "PHIL"     "PHYSICS"  "POLSCI"   "PSYCH"    "RELIGION"
## [43] "SMS"      "SOC"      "SPANISH"  "STATS"    "STRATEGY" "UC"
## [49] "WOMENSTD"
```

To manipulate large dataset, I would like to use dplyr since it's very fast(based on C++) and many operation similar to *SQL*.

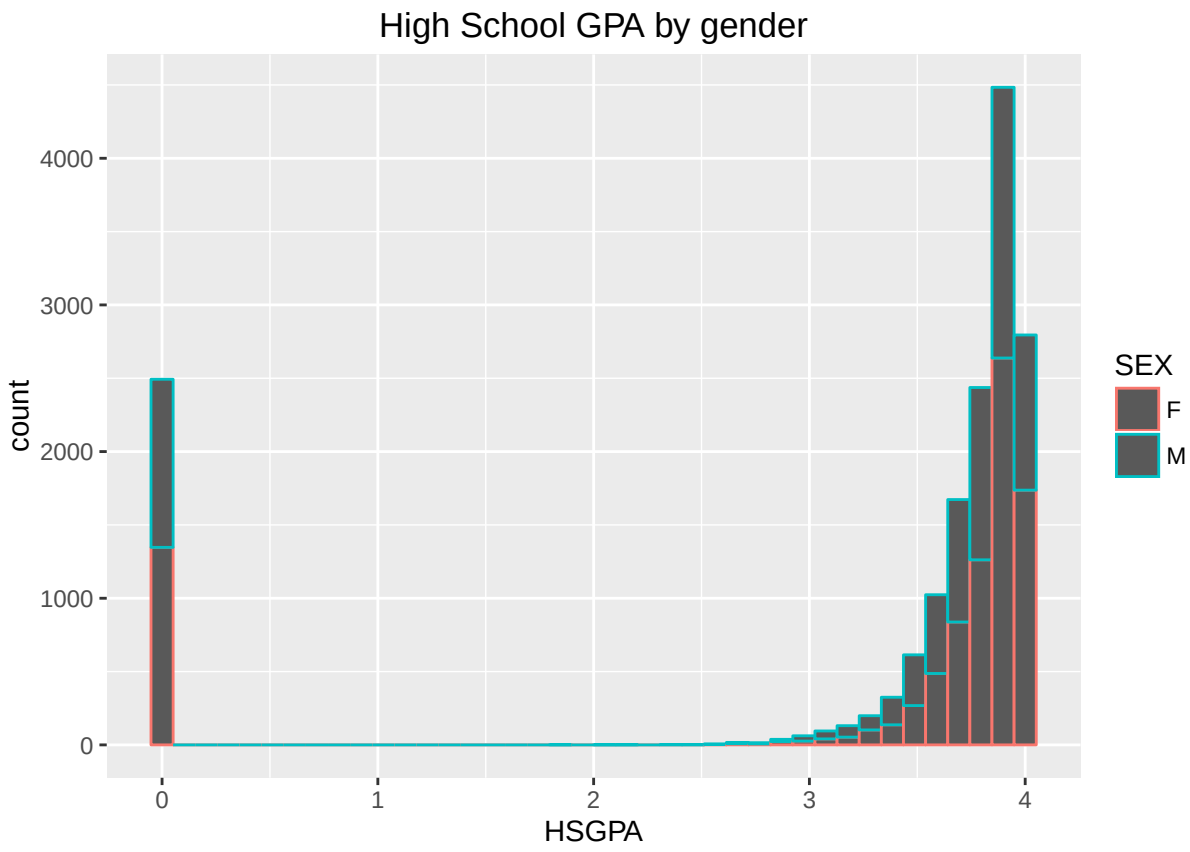
```
#select only the data required for look into STATS 250
STATS250 = filter(Course,SUBJECT =='STATS' & CATALOG_NBR == '250')
#aggregate the 'Record' and 'Course' table by the Anonymous ID
S.A = STATS250.Aggregate <- inner_join(STATS250,Record,by = 'ANONID')
attach(S.A)
#And we could write STATS250.Aggregate data to "SA.csv" which will save
#alot of time if we want to focus on STATS250
write.table(S.A, file = "SA.csv")


