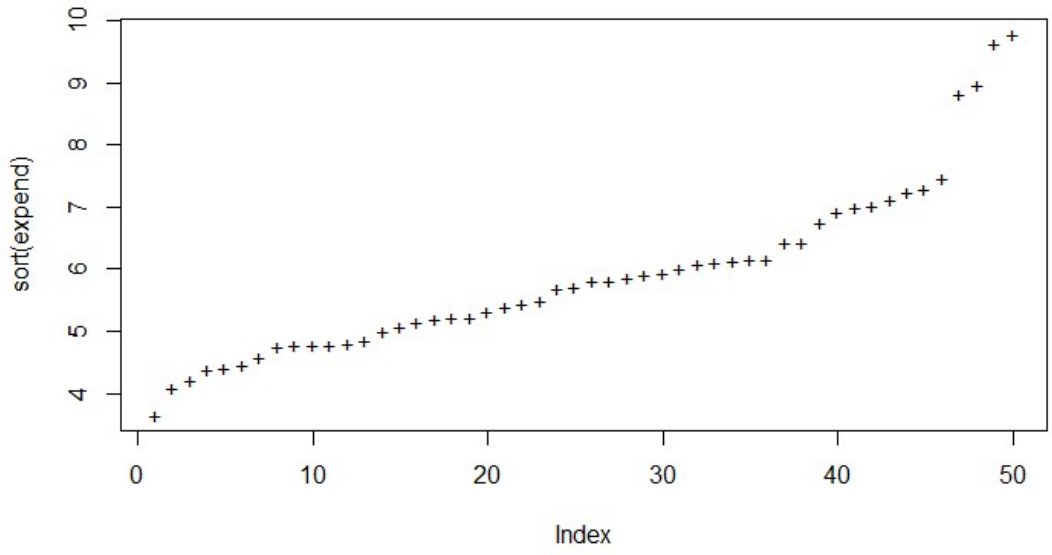
NTHU STAT 5410 - Linear Models

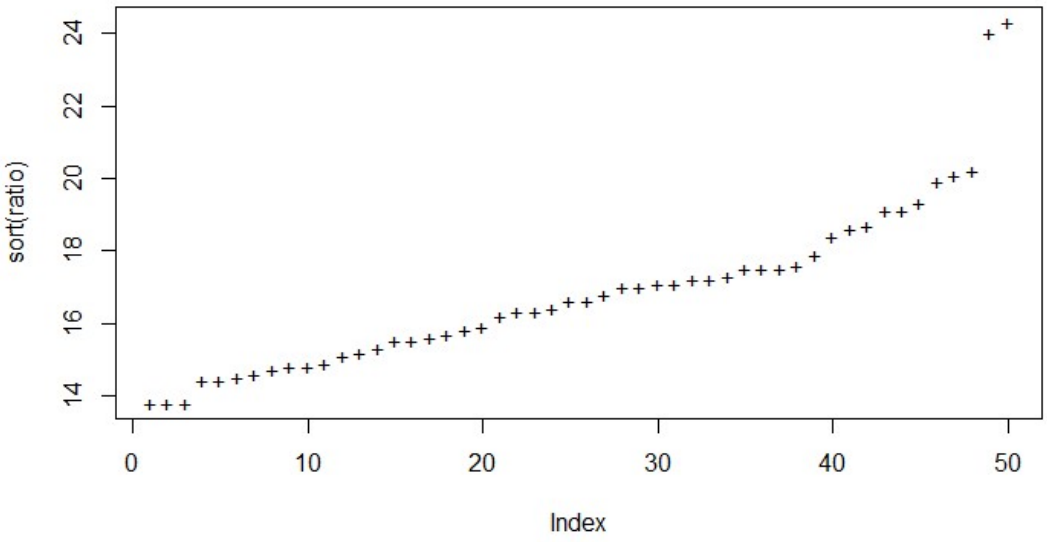
Assignment 1

105061110 周柏宇

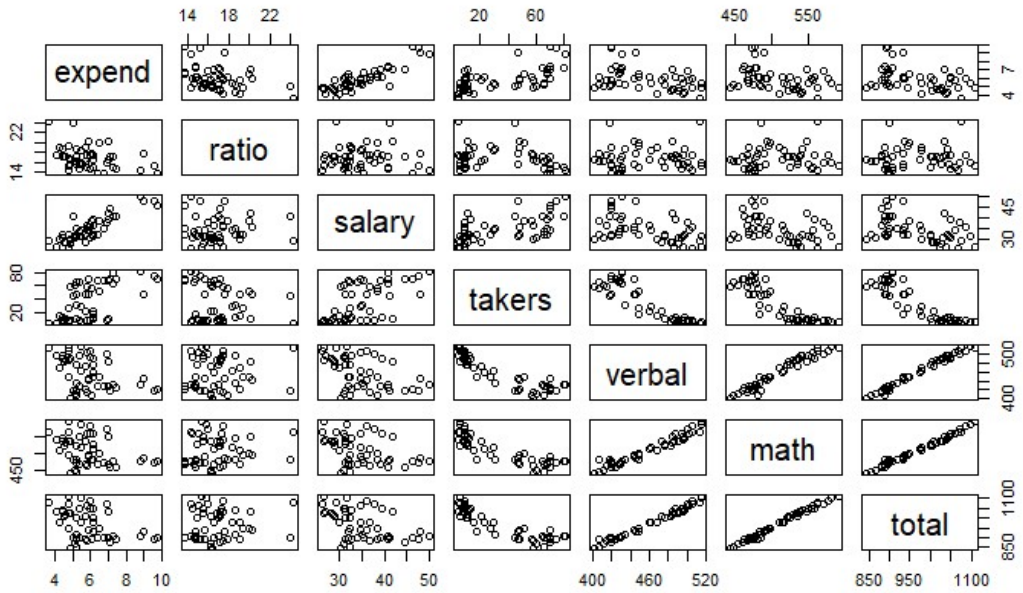
1.



The plot above is the sorted average expenditure on a student. There are four outliers which correspond to Connecticut (8.817), Alaska (8.963), New York (9.623) and New Jersey (9.774).



There are also some outliers for the pupil to teacher ratio, which is California (24.0) and Utah (24.3). Furthermore, the minimal value is 13.8 for three states, which is kind of suspicious because the ratio should be a continuous data, such chance to have 3 exactly the same value is very small.



With the figure above we can visually identify some correlation between some variables. Specifically, like verbal-math, verbal-total, total-math, expend-salary are clearly positively correlated. We can validate our observation with some numeric:

cor(verbal, math) = 0.970256

cor(total, math) = 0.9915033

cor(total, verbal) = 0.9935024

cor(expend, salary) = 0.8698015

We can see that verbal score and math score is strongly positively correlated, which may break our stereotype of a person can either be an art person or science person. Since verbal and math are positively correlated, the total score should still be positively correlated, unsurprisingly.

As for the expenditure and salary pair, understanding the meaning of the variable is enough to explain the correlation. Because the expenditure on a student includes the textbook cost and the tuition, therefore, the higher a teacher's salary is, the higher the expenditure on a student will be.

Also, between eligible takers percentage and total score (also for verbal and math scores) there seemed to be strongly negatively correlated:

