|  |
| --- |
| SFU |
| BUEC333 Assignment 1 |
| Kun Yang 301178299 |
|  |
|  |
| **[Pick the date]** |

|  |
| --- |
|  |

**I attach the R code and the copy of output of R in each question. And I am a STAT minor student, so I use the knowledge of R from my previous STAT courses (STAT 445 and 475).**

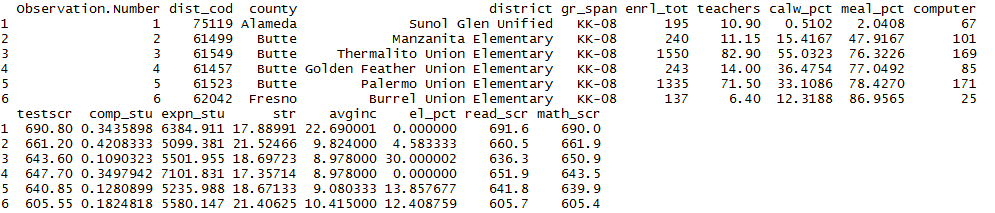
1. ***From the output from that last command, you can read the maximum value for “comp\_stu". What is that value? What does it mean? Are you surprised?***

**> rm(list=ls())**

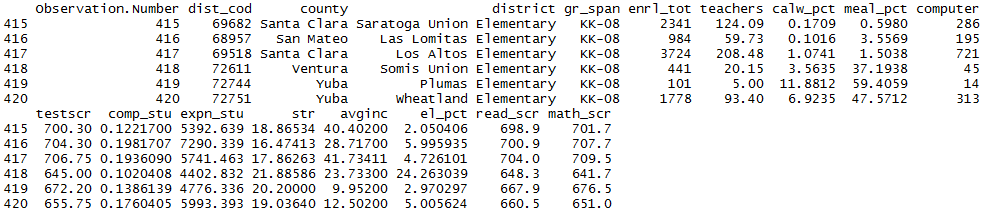
**> setwd("C:\\Users\\Kun\\Desktop\\333\\1")**

**> csdata<-read.csv("testscores\_california\_1999.csv")**

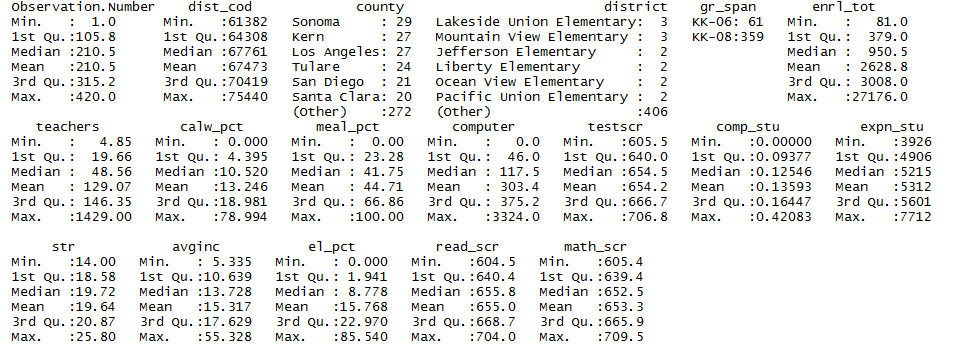
**> head(csdata)**



**> tail(csdata)**



**> summary(csdata)**



The maximum value of “comp\_stu" is 0.42083. It means that there are 0.42083 computers per student in that school. And it is the highest value among the dataset. I am not surprised because nearly two students share one public computer is pretty normal.

1. ***Give the sample mean, minimum, maximum, and sample standard deviation of “math\_scr”.***

**> summary(csdata$math\_sc)**

Sample mean: 653.3

Sample minimum: 605.4

Sample maximum: 709.5

Sample standard deviation: 18.7542

Min. 1st Qu. Median Mean 3rd Qu. Max.

605.4 639.4 652.4 653.3 665.8 709.5

**> sd(csdata$math\_sc)**

[1] 18.7542

1. ***Is the sample standard deviation you computed under (2) an estimand, an estimator, or an estimate?***

Estimand is the fixed but unknown population quantity. Estimator is the function of the data used to make a guess for the estimand. An estimate is what results from applying the estimator to specific measurements. (Definition from the lecture slides)

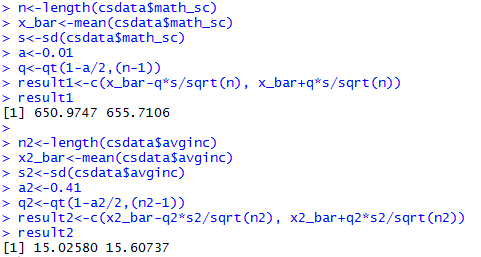
Therefore, the value of the sample standard deviation 18.7542 is an estimate.

1. ***Is the number you got a random variable?***

A random variable is a numerical measure of a random outcome. (Definition from the lecture slides)

Thus, the number which I got for the sample standard deviation (18.7542) is a random variable.