#we can check the data use some basic commands
#Check
# glimpse(S.A)
# summary(S.A)
table(SEX)
```

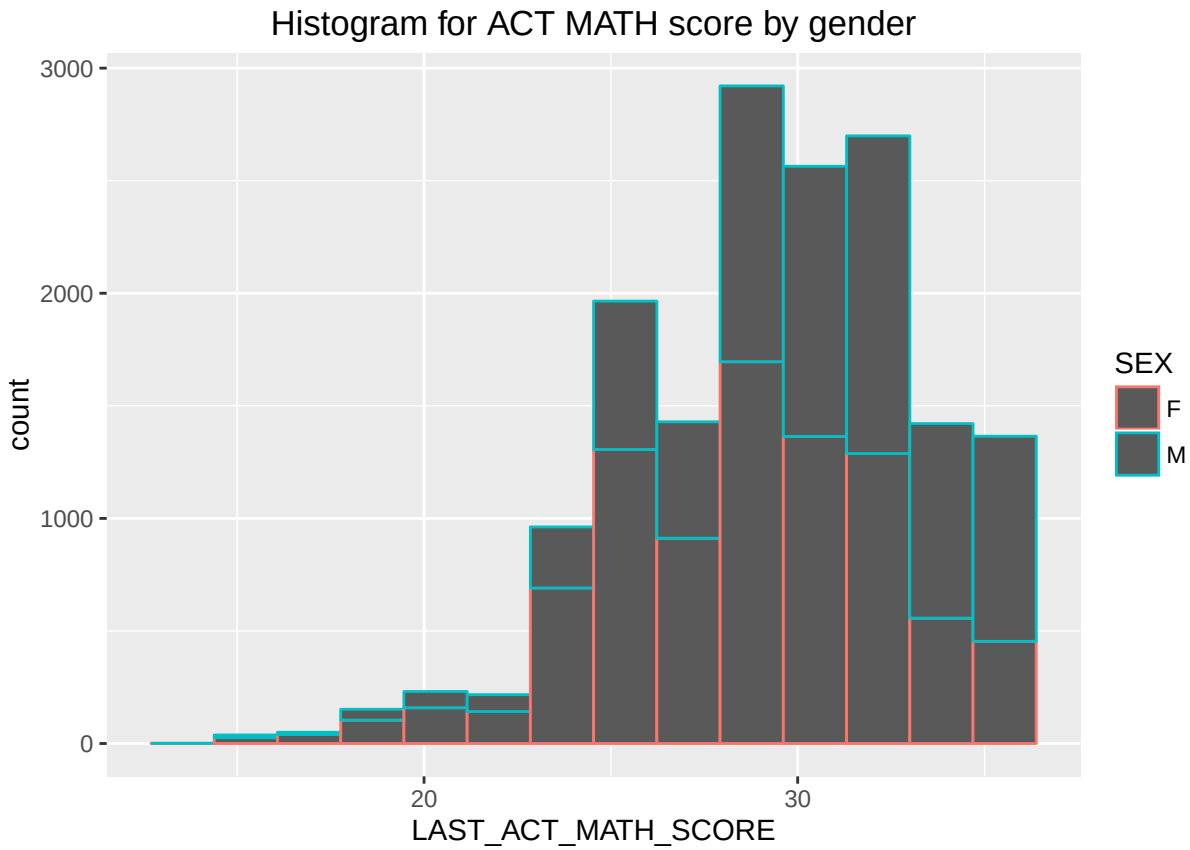
```
## SEX
##    F    M
## 8983 7461
```

