

GER1000 2018 Semester 1  
Quiz 3 and solutions

1. Prof Ng attempted to find the correlation between the number of hours NUS hall residents spent in hall activities and their exam scores. He recorded the average number of hours spent on hall activities and average score from each of the seven halls of residence. He then used it to calculate a correlation of -0.2.

In his report, Prof Ng wrote “I conclude that there is a weak negative correlation between the number of hours spent on hall activities and exam score” among NUS hall residents.

Which of the following may be present in this conclusion?

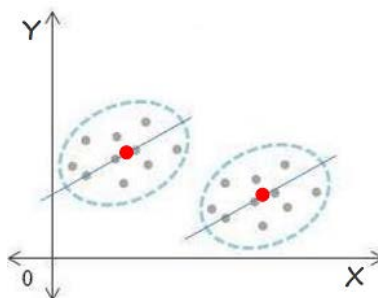
- (a) Atomistic fallacy
- (b) Ecological fallacy

Answer: (b). The correlation is calculated based on average values, while Prof Ng was interested in the correlation at the level of individual hall resident. The latter correlation may not be equal to -0.2.

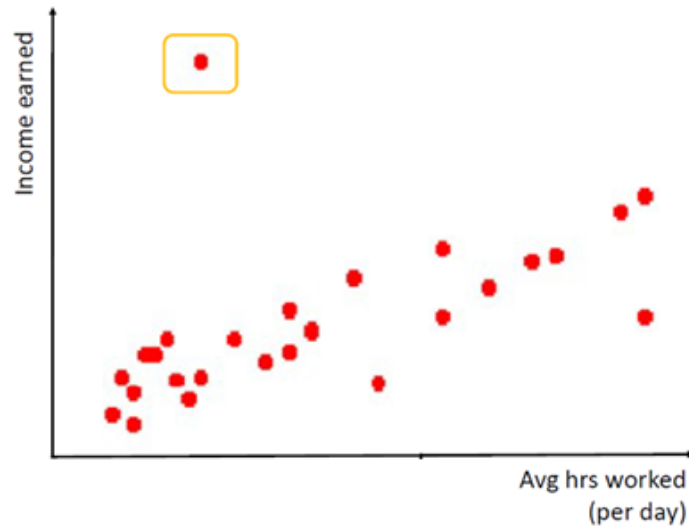
2. Prof Tan is interested in finding the correlation between midterm and final scores of his students in the second semester of 2017/18. In that semester, he taught the same module to students in two classes,  $U_1$  and  $U_2$ . The exam questions for classes  $U_1$  and  $U_2$  were the same. It was observed that the correlation coefficient for class  $U_1$  was 0.8 and the correlation coefficient for class  $U_2$  was 0.8 too. Which of the following is true for the correlation coefficient of all the students, from both classes  $U_1$  and  $U_2$  put together?

- (a) It must be 0.8.
- (b) Considering the attenuation effect, it must be bigger than 0.8.
- (c) It cannot be determined based on the information provided.

Answer: (c). For example, in the diagram below, the correlation obtained if we combine the 2 sets of data is very different from the correlation coefficient obtained for each set of data alone. In this case, in fact, it becomes negative when all the data points are combined.



3. Suji was interested in the relationship between average hours worked and income earned among a few selected graduates. The scatter plot below shows the data she collected. Each data point represents one selected graduate.