cor(takers, total) = -0.8871187

cor(takers, verbal) = -0.893263

cor(takers, math) = -0.8693839

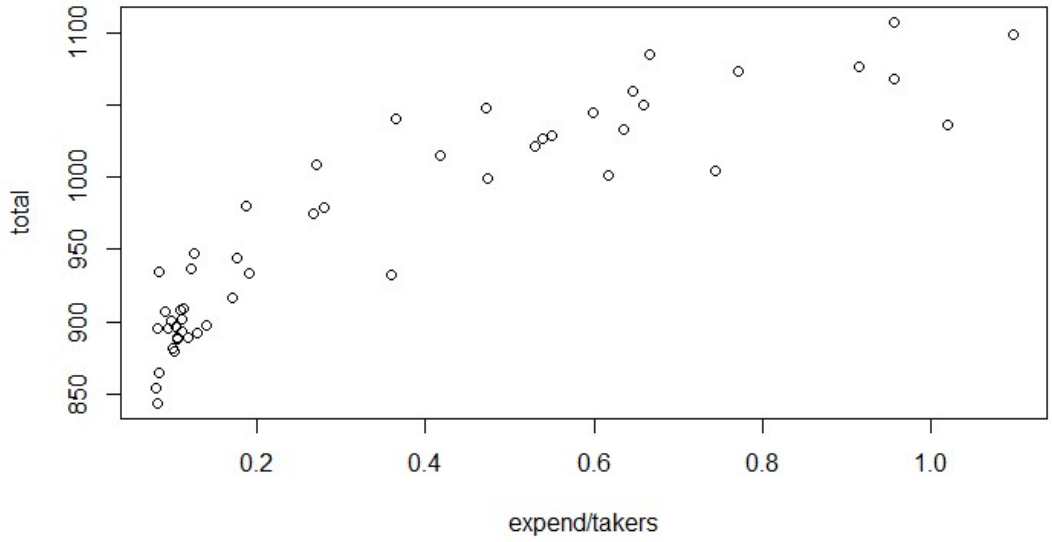
My explanation is that for the states with low takers percentage, the education mainly focus on those students with high potential, therefore they are more likely to score high on SAT; on the contrary, the states with high takers percentage, their education may be more universal: they also focus the students below average, thus less resource for already good students to become better.

Also, we expect the more we spend on a student, we better he or she can scores. At first glance, the expenditure versus total score plot seems uncorrelated, even slightly negative.

cor(expend, total) = -0.380537

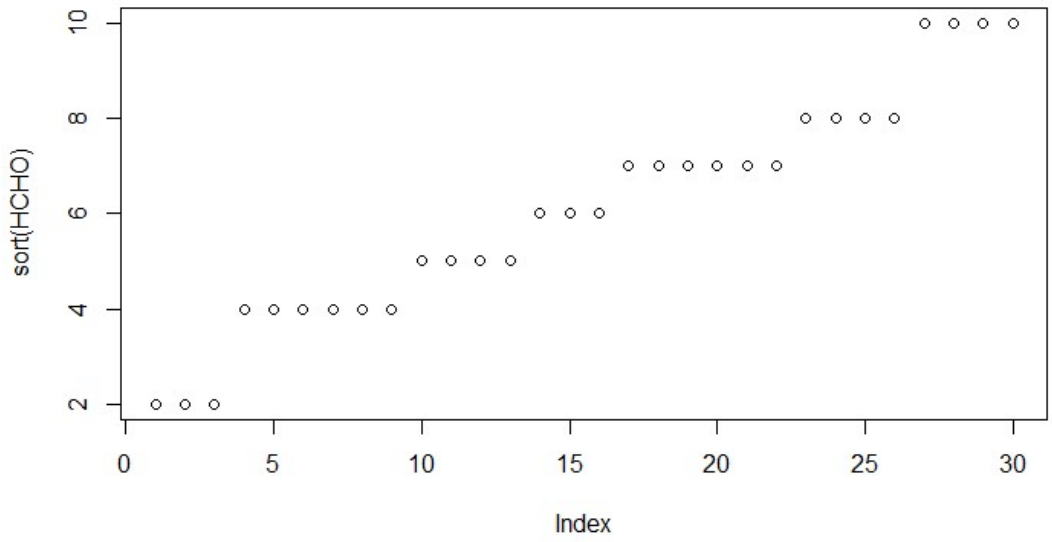
But if we normalize the expenditure with the eligible taker percentage, we get a result fits our expectation (and it almost looks like a logarithmic function).