```
qplot(HSGPA,colour = SEX,bins = 40,main = "High School GPA by gender")
```

```
## Warning: Removed 28 rows containing non-finite values (stat_bin).
```



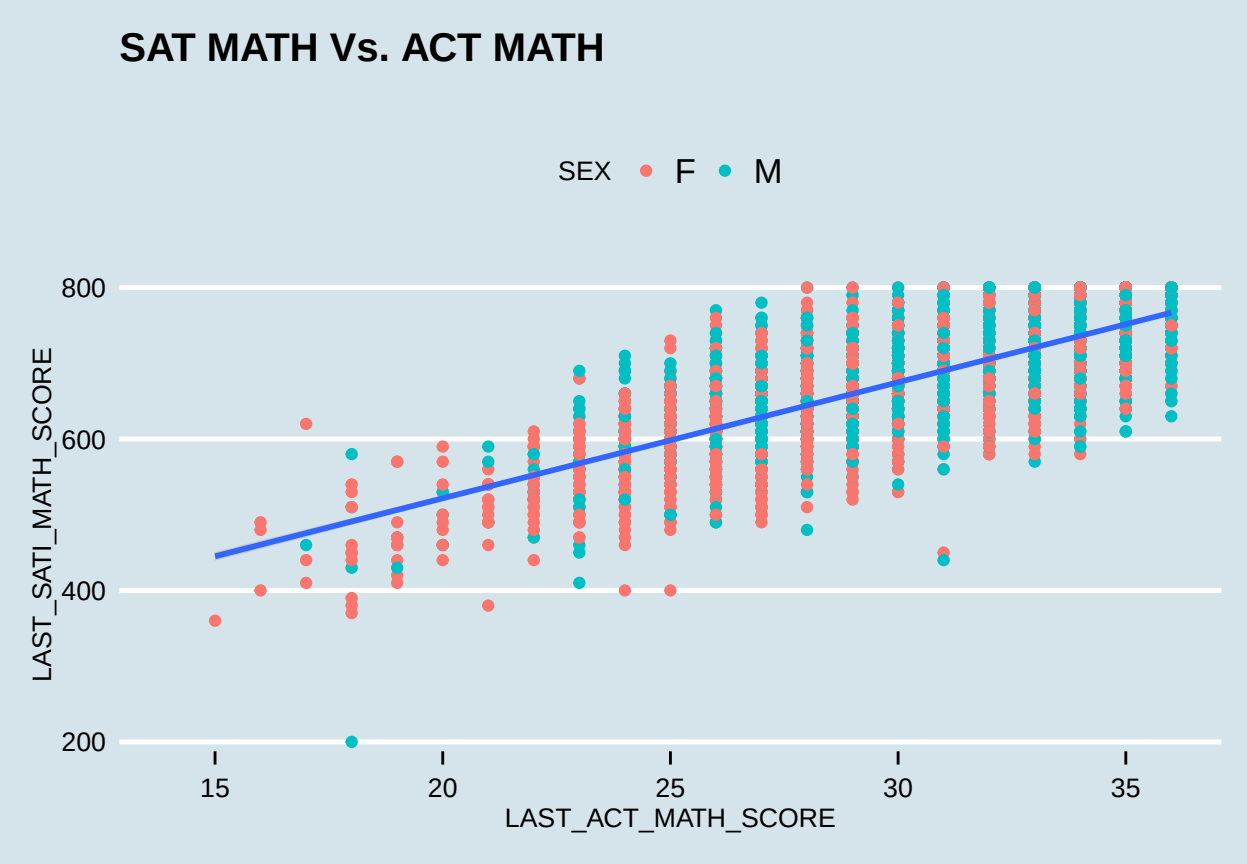

```
qplot(LAST_ACT_MATH_SCORE,colour = SEX,bins = 14,main = "Histogram for ACT MATH score by gender")
```




```
p = ggplot(data = S.A, aes(y = LAST_SATI_MATH_SCORE, x = LAST_ACT_MATH_SCORE))
p + geom_point(aes(color = SEX)) +
  labs(title = "SAT MATH Vs. ACT MATH")+
  geom_smooth(method = "lm")+  theme_economist()
```

```
## Warning: Removed 10986 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 10986 rows containing missing values (geom_point).
```




```
#normality check before model fitting
lillie.test(LAST_ACT_MATH_SCORE)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  LAST_ACT_MATH_SCORE
## D = 0.089798, p-value < 2.2e-16
```

```
lillie.test(HSGPA)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  HSGPA
## D = 0.37327, p-value < 2.2e-16
```

