Predicting SAT Performance from the TOEFL

Zachary Herold

December 13, 2018

Part 1 - Introduction

The TOEFL examination, or Test of English as a Foreign Language, is a test which measures people's English language skills to see if they are good enough to take a course at university or graduate school in English-speaking countries. it is for people whose native language is not English, and is typically submitted along with the SAT in high school students' unievrsity applications.

The TOEFL is divided into four components (Reading, Listening, Speaking, Writing). Usually taken after the TOEFL, the SAT is divided into three sections (four, when including the optional writing section), colloquially referred to here as Reading, Grammar and Math.

Problem: As the Academic Principal of an International High School in Wuhan, China, I wish to be able to use the data collected to:

(1) Compare the current Class of 2018 to last year's graduating class of 2017. Consider whether differences in SAT and TOEFL performance are stastically significant.

(2) Predict students' SAT scores from TOEFL results. Students will begin a TOEFL-related curriculum in Grade 10 and begin an SAT one in Grade 11.

(3) Determine which TOEFL component most accurately predicts SAT performance, to inform decisions about overall course hours for teachers and more optimal arrangement of teaching staff, assuming allocation of higher quality staffing resources to courses with bigger payoff.

(4) Build an assessment metric for English subject teachers based on student progress in standardized examinations.

```r
knitr::opts_chunk$set(echo = TRUE)
library(readxl)
library(dplyr)
library(ggplot2)
library(kableExtra)
```

Part 2 - Data

I use data collected from the students of the International Department of Wuhan No. 6 High School over the years of 2016 to 2018. I limit the observation subjects to students from the Class of 2017 and 2018 who have verifiable records of having taken both the TOEFL and the SAT exams during their high school study.

The data is recorded in four separate .csv files (one for each class and exam type), which is first cleaned in Excel and then again in R.

The dyplr package is used to summarize the data according to the table below, grouping student results according to their ID number. Altogether there are 23 TOEFL takers from the Class of 2018 and 46 from the class of 2017.

variable	description
`ID`	student ID no.
`no.att.ibt`	number of attempts of TOEFL exam for each student
`max.R`	maximum score on TOEFL reading section for each student
`max.L`	maximum score on TOEFL listening section for each student
`max.S`	maximum score on TOEFL speaking section for each student
`max.W`	maximum score on TOEFL writing section for each student
`max.ibt`	maximum total TOEFL score for each student
`avg.ibt`	mean total TOEFL score for each student
`year`	year of graduation

```
## Cleaning the Class of 2018 TOEFL data

ibt18 <-
  data.frame(read.csv("C:/Users/ZacharyHerold/Documents/DATA606/Final.Project/Class18.ibt.csv"))


names(ibt18)[1] <- c("ID")

ibt18sum <- ibt18 %>%
  group_by(ID) %>%
  summarise(no.att.ibt = max(attempt), max.R = max(na.omit(R)), max.L = max(na.omit(L)),
  max.S = max(na.omit(S)), max.W = max(na.omit(W)), max.ibt = max(T), avg.ibt =
  round(mean(T),3))

ibt18sum$year <- rep("2018", nrow(ibt18sum))

ibt18sum
```

```
## # A tibble: 23 x 9
##       ID no.att.ibt max.R max.L max.S max.W max.ibt avg.ibt year
##    <dbl>      <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl> <chr>
## 1     1           3    30    29    19    29     107    95.7 2018
## 2     2           3    29    29    24    27     109   104.  2018
## 3     4           4    25    21    18    23      86    72.8 2018
## 4     5           5    22    21    19    23      79    74.6 2018
## 5     6           6    23    26    22    22      90    74.8 2018
## 6     7           5    25    24    27    25      94    85.2 2018
## 7     8           1    22    18    16    22      78    78   2018
## 8     9           3    13    18    17    20      65    57.3 2018
## 9    10           6    24    28    22    24      94    78   2018
## 10   11           3    13    16    20    18      67    58   2018
## # ... with 13 more rows
```

```
## Cleaning the Class of 2017 TOEFL data

ibt17 <-
  data.frame(read.csv("C:/Users/ZacharyHerold/Documents/DATA606/Final.Project/Class17.ibt.csv")


ibt17 <- ibt17[1:7]
names(ibt17)[1] <- c("ID")

ibt17sum <- ibt17 %>%
  group_by(ID) %>%
  summarise(no.att.ibt = max(attempt), max.R = max(na.omit(R)), max.L = max(na.omit(L)),
  max.S = max(na.omit(S)), max.W = max(na.omit(W)), max.ibt = max(T), avg.ibt =
  round(mean(T),3))

ibt17sum$year <- rep("2017", nrow(ibt17sum))

ibt17sum
```

```
## # A tibble: 46 x 9
##       ID no.att.ibt max.R max.L max.S max.W max.ibt avg.ibt year
##    <int>      <dbl> <int> <int> <int> <int>   <dbl>   <dbl> <chr>
## 1     1           3    30    28    24    26     108   102.  2017
## 2     2           4    27    29    24    28     103   102.  2017
## 3     3           7    27    25    23    26      98    89   2017
## 4     4           3    28    24    22    28     102    87   2017
## 5     5          10    26    29    24    26     102    93.2 2017
## 6     6           3    27    26    23    24      98    91.7 2017
## 7     7           7    23    19    19    22      74    67.1 2017
## 8     8           5    24    25    23    25      97    86.6 2017
## 9     9           2    30    29    24    28     111   106.  2017
## 10   10           3    26    25    27    28     106    95.7 2017
## # ... with 36 more rows
```

