Determination of the covariance and correlation of quizzes, exam scores, and absences

Kenneth Domingo, Mary Chris Go, and Marc Arvie Talavera

National Institute of Physics, University of the Philippines, Diliman, Quezon City
*Corresponding author: mvtalavera1@up.edu.ph

Abstract

Covariance and correlation help in treating and interpreting the data. Between two sets, knowing these two will let us know their relationship. Using the right calculations and Python, values were obtained. This experiment was done for 61 measurements. For all pairs of variables, a linear correlation was observed and showed either highly significant or significant. Plots show a direct relationship except for the plots of the long exams 1,2,3, long quizzes 1,2,3, and final exam versus the number of student absences.

Keywords: covariance, correlation

1 Introduction

Obtained values are not helpful when you do not treat them right and interpret them. One popular way of interpreting values is obtaining the covariance and correlation. This helps to see the relationship between the values.

Covariance and correlation shows information about the measurement of strength between the sets of random variates [1]. We see how they vary together. Given two values X and Y, each had a sample size of N. Both of them show whether variables are directly or indirectly proportional. The relationship between variables can be illustrated through a graph. Covariance shows how these variables are related. A positive covariance shows a direct relationship while a negative shows an inverse relationship. To check the covariance, we use this formula [2]:

$$\sigma_{xy}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{1}$$

where N is the number of data points, x_i is the independent variable, y_i is the dependent variable, while $\bar{x}$ and $\bar{y}$ are the mean of the variables respectively.

On the other hand, correlation is another way to show the relationship of two sets. As covariance tells if the sets are positively or negatively related, correlation tells the degree to how these sets move together. Correlation standardizes the measure of the dependence between these sets and shows how closely the two sets move. We introduce the idea of correlation coefficient, a value between 1 and -1. From these coefficient we are able to interpret these variables.

If the coefficient is $+1$, a direct relationship can be observed. A positive correlation that is less than one shows a less than perfect correlation, where the strength of the correlation grows as the number approaches one. Furthermore, if the coefficient is zero, we cannot claim a relationship between these variables. If one variable moves, we cannot

detect how would the other variable move too. Lastly, if the coefficient is -1, what happens here is just the opposite of the result for $+1$ [2].

In this experiment, we were given a data set which the goal is to be able to find the correlation between the following: long quizzes and long exams, average long quizzes and average long exams, average long quizzes and final exam scores, average of long exams and final exam scores, student absences and long quiz scores, student absences and long exam scores, and student absences and final exam scores.

The probability asserting the hypothesis that rejects a linear correlation between two variables can be calculated using the following equation

$$\text{Prob}_N\left(|r| \geq |r_0|\right) = \frac{2\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N-2}{2}\right)} \int_{|r_0|}^{1} (1-r^2)^{(N-4)/2} \mathrm{d}r \tag{2}$$

where N is the number of measurements and r_0 is the correlation coefficient [3]. Performing the integration in the equation above using the *Wolfram Mathematica 11.0* software, equation (2) becomes

$$\text{Prob}_N\left(|r| \geq |r_0|\right) = 1 - \frac{2\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{N-2}{2}\right)} |r_0| \, {}_2F_1\left(\frac{1}{2}, 2 - \frac{N}{2}; \frac{3}{2}; r_0^2\right) \tag{3}$$

where $|r_0| < 1$ and $\text{Re}[N] > 2$.

This paper is organized as follows: in Section 2 we describe and detail the methods used in performing the experiment. In Section 3, we analyze and discuss the results. We conclude and summarize the paper and give recommendations to improve similar experiments in Section 4.

2 Methodology

All the given data was processed and interpreted in Python. The desired variables were plotted against each other, and the linear best-fit lines, as well as the linear correlation coefficients, were obtained using the `stats` library from the `scipy` package.

3 Results and Discussion

Using (3) and $N = 61$ measurements, the probability that the hypothesis that does not support a linear correlation between the variables in question will be determined. A linear correlation will be asserted if the probability computed is less than 1%, or 5%. If it is less than 5%, then the correlation is significant; if it is less than 1%, then the correlation is highly significant [3].

3.1 Correlation between long exams and long quizzes

From the plot of the long exam 1 versus long quiz 1, the r^2 value is given by 0.286, having a probability equal to 9.0×10^{-4}%, which is less than 1% implying that a linear correlation between long exam 1 and long quiz 1 is highly significant.

For the long exam 2 versus long quiz 2 plot, the r^2 value is given by 0.253, resulting to a probability equal to 3.6×10^{-3}%, which is less than 1%, implying that a linear correlation between the two variables in question is highly significant.

For the long exam 3 versus long quiz 3 plot, the r^2 value is given by 0.474, resulting the probability equal to 8.7×10^{-8}%. This probability is less than 1%, which implies a linear correlation between long exam 3 and long quiz 3 is highly significant.

For this part, the averages for the three long exams and three long quizzes were calculated, and the resulting data were plotted, giving the r^2 value equal to 0.684. This gives a probability equal to $6.8 \times 10^{-13}\%$, which is less than 1%; therefore, a linear correlation between the averages of the long exams and the long quizzes is highly significant. The plot of the averages for the three long exams versus the three long quizzes shows a direct relationship between the two.