```
lillie.test(LAST_SATI_MATH_SCORE)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  LAST_SATI_MATH_SCORE
## D = 0.054764, p-value < 2.2e-16
```

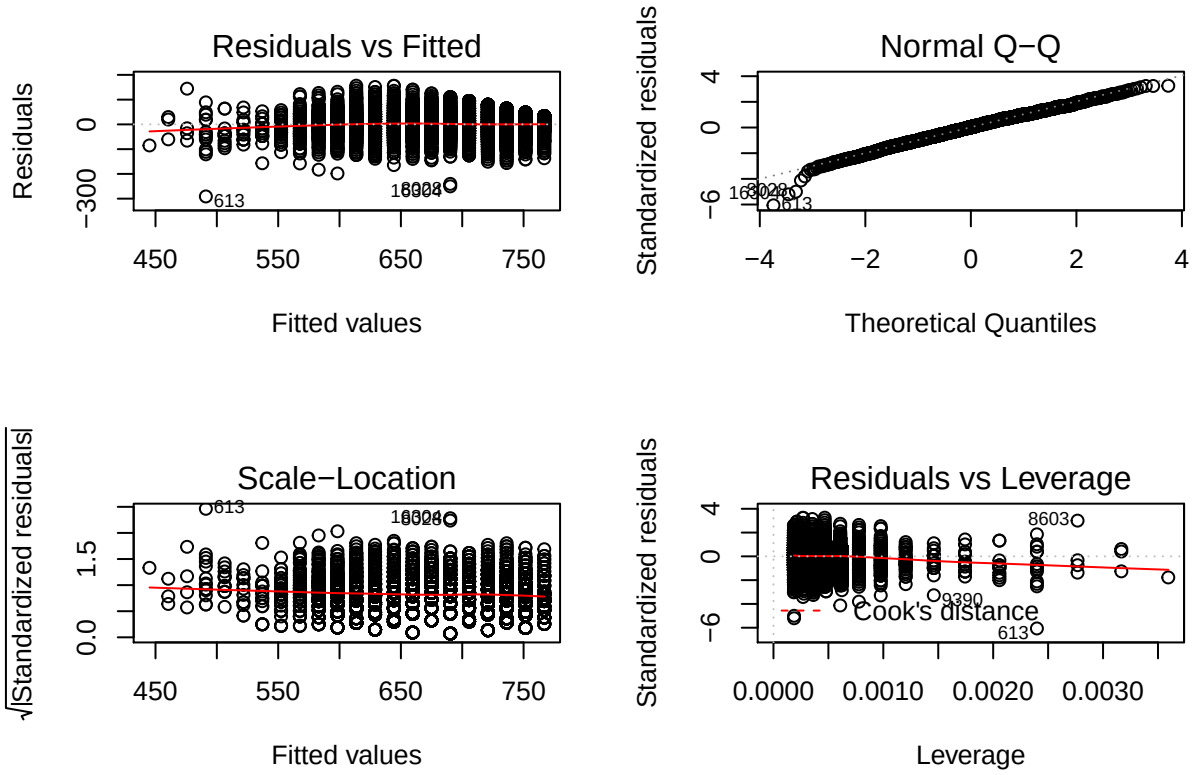
```r
lillie.test(GRD_PTS_PER_UNIT)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  GRD_PTS_PER_UNIT
## D = 0.17504, p-value < 2.2e-16
```

```r
fit.SatVsAct <- lm(LAST_SATI_MATH_SCORE~LAST_ACT_MATH_SCORE)
summary(fit.SatVsAct)
```

```
##
## Call:
## lm(formula = LAST_SATI_MATH_SCORE ~ LAST_ACT_MATH_SCORE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -291.051  -31.539    1.043   33.784  156.366
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         215.2411     5.5708   38.64   <2e-16 ***
## LAST_ACT_MATH_SCORE  15.3228     0.1817   84.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.07 on 5456 degrees of freedom
##   (10986 observations deleted due to missingness)
## Multiple R-squared:  0.5659, Adjusted R-squared:  0.5659
## F-statistic:  7114 on 1 and 5456 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(fit.SatVsAct)
```



There are 16015 students have ACT math score, 5512 students have SAT math score, but only 5458 students have both. It will be a huge lose if we use both SAT and ACT math score as explanary variables.

Since we could find a strong correlation between SAT and ACT math score, we can use one to predict the other, or even calculate a mathematic ability score as a predictor.

The paper Dorans, Neil J. "Correspondences between ACTT and SAT® I scores." ETS Research Report Series 1999.1 (1999): i-18. gave a similar result on page 20.