For the SAT

variable	description
`ID`	student ID no.
`no.att.sat`	number of attempts of SAT exam for each student
`max.V`	maximum score on SAT verbal section for each student
`max.sat`	maximum score on SAT for each student
`avg.sat`	mean score on SAT for each student

```
## Cleaning the Class of 2018 SAT data

sat18 <-
  data.frame(read_excel("C:/Users/ZacharyHerold/Documents/DATA606/Final.Project/Class18.sat.xls

sat18 <- sat18[-c(5,37,38,39),]
sat18$attempt <- as.numeric(sat18$attempt)
sat18$V <- sat18$T - sat18$M

colnames(sat18)[1] <- "ID"
sat18sum <- sat18 %>%
  group_by(ID) %>%
  summarise(no.att.sat = max(attempt), max.V = max(na.omit(V)), max.sat = max(T),
  avg.sat = mean(T))

sat18sum
```

```
## # A tibble: 24 x 5
##       ID no.att.sat max.V max.sat avg.sat
##    <dbl>      <dbl> <dbl>   <dbl>   <dbl>
##  1     1          2   690    1490    1485
##  2     2          2   660    1440    1405
##  3     3          1   620    1410    1410
##  4     4          2   570    1350    1345
##  5     5          2   570    1330    1295
##  6     6          1   540    1330    1330
##  7     7          2   540    1280    1280
##  8     8          2   540    1270    1235
##  9     9          2   510    1250    1235
## 10    10          2   530    1310    1255
## # ... with 14 more rows
```

```
## Cleaning the Class of 2018 SAT data
```

```
sat17 <-
  data.frame(read_excel("C:/Users/ZacharyHerold/Documents/DATA606/Final.Project/Class17.sat.xls

sat17 <- sat17[-c(1, 5,37,38,39),]

sat17$attempt <- as.numeric(sat17$attempt)
sat17$R <- as.numeric(sat17$R)
sat17$G <- as.numeric(sat17$G)
sat17$M <- as.numeric(sat17$M)

sat17sum <- sat17 %>%
  group_by(ID) %>%
  summarise(no.att.sat = max(attempt), max.V = max(na.omit(V)), max.sat = max(ttl),
  avg.sat = mean(ttl))

sat17sum
```

```
## # A tibble: 25 x 5
##    ID    no.att.sat max.V max.sat avg.sat
##    <chr>      <dbl> <dbl>   <dbl>   <dbl>
##  1 1              2   690    1490   1490
##  2 10             3   570    1350   1307.
##  3 11             2   640    1400   1385
##  4 12             3   670    1460   1420
##  5 13             2   670    1470   1355
##  6 15             3   580    1350   1293.
##  7 16             1   580    1260   1260
##  8 17             3   620    1390   1390
##  9 18             2   510    1270   1255
## 10 19             3   630    1380   1327.
## # ... with 15 more rows
```

Finally, I merge the class data int two dataframes by student ID, and then rbind it into a single one.

```
## Merging Class Data

class18 <- merge(ibt18sum, sat18sum, by = "ID")
nrow(class18)
```

```
## [1] 22
```

```
class17 <- merge(ibt17sum, sat17sum, by = "ID")
nrow(class17)
```

```
## [1] 24
```

```
ibt.sat <- rbind(class17, class18)
```

Part 3 - Exploratory data analysis

First, I wish to check how reasonable it is to apply the least squares regression, using residual plots, regressing maximum SAT score over maximum TOEFL score.

For the current grade 12:

```
summary(class18)
```

```
##        ID          no.att.ibt        max.R           max.L      
##  Min.   : 1.00   Min.   :1.000   Min.   :10.00   Min.   : 4.00  
##  1st Qu.: 7.25   1st Qu.:2.250   1st Qu.:15.00   1st Qu.:16.00  
##  Median :12.50   Median :3.000   Median :22.00   Median :20.00  
##  Mean   :12.64   Mean   :3.545   Mean   :20.05   Mean   :19.41  
##  3rd Qu.:17.75   3rd Qu.:5.000   3rd Qu.:24.00   3rd Qu.:23.50  
##  Max.   :25.00   Max.   :6.000   Max.   :30.00   Max.   :29.00  
##      max.S           max.W       max.ibt          avg.ibt      
##  Min.   :15.00   Min.   :15   Min.   : 51.00   Min.   : 47.50  
##  1st Qu.:17.00   1st Qu.:18   1st Qu.: 64.25   1st Qu.: 57.50  
##  Median :19.00   Median :21   Median : 78.50   Median : 73.21  
##  Mean   :19.27   Mean   :21   Mean   : 77.68   Mean   : 69.86  
##  3rd Qu.:21.50   3rd Qu.:23   3rd Qu.: 89.50   3rd Qu.: 78.00  
##  Max.   :27.00   Max.   :29   Max.   :109.00   Max.   :103.67  
##      year             no.att.sat        max.V          max.sat    
##  Length:22          Min.   :1.000   Min.   :370.0   Min.   : 930  
##  Class :character   1st Qu.:2.000   1st Qu.:482.5   1st Qu.:1135  
##  Mode  :character   Median :2.000   Median :510.0   Median :1235  
##                     Mean   :1.818   Mean   :517.3   Mean   :1229  
##                     3rd Qu.:2.000   3rd Qu.:540.0   3rd Qu.:1325  
##                     Max.   :2.000   Max.   :690.0   Max.   :1490  
##     avg.sat    
##  Min.   : 930  
##  1st Qu.:1125  
##  Median :1215  
##  Mean   :1212  
##  3rd Qu.:1291  
##  Max.   :1485
```

The mean maximum SAT score by student is 1229. The overall average when combining the average results of each student is 1212.

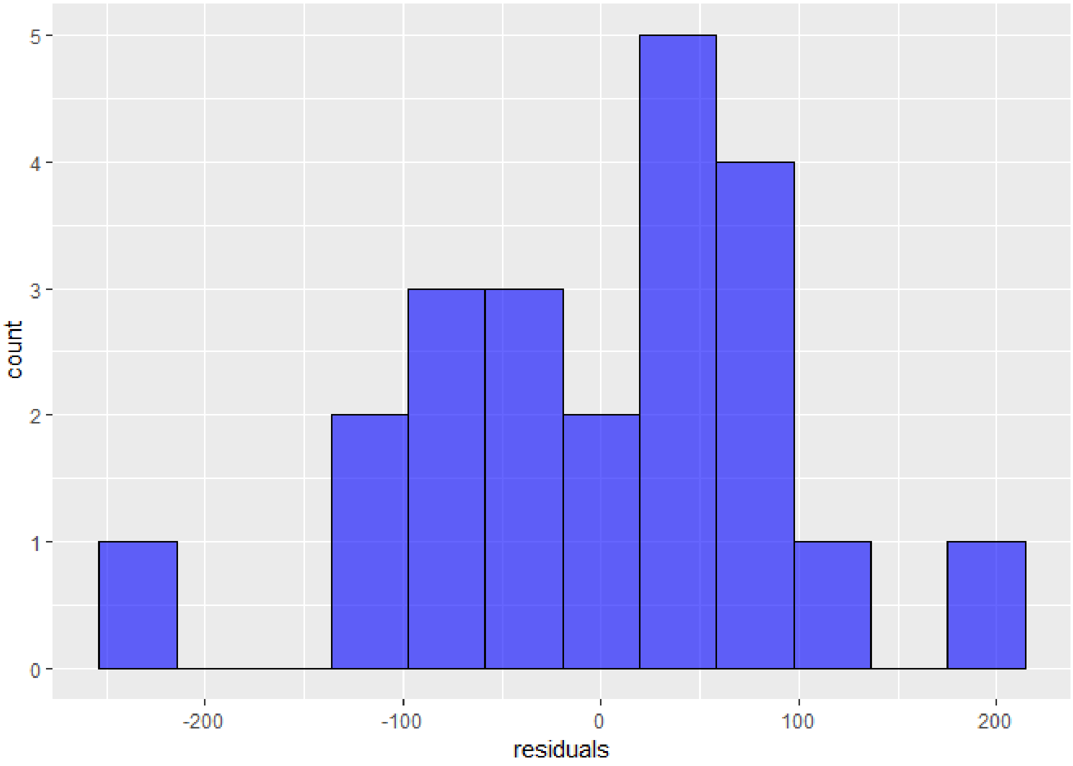
```
m_sat <- lm(max.sat ~ max.ibt, data = class18)
summary(m_sat)
```

```
##
## Call:
## lm(formula = max.sat ~ max.ibt, data = class18)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -220.27  -59.93   21.08   68.77  208.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  814.007     99.584   8.174 8.35e-08 ***
## max.ibt        5.338      1.253   4.261 0.000382 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99.2 on 20 degrees of freedom
## Multiple R-squared:  0.4758, Adjusted R-squared:  0.4496
## F-statistic: 18.15 on 1 and 20 DF,  p-value: 0.0003823
```