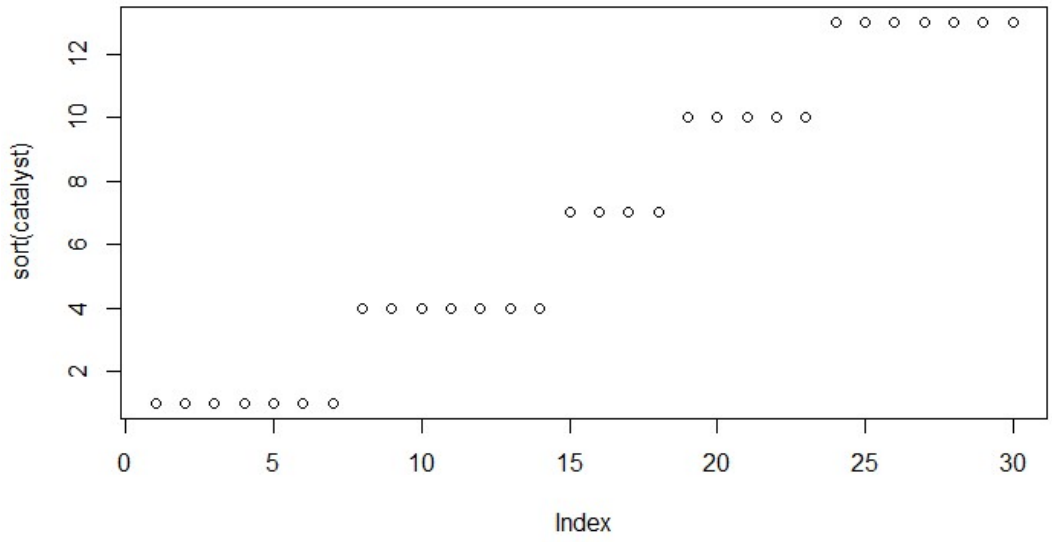
cor(expend/takers, total) = 0.9014098

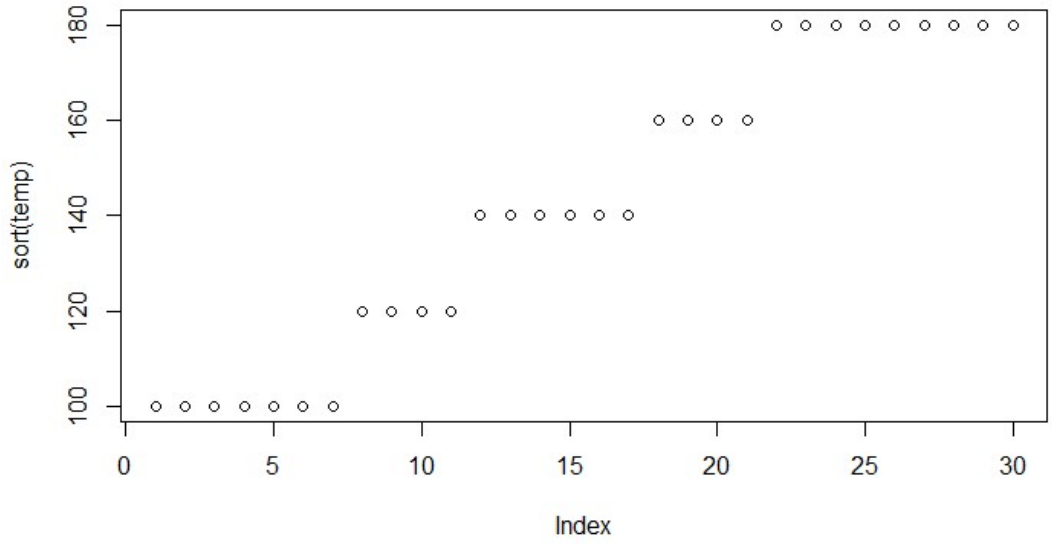

total
850
900
950
1000
1100
0.2
0.4
0.6
0.8
1.0
expend/takers

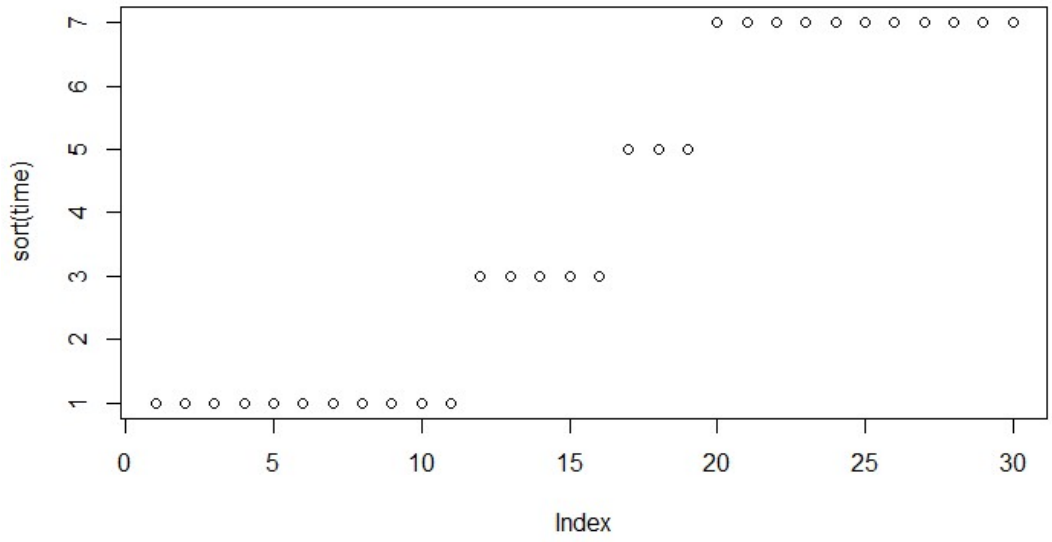
2.