1. ***Construct a 99% confidence interval for the mean of “math\_scr”. Also, construct a 59% confidence interval for the mean of “avg\_inc".***



A 99% confidence interval for the mean of “math\_scr” is (650.9747, 655.7106)

A 59% confidence interval for the mean of “avg\_inc" is (15.02580, 15.60737)

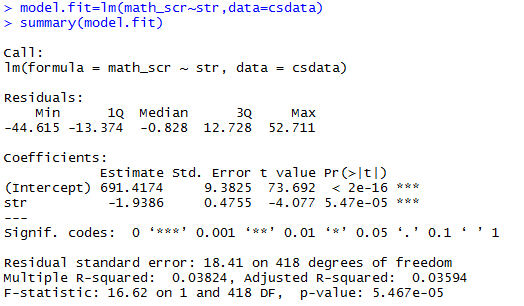
1. ***For the previous question, what is the population you chose?***

The population is the entire group of units of interest. In this assignment, I am investigating the effect of class size on student performance. (In the assignment description, it does not mention that our investigation focus on the whole world or just on united states) Therefore, the population is all students in general.

1. ***From question (5): What is the interpretation of the 99% confidence interval that you computed?***

We are 99% confident to say that the true population mean of average math score is between 650.9747 and 655.7106. Or in other words, if we repeat this procedure many times, the confidence interval will contain the true population mean 99% of the time.

1. ***Estimate the coefficients in the linear regression of “math\_scr" on “str". Include a constant. Report the coefficient estimates and standard errors, and interpret the coefficient estimates.***



The Y-intercept, the expected value for the average math score is 691.4174 when student teacher ratio is zero. It is meaningless because when str equals to zero then there are no students. So it does not have a useful interpretation.

The average math scores are expected to decline by 1.9386 for a one unit increase in student teacher ratio.

1. ***Construct a 95% confidence interval for the regression coefficient of “str".***

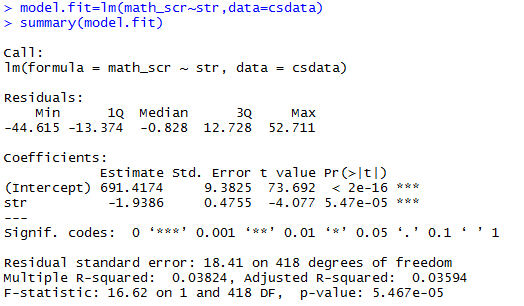
**> confint(model.fit,'str',level=0.95)**

A 95% confidence interval for the regression coefficient of “str" is (-2.873292, -1.003889)

2.5 % 97.5 %

str -2.873292 -1.003889

1. ***You expect the coefficient of str to be negative. Formulate an appropriate null and alternative hypothesis; formulate a decision rule (use significance level of 5%); use R to compute the necessary values; draw the conclusion.***



1. ***What is the interpretation of the conclusion of the test in (10)? What do you conclude about the effect of class size on student test performance?***

The null hypothesis is rejected. We are 95% confident to say that the student teacher ratio and average math scores have a linear relationship. In this case, the student teacher ratio has negative effect on average math scores, so as the class size increase (other variables remain constant) the average math scores decrease.

1. ***There other variables that could have an effect on student test scores. List two or three such variables in this data set, and explain why they could have an effect on test scores, Do you expect the effects to be positive or negative?***

COMP\_STU: computers per student

As the computer per student increases, the students will have more opportunity to study on-line or do assignments on computer. So we can expect it has **positive** effect on the test scores.

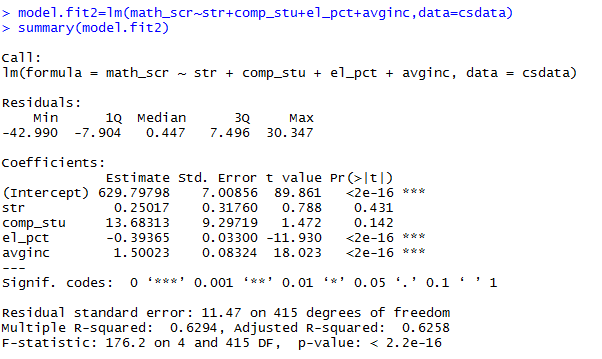
el pct: Percentage of English learners

The percentage of English learners increases means that there are less percent of English native speakers. So we can expect it has **negative** effect on the test scores.

AVGINC: District average income (IN $1000'S)

As the district average income increase, we expect that the students’ family will spend more money on children’s education. So the test scores will increase. The average income has **positive** effect.

1. ***Estimate the coefficients in the linear regression of “math\_scr" on “str" and the variables you came up with under (12). Include a constant. Report the coefficient estimate and standard error for “str".***



1. ***For bonus points, explain (or speculate about) the difference between your answer under (13) versus (8).***

In question (8) the average math scores has negative linear relationship with the student teacher ratio and the null hypothesis that beta1=0 was rejected. So we are confident to say the relationship exists.

In part (13), however, when we set 4 independent variables: “student teacher ratio”, “computers per student”, “percentage of English learners” and “district average income”, the student teacher ratio are expected to has positive relationship with the average math scores, which is opposite to the result in part(8). And the null hypothesis that beta1=0 was not rejected. So there are NOT enough evidences show that the relationship between the student teacher ratio and the average math scores exist. It means that when we consider more factors of effect in this regression problem, the effect of student teacher ratio become less significant.

1. ***If you could gather additional data to answer this question in a better way, what data would you gather?***

Because our investigation is focus on the effect of class size on scores (in general not just US.), I would like to gather more data from Canada, Britain, Australia, etc. The reason is that if we only focus on the data from California, the data may be biased. If our topic only focuses on US domestic students, we would better add some datasets from different states across different geographic regions.