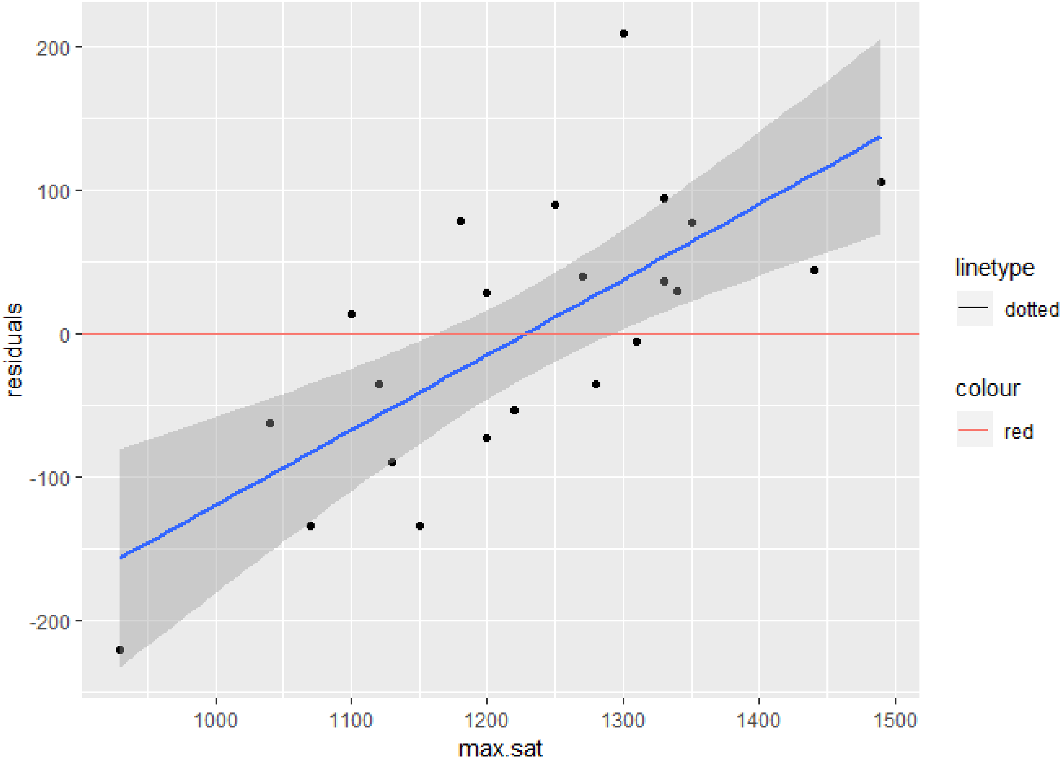
$$\hat{y} = 814.007 + 5.338 * max.ibt$$

```
class18$pred_sat <- 814.007 + 5.338 * class18$max.ibt
class18$residuals <- class18$max.sat - class18$pred_sat

ggplot(class18, aes(residuals)) +
geom_histogram(bins = 12, fill="blue", colour="black", alpha =0.6)
```



```
sp <- ggplot(class18, aes(x=max.sat, y=residuals))
sp + geom_point() + stat_smooth(method=lm, se=T) + geom_hline(aes(yintercept=0,
  colour="red", linetype="dotted"))
```



Linearity: With a moderate R-squared (0.4758), there is some support for goodness of fit of this linear model

Nearly Normal Residuals: Residual distribution appears more of less normal.

Constant Variability: This condition appears not to be met, with some evidence of heteroskedasticity in the scatterplot above. Residuals are increasing positively as SAT Max increases.

For last year’s grade 12:

```
summary(class17)
```

```
##        ID          no.att.ibt         max.R           max.L      
##  Min.   : 1.00   Min.   : 2.000   Min.   :19.00   Min.   :17.00  
##  1st Qu.: 6.75   1st Qu.: 3.000   1st Qu.:23.75   1st Qu.:24.00  
##  Median :12.50   Median : 3.000   Median :27.00   Median :25.00  
##  Mean   :15.12   Mean   : 3.917   Mean   :25.67   Mean   :24.62  
##  3rd Qu.:19.25   3rd Qu.: 5.000   3rd Qu.:28.00   3rd Qu.:26.25  
##  Max.   :43.00   Max.   :10.000   Max.   :30.00   Max.   :30.00  
##      max.S           max.W          max.ibt          avg.ibt      
##  Min.   :19.00   Min.   :18.00   Min.   : 74.00   Min.   : 65.00  
##  1st Qu.:22.00   1st Qu.:24.00   1st Qu.: 91.75   1st Qu.: 82.35  
##  Median :23.00   Median :25.00   Median : 98.00   Median : 89.00  
##  Mean   :22.71   Mean   :24.83   Mean   : 95.92   Mean   : 87.93  
##  3rd Qu.:24.00   3rd Qu.:26.25   3rd Qu.:103.00   3rd Qu.: 93.46  
##  Max.   :27.00   Max.   :28.00   Max.   :111.00   Max.   :106.50  
##      year             no.att.sat     max.V          max.sat    
##  Length:24          Min.   :1    Min.   :510.0   Min.   :1210  
##  Class :character   1st Qu.:1    1st Qu.:567.5   1st Qu.:1310  
##  Mode  :character   Median :2    Median :595.0   Median :1370  
##                     Mean   :2    Mean   :597.9   Mean   :1362  
##                     3rd Qu.:3    3rd Qu.:640.0   3rd Qu.:1408  
##                     Max.   :3    Max.   :690.0   Max.   :1490  
##     avg.sat    
##  Min.   :1210  
##  1st Qu.:1292  
##  Median :1320  
##  Mean   :1337  
##  3rd Qu.:1386  
##  Max.   :1490
```