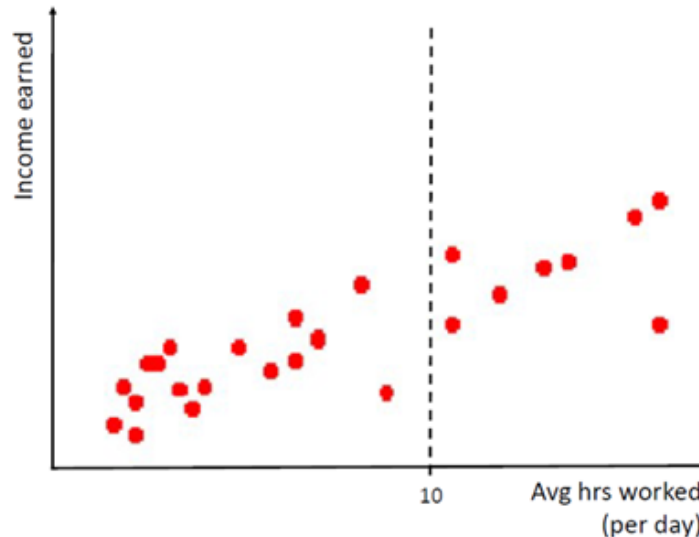


At first, Suji calculated the correlation coefficient ( $r_1$ ) from all the data she collected. However, on taking a closer look at the scatter plot, she calculated another correlation coefficient ( $r_2$ ) after removing the data point in the yellow box. Choose the most appropriate option that compares the correlation coefficients  $r_1$  and  $r_2$ ; and the reason involved in calculating  $r_2$ .

- (a)  $r_1 > r_2$ , after removing an outlier
- (b)  $r_1 > r_2$ , attenuation effect
- (c)  $r_1 < r_2$ , after removing an outlier
- (d)  $r_1 < r_2$ , attenuation effect

Answer: (c). From the two examples illustrated in the lecture notes (Chapter 2 - slides 60,61), the presence of outliers seemed to decrease the correlation. However, note that there are situations where the outliers will, actually, increase the correlation.

4. Suji was interested in the relationship between average hours worked and income earned among a few selected graduates. The scatter plot below shows the data she collected. Each data point represents one selected graduate.

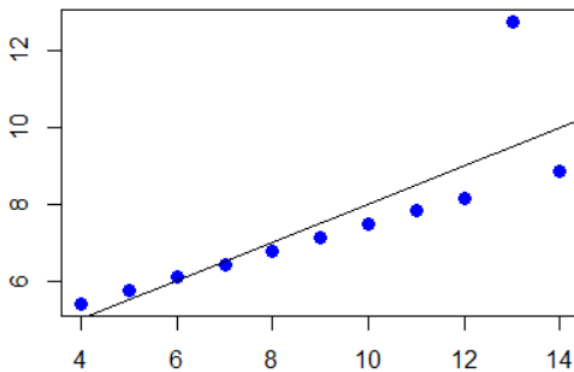


Suji calculated the correlation coefficient ( $R_1$ ) for all the data she collected above. She also wanted to check the relationship between hours worked and the income earned only among those who over-work. (Here she defined 'overworking' as working equal to or more than 10 hours per day on average.) Hence she calculated the correlation coefficient ( $R_2$ ) for those who 'overwork' from the above data set. Choose the most appropriate option that compares the correlation coefficients  $R_1$  and  $R_2$ ; and the concept involved in the comparison.

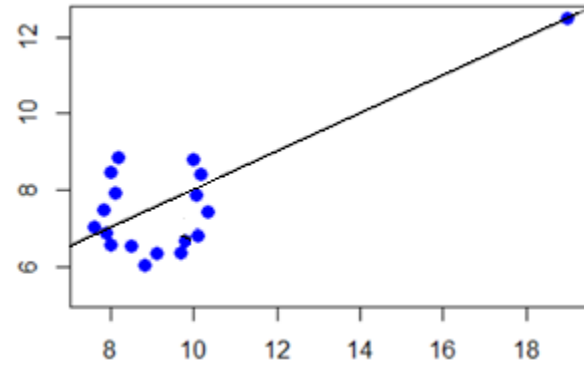
- (a)  $R_1 > R_2$ , atomistic fallacy
- (b)  $R_1 > R_2$ , attenuation effect
- (c)  $R_1 < R_2$ , ecological fallacy
- (d)  $R_1 < R_2$ , attenuation effect
- (e)  $R_1 < R_2$ , atomistic fallacy

Answer: (b). As discussed in the lecture notes (Chapter 2 - slide 79), attenuation effect occurs when we restrict the range of a variable.

5. Please look at the two scatter plots below.



(i)



(ii)

What will the new correlation coefficients be in the cases of (i) and (ii) respectively, after removing the outliers?

- (a) close to 1, close to 1
- (b) close to 0, close to 0
- (c) close to 1, close to 0
- (d) close to 0, close to 1
- (e) No change in the correlation coefficients.

Answer: (c). For case (i), the remaining points lie almost on a straight line, hence we have an almost perfect correlation. For case (ii), the linear relationship between the remaining points is very weak and is close to 0.