```
fit.GRD <- lm(data = S.A, GRD_PTS_PER_UNIT~HSGPA + LAST_ACT_MATH_SCORE+SEX)
summary(fit.GRD)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ HSGPA + LAST_ACT_MATH_SCORE +
##     SEX, data = S.A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8205 -0.3510  0.2449  0.5594  2.2225
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.101051   0.054735   1.846   0.0649 .
## HSGPA                0.048385   0.005437   8.899   <2e-16 ***
## LAST_ACT_MATH_SCORE  0.098615   0.001863  52.933   <2e-16 ***
## SEXM                -0.194236   0.014462 -13.431   <2e-16 ***
```

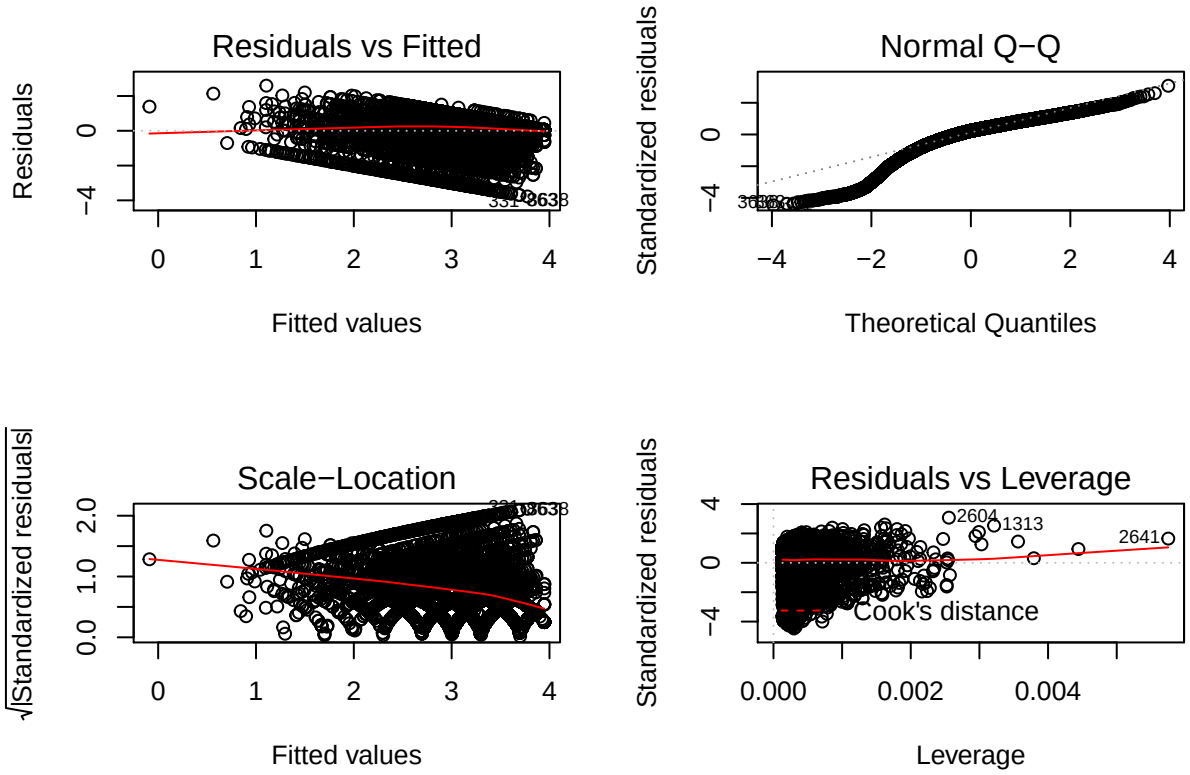
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8889 on 16002 degrees of freedom
##   (438 observations deleted due to missingness)
## Multiple R-squared:  0.1613, Adjusted R-squared:  0.1612
## F-statistic:  1026 on 3 and 16002 DF,  p-value: < 2.2e-16
```