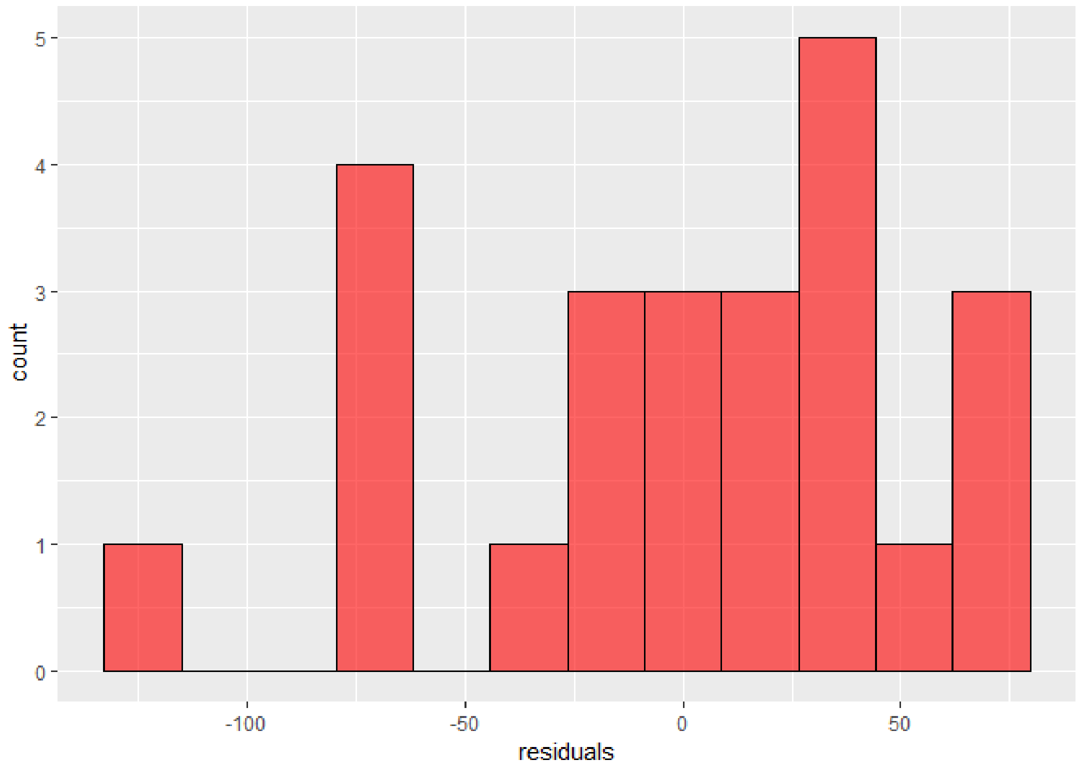
The mean maximum SAT score by student is 1362. The overall average when combining the average results of each student is 1337.

```
m_sat <- lm(max.sat ~ max.ibt, data = class17)
summary(m_sat)
```

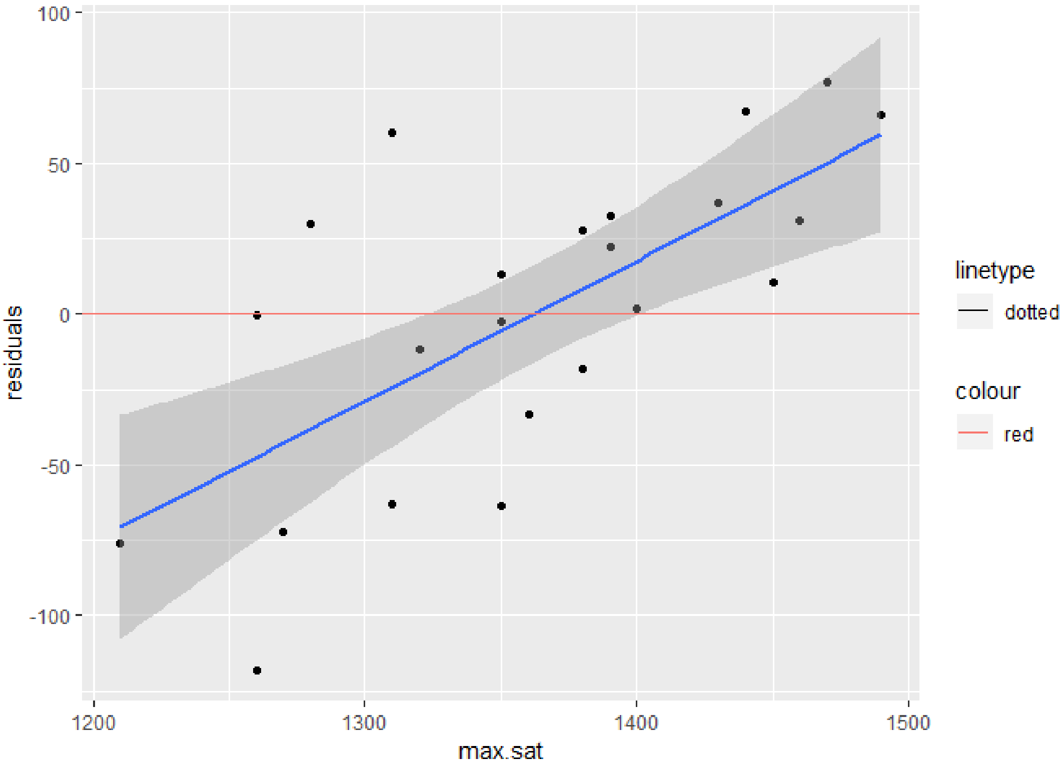
```
##
## Call:
## lm(formula = max.sat ~ max.ibt, data = class17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.863  -22.055    6.195   31.371   76.784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  871.204     98.146   8.877 1.01e-08 ***
## max.ibt        5.118      1.017   5.031 4.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51.77 on 22 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.5138
## F-statistic: 25.31 on 1 and 22 DF,  p-value: 4.89e-05
```

```
class17$pred_sat <- 871.204 + 5.118 * class17$max.ibt
class17$residuals <- class17$max.sat - class17$pred_sat

ggplot(class17, aes(residuals)) +
geom_histogram(bins = 12, fill="red", colour="black", alpha =0.6)
```



```
sp <- ggplot(class17, aes(x=max.sat, y=residuals))
sp + geom_point() + stat_smooth(method=lm, se=T) + geom_hline(aes(yintercept=0,
  colour="red", linetype="dotted"))
```



Linearity: With a moderate R-squared (0.535), there is some support for goodness of fit of this linear model

Nearly Normal Residuals: Residual distribution appears to be more uniform that normal.