i.


sort(HCHO)
2
4
6
8
10
0
5
10
15
20
25
30
Index


sort(catalyst)
2
4
6
8
10
12
0
5
10
15
20
25
30
Index

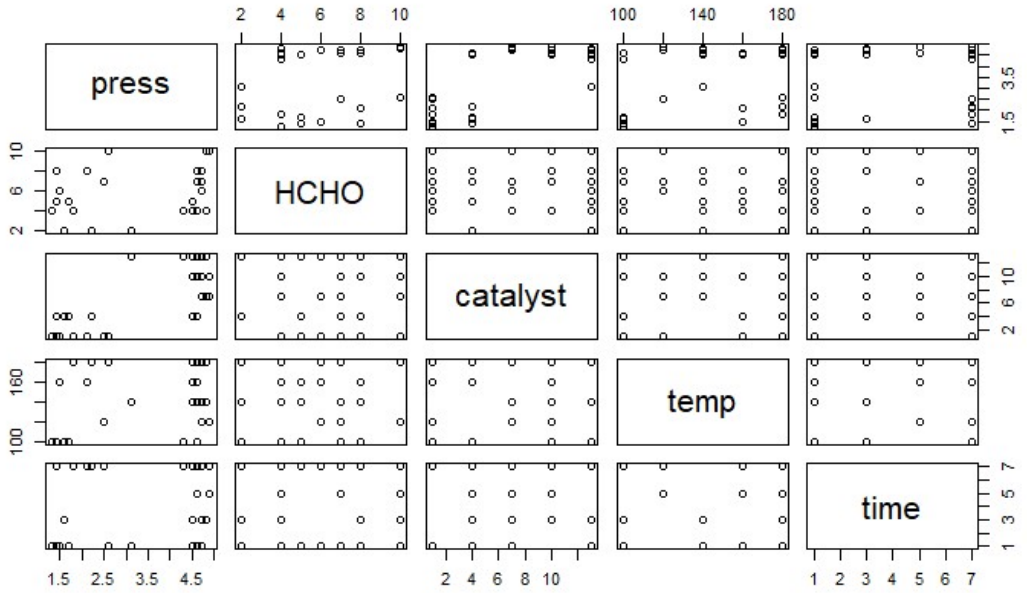

sort(temp)
100
120
140
160
180
0
5
10
15
20
25
30
Index



Although we often think of temperature, time and concentration as continuous values. In this case, they are confined to finite choices and have an almost uniform interval between one another.

ii.

Judging by the meaning of the variables, I posit that the “press” should be the y and all the other variables are the potential factors that affect “press”. Once we observe the figures above, we can tell that all the variables we consider them factors are all discrete. Especially for the “temp”, which represents the curing temperature and generally will not be the same, not to mention that they have a perfect interval of 20 between different temperatures. Moreover, we can observe the plot of catalyst vs temp, catalyst vs time and HCHO vs catalyst from the figure below. It looks kind of like a grid search, trying all the possible combination of factors and see how the result responses. For this reason, I believed the data are collected under controlled environment, therefore, they are experimental data.


press
HCHO
catalyst
temp
time
2 4 6 8 10
100 140 180
1.5 3.5
2 6 10
100 160
1 3 5 7
1.5 2.5 3.5 4.5
2 4 6 8 10
1 2 3 4 5 6 7