The r^2 is very low, I believe the wrongly placed 2000 '0's under HSGPA should be one of the reason. Let's **try to remove those HSGPA = 0**. This is not the best way to haddle this situation. Since we don't know what casued this error, it might be influce the result if the error were not happens randomly.

```
# filter out those have HSGPA = 0 and ACT math = 0
S.A.c<- filter(S.A, HSGPA>0 & LAST_ACT_MATH_SCORE > 0)
fit.GRDc <- lm(data = S.A.c, GRD_PTS_PER_UNIT~HSGPA + LAST_ACT_MATH_SCORE+SEX)
summary(fit.GRDc)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ HSGPA + LAST_ACT_MATH_SCORE +
##     SEX, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7754 -0.3342  0.1913  0.5481  2.5941
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.677322   0.136383 -26.963   <2e-16 ***
## HSGPA                1.134584   0.036131  31.402   <2e-16 ***
## LAST_ACT_MATH_SCORE  0.085718   0.001987  43.140   <2e-16 ***
## SEXM                -0.138187   0.015010  -9.206   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8451 on 13864 degrees of freedom
## Multiple R-squared:  0.2113, Adjusted R-squared:  0.2112
## F-statistic:  1238 on 3 and 13864 DF,  p-value: < 2.2e-16
```

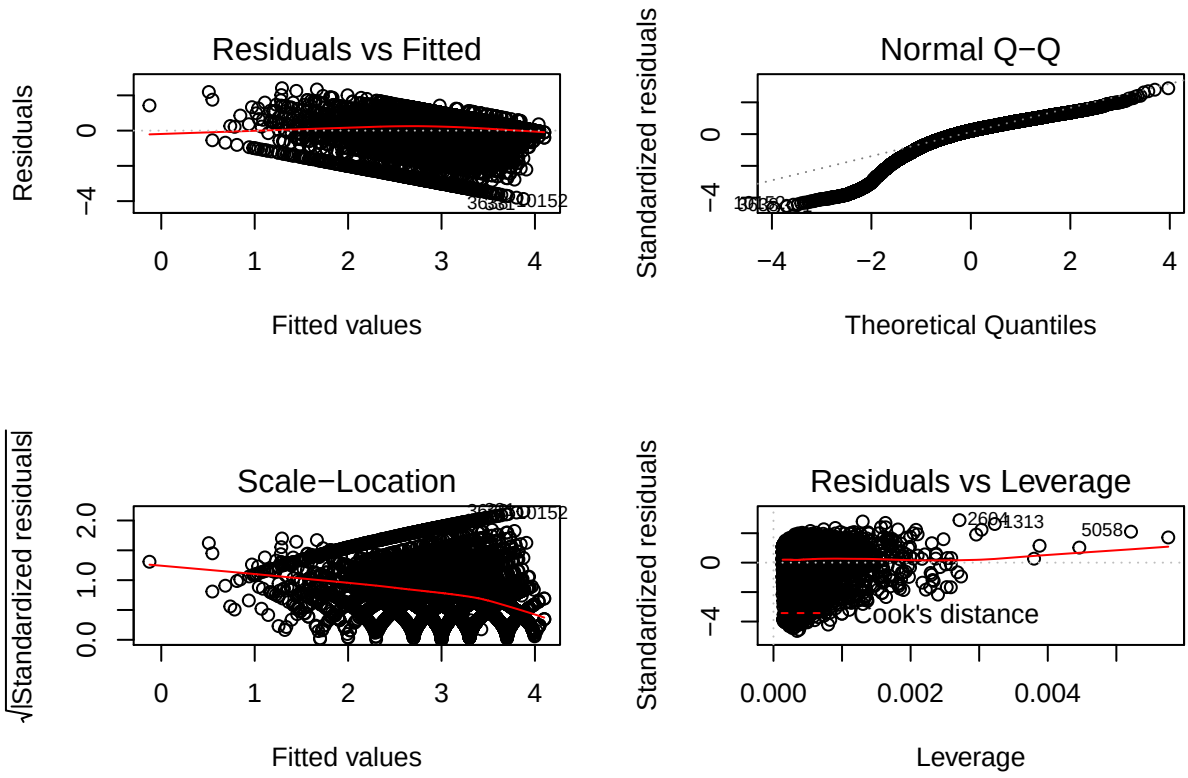
```
par(mfrow=c(2,2))
plot(fit.GRDc)
```




```
fit.GRDb <- lm(data = S.A.c, GRD_PTS_PER_UNIT~HSGPA +LAST_ACT_COMP_SCORE+ LAST_ACT_MATH_SCORE +SEX)
summary(fit.GRDb)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ HSGPA + LAST_ACT_COMP_SCORE +
##     LAST_ACT_MATH_SCORE + SEX, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8766 -0.3184  0.1836  0.5382  2.4078
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.107503   0.137041 -29.973   <2e-16 ***
## HSGPA                1.047822   0.036065  29.054   <2e-16 ***
## LAST_ACT_COMP_SCORE  0.060262   0.003403  17.709   <2e-16 ***
## LAST_ACT_MATH_SCORE  0.051293   0.002764  18.557   <2e-16 ***
## SEXM                -0.131470   0.014849  -8.854   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8358 on 13863 degrees of freedom
## Multiple R-squared:  0.2288, Adjusted R-squared:  0.2286
## F-statistic:  1028 on 4 and 13863 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(fit.GRDb)
```




```r