Constant Variability: This condition appears not to be met, likely due to the bound range of possible SAT scores (400-1600).

Part 4 - Inference

(1) Compare the current Class of 2018 to last year’s graduating class of 2017. Consider whether differences in SAT and TOEFL performance are stastically significant.

I wish to determine if the difference in mean max SAT between the two classes is likely due to chance or attributes inherent to each cohort.

My null hypothesis is that they are equivalent.

```
mean.sat.diff <- mean(class17$max.sat) - mean(class18$max.sat)
print(paste0("The difference in mean max SAT before the two grades is: ",
  mean.sat.diff))
```

```
## [1] "The difference in mean max SAT before the two grades is: 133.44696969697"
```

I construct a 95% confidence interval, using my most conservative estimate of a t-score with 20 degrees of freedom. I use t-scores due to the relatively low sample sizes. The 95% confidence interval does not intersect with 0, providing us with enough evidence to reject the null hypothesis. In other words there is sufficient evidence to lead us to believe the difference in means is not due to random noise.

```
var18 <- round(var(class18$max.sat, na.rm=T),3)
var17 <- round(var(class17$max.sat, na.rm=T),3)
se <- round(sqrt((var17/ 24) + (var18/22)),3)
t.score <- round(abs(qt(0.025, 20)),3)
print(paste0("estimated critical t-score: ", t.score))
```

```
## [1] "estimated critical t-score: 2.086"
```

```
print(paste0("standard error: ", se))
```

```
## [1] "standard error: 32.286"
```

```
low.bound <- round((mean.sat.diff - t.score * se), 3)
high.bound <- round((mean.sat.diff + t.score * se), 3)
print(paste0("95% confidence interval: ", low.bound, ", ", high.bound))
```

```
## [1] "95% confidence interval: 66.098, 200.796"
```

(2) Predict students’ SAT scores from TOEFL results. Students will begin a TOEFL-related curriculum in Grade 10 and begin an SAT one in Grade 11.

```
str(ibt.sat)
```

```
## 'data.frame':    46 obs. of  13 variables:
##  $ ID        : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ no.att.ibt: num  3 4 7 3 10 3 7 5 2 3 ...
##  $ max.R     : num  30 27 27 28 26 27 23 24 30 26 ...
##  $ max.L     : num  28 29 25 24 29 26 19 25 29 25 ...
##  $ max.S     : num  24 24 23 22 24 23 19 23 24 27 ...
##  $ max.W     : num  26 28 26 28 26 24 22 25 28 28 ...
##  $ max.ibt   : num  108 103 98 102 102 98 74 97 111 106 ...
##  $ avg.ibt   : num  101.7 101.8 89 87 93.2 ...
##  $ year      : chr  "2017" "2017" "2017" "2017" ...
##  $ no.att.sat: num  2 2 2 2 3 1 2 3 2 3 ...
##  $ max.V     : num  690 620 640 650 580 570 560 610 650 570 ...
##  $ max.sat   : num  1490 1380 1440 1430 1360 1310 1310 1390 1450 1350 ...
##  $ avg.sat   : num  1490 1335 1425 1425 1320 ...
```

```
ibt.sat$year <- as.factor(ibt.sat$year)
levels(ibt.sat$year)
```

```
## [1] "2017" "2018"
```

```
ibt.sat$ID <- as.factor(ibt.sat$ID)

head(ibt.sat,10)
```

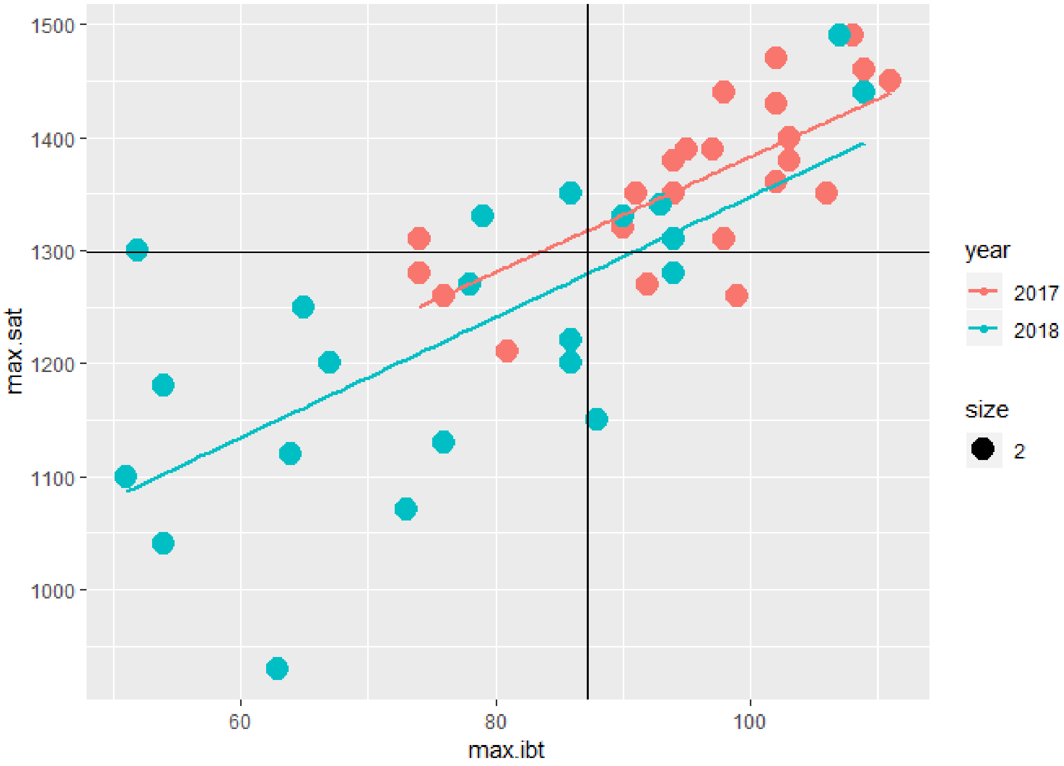
```
##    ID no.att.ibt max.R max.L max.S max.W max.ibt avg.ibt year no.att.sat
## 1   1          3    30    28    24    26     108 101.667 2017          2
## 2   2          4    27    29    24    28     103 101.750 2017          2
## 3   3          7    27    25    23    26      98  89.000 2017          2
## 4   4          3    28    24    22    28     102  87.000 2017          2
## 5   5         10    26    29    24    26     102  93.200 2017          3
## 6   6          3    27    26    23    24      98  91.667 2017          1
## 7   7          7    23    19    19    22      74  67.143 2017          2
## 8   8          5    24    25    23    25      97  86.600 2017          3
## 9   9          2    30    29    24    28     111 106.500 2017          2
## 10 10          3    26    25    27    28     106  95.667 2017          3
##    max.V max.sat  avg.sat
## 1    690    1490 1490.000
## 2    620    1380 1335.000
## 3    640    1440 1425.000
## 4    650    1430 1425.000
## 5    580    1360 1320.000
## 6    570    1310 1310.000
## 7    560    1310 1300.000
## 8    610    1390 1316.000
## 9    650    1450 1445.000
## 10   570    1350 1306.667
```

First we check the correlation between maximum TOEFL and maximum SAT results.

```
cor(ibt.sat$max.ibt,y=ibt.sat$max.sat)
```

```
## [1] 0.7887599
```

```
p0 <- ggplot(ibt.sat, aes(x=max.ibt, y=max.sat, colour = year, size = 2)) +
  geom_point()  + stat_smooth(method=lm, se=F, size = 1) +
  geom_vline(xintercept = mean(ibt.sat$max.ibt, na.rm = T), mapping=NULL) +
  geom_hline(yintercept = mean(ibt.sat$max.sat, na.rm = T), mapping=NULL)
p0
```



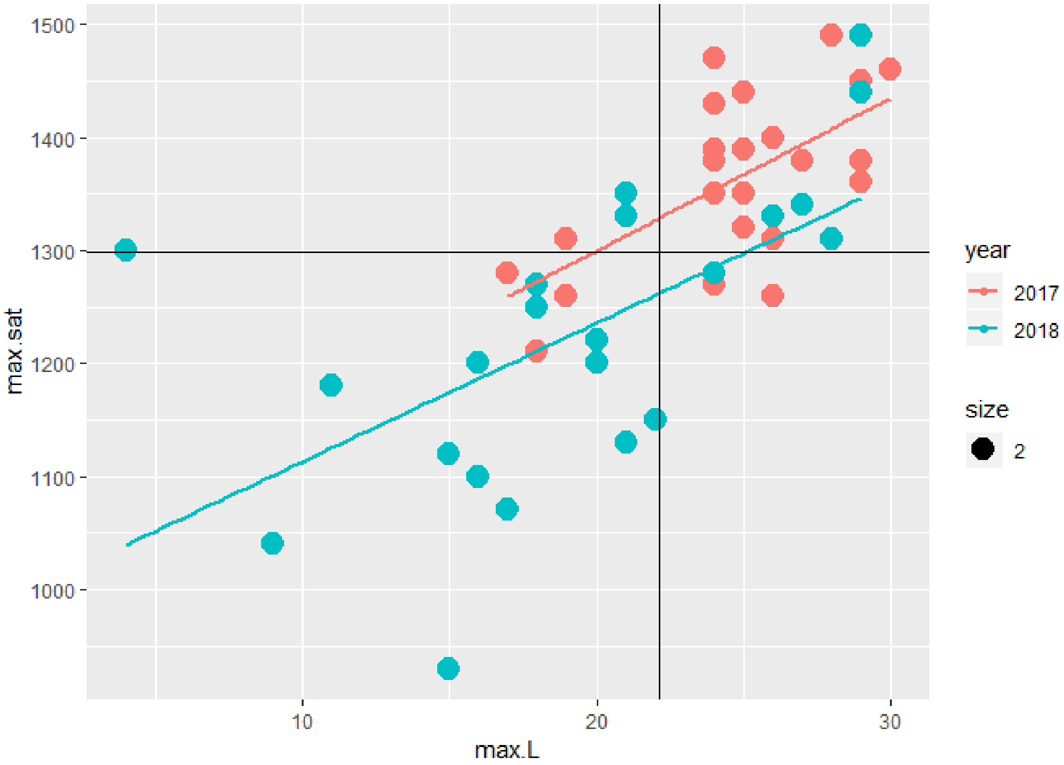
```
cor(ibt.sat$max.S,y=ibt.sat$max.sat)
```

```
## [1] 0.52249
```

A positive correlation between max TOEFL speaking and max overall SAT scores.

When testing for correlation, one notes that most of the overall, as well as the Listening results for the Class of 2018 are below the mean.

```
p1 <- ggplot(ibt.sat, aes(x=max.L,y=max.sat, colour = year, size = 2)) +
  geom_point()  + stat_smooth(method=lm, se=F, size = 1) +
  geom_vline(xintercept = mean(ibt.sat$max.L, na.rm = T), mapping=NULL) +
  geom_hline(yintercept = mean(ibt.sat$max.sat, na.rm = T), mapping=NULL)
p1
```



From the scatterplot of the data, we notice that the Class of 2017 have a more concentrated cluster. The class of 2018 has lower results across the board. The regression lines do not intersect, suggesting that year of graduation is a significant determinant of maximum SAT score.

```
m_sat.all <- lm(max.sat ~ no.att.ibt + max.R + max.L + max.S + max.W + max.ibt + avg.ibt
  + no.att.sat + year, data = ibt.sat)
summary(m_sat.all)
```

```
##
## Call:
## lm(formula = max.sat ~ no.att.ibt + max.R + max.L + max.S + max.W +
##     max.ibt + avg.ibt + no.att.sat + year, data = ibt.sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.214  -44.526   -0.522   45.362  121.099
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  829.997    112.238   7.395    1e-08 ***
## no.att.ibt    -4.286      8.317  -0.515   0.6095  
## max.R          7.739      7.586   1.020   0.3145  
## max.L         -1.664      6.127  -0.272   0.7875  
## max.S        -10.351      8.106  -1.277   0.2098  
## max.W         21.485      7.950   2.703   0.0104 *
## max.ibt        2.769      4.583   0.604   0.5494  
## avg.ibt       -2.314      2.869  -0.806   0.4253  
## no.att.sat    15.282     18.851   0.811   0.4229  
## year2018     -41.958     25.116  -1.671   0.1035  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 66.31 on 36 degrees of freedom
## Multiple R-squared:  0.776,  Adjusted R-squared:   0.72 
## F-statistic: 13.86 on 9 and 36 DF,  p-value: 2.804e-09
```

I now perform a multivariate linear regression on the maximum SAT results, using 9 variables, removing the collinear max.V and avg.sat. Year is the one categorical variable. Adjusted R-squared is 0.72 given this set of explanatory variables.