fit.GRDe <- lm(data = S.A.c, GRD_PTS_PER_UNIT~LAST_ACT_COMP_SCORE);summary(fit.GRDe)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ LAST_ACT_COMP_SCORE, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7740 -0.3474  0.2260  0.5893  2.4057
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.464238   0.071749   -6.47 1.01e-10 ***
## LAST_ACT_COMP_SCORE  0.121091   0.002424   49.96  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.876 on 13866 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1525
## F-statistic:  2496 on 1 and 13866 DF,  p-value: < 2.2e-16
```

```r
fit.GRDf <- lm(data = S.A.c, GRD_PTS_PER_UNIT~HSGPA);summary(fit.GRDf)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ HSGPA, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4074 -0.4073  0.1927  0.5927  2.3929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.59329    0.14119  -18.37   <2e-16 ***
## HSGPA        1.50016    0.03714   40.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9001 on 13866 degrees of freedom
## Multiple R-squared:  0.1053, Adjusted R-squared:  0.1052
## F-statistic:  1631 on 1 and 13866 DF,  p-value: < 2.2e-16
```

```r
fit.GRDg <- lm(data = S.A.c, GRD_PTS_PER_UNIT~LAST_ACT_MATH_SCORE);summary(fit.GRDg)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ LAST_ACT_MATH_SCORE, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7076 -0.3478  0.2522  0.5752  2.0834
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.314024   0.058365    5.38 7.55e-08 ***
## LAST_ACT_MATH_SCORE 0.094267   0.001958   48.15  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8808 on 13866 degrees of freedom
## Multiple R-squared:  0.1432, Adjusted R-squared:  0.1432
## F-statistic:  2318 on 1 and 13866 DF,  p-value: < 2.2e-16
```

```r
fit.GRDe <- lm(data = S.A.c, GRD_PTS_PER_UNIT~LAST_ACT_COMP_SCORE);summary(fit.GRDe)
```

```
##
## Call:
## lm(formula = GRD_PTS_PER_UNIT ~ LAST_ACT_COMP_SCORE, data = S.A.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7740 -0.3474  0.2260  0.5893  2.4057
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.464238   0.071749   -6.47 1.01e-10 ***
## LAST_ACT_COMP_SCORE  0.121091   0.002424   49.96  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.876 on 13866 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1525
## F-statistic:  2496 on 1 and 13866 DF,  p-value: < 2.2e-16
```