Reviewing the regression coefficients it is surprising to see negative values for total number of attempts, maximum TOEFL Listening score and maximum TOEFL Speaking score. This suggests that students actually perform worse on the SAT the more TOEFL attempts they make, and the HIGHER they performing on the Speaking and Listening sections.

The reduction in SAT score according to number of TOEFL attempts may be explained by the fact they those students spend more time on the TOEFL curriculum, allowing less preparation time for SAT-related content.

I try to optimize the regression model through backward elimination, removing in order the: (1) TOEFL Listening max score (improving to Adjusted R-squared of 0.727), (2) overall TOEFL max score (to 0.7318), (3) number of attempts on TOEFL (to 0.7349), (4) average TOEFL score (to 0.7401), and (5) number of attempts on TOEFL (to 0.7426).

The model performs worse when the year factor is removed. We are left with a robust function of moderately strong evidence of goodness of fit with the explanatory variables of TOEFL Reading, TOEFL Speaking, TOEFL Writing and class.

```
m_sat.all2 <- lm(max.sat ~ max.R + max.S + max.W + year, data = ibt.sat)
summary(m_sat.all2)
```

```
## 
## Call:
## lm(formula = max.sat ~ max.R + max.S + max.W + year, data = ibt.sat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -147.59  -40.47   -3.40   42.76  132.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  841.062     96.595   8.707 7.24e-11 ***
## max.R          7.681      3.523   2.180 0.035035 *
## max.S        -10.955      4.683  -2.339 0.024271 *
## max.W         23.059      5.650   4.081 0.000202 ***
## year2018     -39.512     23.427  -1.687 0.099275 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.58 on 41 degrees of freedom
## Multiple R-squared:  0.7655, Adjusted R-squared:  0.7426
## F-statistic: 33.45 on 4 and 41 DF,  p-value: 2.053e-12
```

The year2018 variable if given a value of 1 if true and 0 if false (Class of 2017).

The regression line is determined as:

$$\hat{y} = 841.062 + 7.681 * max.R + -10.955 * max.S + 23.059 * max.W + -39.512 * year2018$$

Again we see this peculiar phenonemon of students expecting a lower maximum SAT score the better they perform on the TOEFL Speaking section! Clearly being in the Class of 2018 also has a negative impact, due to characteristics inherent to that class or perhaps reflecting the quality of instruction for that cohort.

(3) Determine which TOEFL component most accurately predicts SAT performance, to inform decisions about overall course hours for teachers and more optimal arrangement of teaching staff, assuming allocation of higher quality staffing resources to courses with bigger payoff.

I seek to decompose the results a bit further on a class-by-class basis.

CLass of 2017 Regression of SAT max on TOEFL max

```
m_ibt17 <- lm(class17$max.sat ~ class17$max.S)
summary(m_ibt17)
```

```
##
## Call:
## lm(formula = class17$max.sat ~ class17$max.S)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.823  -48.980   -0.897   65.251  111.030
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1065.198    178.976   5.952 5.46e-06 ***
## class17$max.S   13.074      7.855   1.664     0.11    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 71.55 on 22 degrees of freedom
## Multiple R-squared:  0.1118, Adjusted R-squared:  0.07146 
## F-statistic:  2.77 on 1 and 22 DF,  p-value: 0.1102
```

TOEFL Reading = Multiple R-squared: 0.5445

TOEFL Total (Max) = Multiple R-squared: 0.535

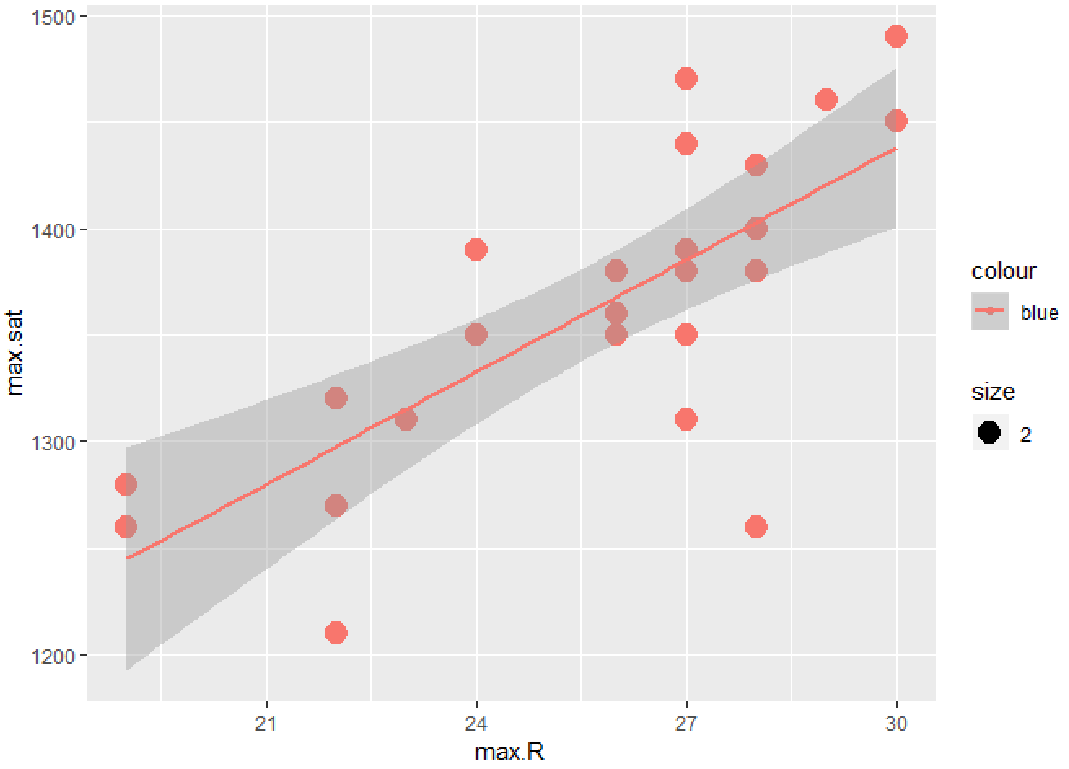
TOEFL Writing = Multiple R-squared: 0.4626

TOEFL Listening = Multiple R-squared: 0.4724

TOEFL Speaking = Multiple R-squared: 0.1118

For the class of 2017, TOEFL Reading score outperforms the overall TOEFL score as a predictor of maximum SAT. This visualization shows a relatively close fit to the regression line.

```
p1 <- ggplot(class17, aes(x=max.R,y=max.sat, colour = "blue", size = 2)) +
  geom_point()  + stat_smooth(method=lm, se=T, size = 1)
p1
```



CLass of 2018 Regression of SAT max on TOEFL max

```
m_ibt18 <- lm(class18$max.sat ~ class18$max.S)
summary(m_ibt18)
```

```
##
## Call:
## lm(formula = class18$max.sat ~ class18$max.S)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295.07  -71.08  -18.54   97.43  264.93
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    976.858    175.355   5.571 1.88e-05 ***
## class18$max.S   13.064      8.984   1.454    0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130.3 on 20 degrees of freedom
## Multiple R-squared:  0.09562,    Adjusted R-squared:  0.0504
## F-statistic: 2.115 on 1 and 20 DF,  p-value: 0.1614
```

TOEFL Writing = Multiple R-squared: 0.687

TOEFL Reading = Multiple R-squared: 0.518

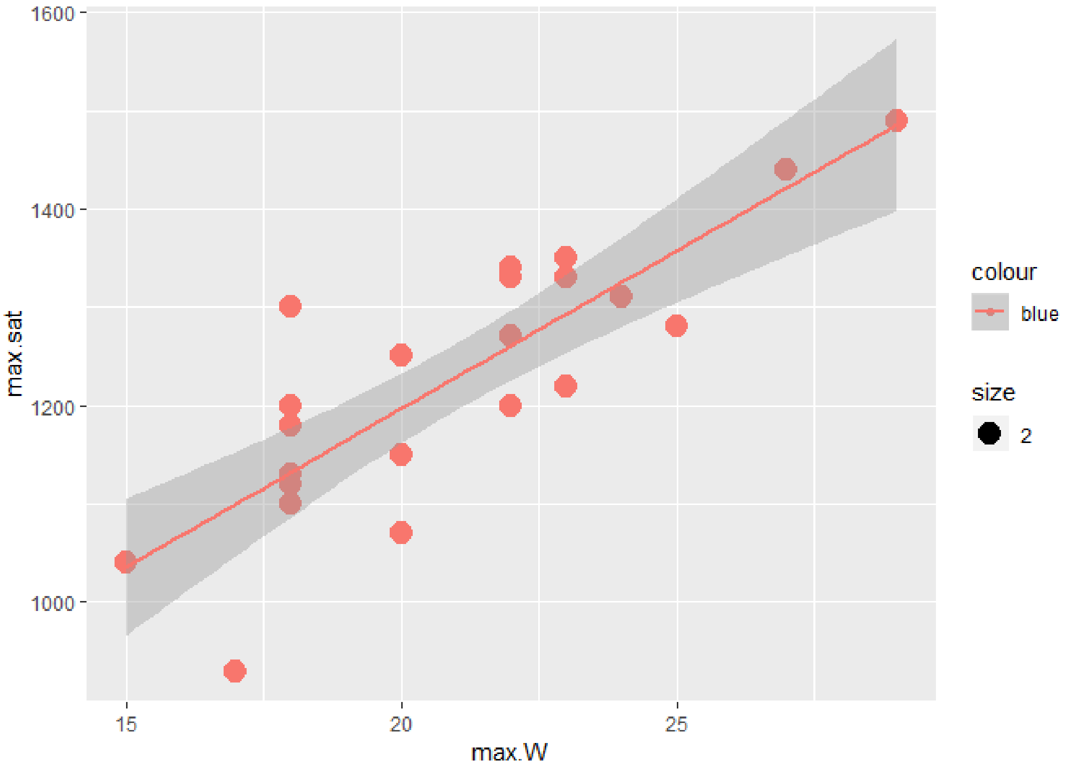
TOEFL Total (Max) = Multiple R-squared: 0.4758

TOEFL Listening = Multiple R-squared: 0.3551

TOEFL Speaking = Multiple R-squared: 0.09562

For the class of 2018, both TOEFL Writing and Reading score outperform the overall TOEFL score as a predictor of maximum SAT. This visualization shows a relatively close fit to the Writing-explained regression line.

```
p2 <- ggplot(class18, aes(x=max.W,y=max.sat, colour = "blue", size = 2)) +
  geom_point()  + stat_smooth(method=lm, se=T, size = 1)
p2
```



(4) Build an assessment metric for English subject teachers based on student progress in standardized examinations.

This fourth research question remains a topic for further investigation. I would want to get a track record to additional years TOEFL and SAT scores to build out the sample size. One would have to read studies related to the implications on Progress metrics in teacher evaluation.

Part 5 - Conclusions

(1) The large disparity in the 2017 and 2018 cohorts is likely attributable to some sort of survivorship basis, and would need to be considered in the context of this particular school. Certain reasons that could exist for the proportionally higher over-performance on the TOEFL may be due to the fact that in years prior “low-TOEFL” students had a direct placement at American universities without needing SAT scores for admissions. In this year,many students were undecided about that program, leading to greater incidence of underperforming students taking the SAT exam. If may also be possible that students may have dropped out in years prior if their TOEFL is below a certain score, but the current class did not witness this phenomenon.

(2) The best-fit model for the data was a multivariate linear regression, with maximum TOEFL Reading, Speaking and Listening scores, as well as class year, being the most prevalent predictors, as determined through backward elimination of high p-score variables. One saw an unexpected result that TOEFL Speaking scores had a negative effect on overall max SAT scores. One would have to run more tests and consider if Simpson’s Paradox occurred.

(3) Different classes may have skills which bear a stronger correlation to the max SAT score. For Class 2017, Reading was the strongest predictor, while for Class 2018, it was Writing. The

R-Squared score of approximately 50% for the overall TOEFL score, suggests that certain skills of the TOEFL suggests that certain TOEFL skills are not relevant to the SAT.

References

Inhouse data-collection.

~Zachary Herold