From the plots of long exams 1, 2, and 3 versus long quizzes 1, 2, and 3, respectively, all these plots show a direct relationship between the corresponding variables in question.

3.2 Correlation between long exams versus student absences and between long quizzes and student absences

For long exams 1, 2, and 3 versus number of student absences plots, the r^2 values are equal to 0.190, 0.189, and 0.170, respectively, resulting to probabilities equal to 0.04%, 0.05%, and 0.1%, respectively. These values are less than 1 %, which makes the correlation between the corresponding variables in question highly significant.

From the plots of long exams 1, 2, and 3 versus the number of student absences, as the number of absences increases, the long exam scores decrease.

For long quizzes 1, 2, and 3 versus number of student absences plots, the r^2 values are given by 0.260, 0.310, 0.284, respectively, resulting to probabilities equal to $2.7 \times 10^{-3}\%$, $3.2 \times 10^{-4}\%$, and $9.8 \times 10^{-4}\%$, respectively. These values are less than 1 %, which makes the correlation between the corresponding variables under study highly significant.

From the plots of long quizzes 1, 2, and 3 versus the number of student absences, as the number of student absences increase, the long quiz scores decrease.

3.3 Correlation between final exam and the average of long exams, final exam and the average of long quizzes, and final exam and the number of student absences

From the plots of final exam versus the average of long exams, final exam versus the average of long quizzes, and final exam versus the number of student absences, the r^2 values are given by 0.572, 0.637, 0.244, respectively, resulting to probabilities equal to $1.8 \times 10^{-10}\%$, $1.6 \times 10^{-12}\%$ and $5.2 \times 10^{-3}\%$, respectively. These values are less than 1%, which leads to a highly significant correlation between the variables in0 question.

The plots of final exam versus the average of three long exams and of final exam versus the average of long quizzes show a direct relationship between the two corresponding variables. The plot of final exam versus the number student absences shows that as the number of student absences increases, final exam scores decrease.

4 Conclusions

All the pairs of variables that were compared and studied in this experiment showed a linear correlation. The correlations are either highly significant or significant, which depends on the value of the caluclated probability using equation (3). The plots show a direct relationship, except for the plots of the long exams 1, 2, and 3, long quizzes 1, 2, and 3, and final exam versus the number of student absences.

References

[1] Covariance. (n.d.). Retrieved April 8, 2019, from http://mathworld.wolfram.com/Covariance.html

[2] Statistical Sampling and Regression: Convariance and Correlation. (n.d.). Retrieved April 8, 2019, from *http://ci.columbia.edu/ci/premba$_t$est/c*0331/*s*7/*s*7$_5$*.html*

[3] Taylor, J. R. An Introduction to Error Analysis: The Study of Uncertainties in Physical measurements 2nd ed. University Science Books, 1982.

Appendix
Figures and Diagrams

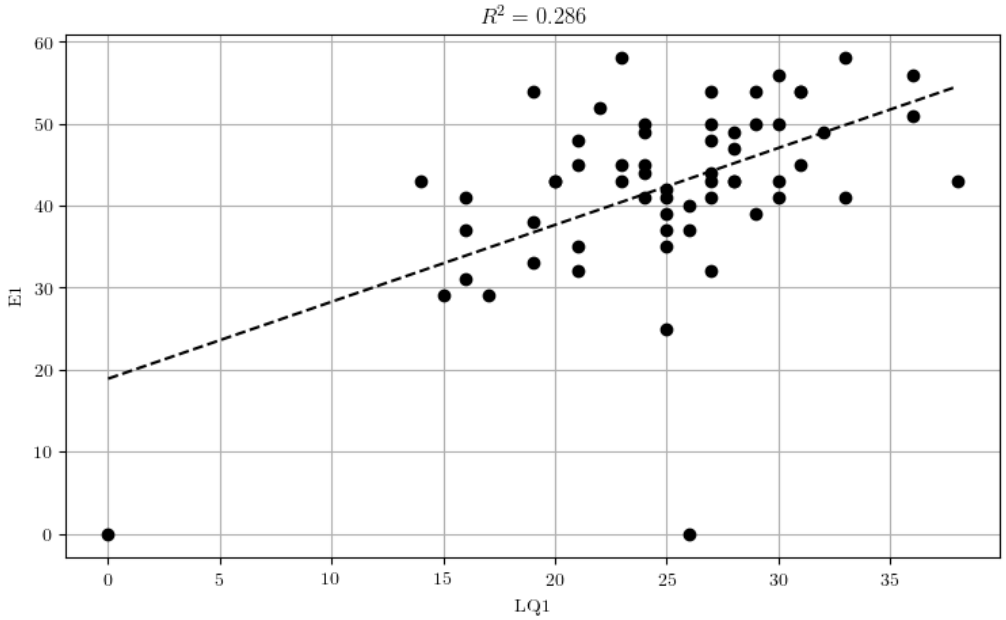

Figure 1

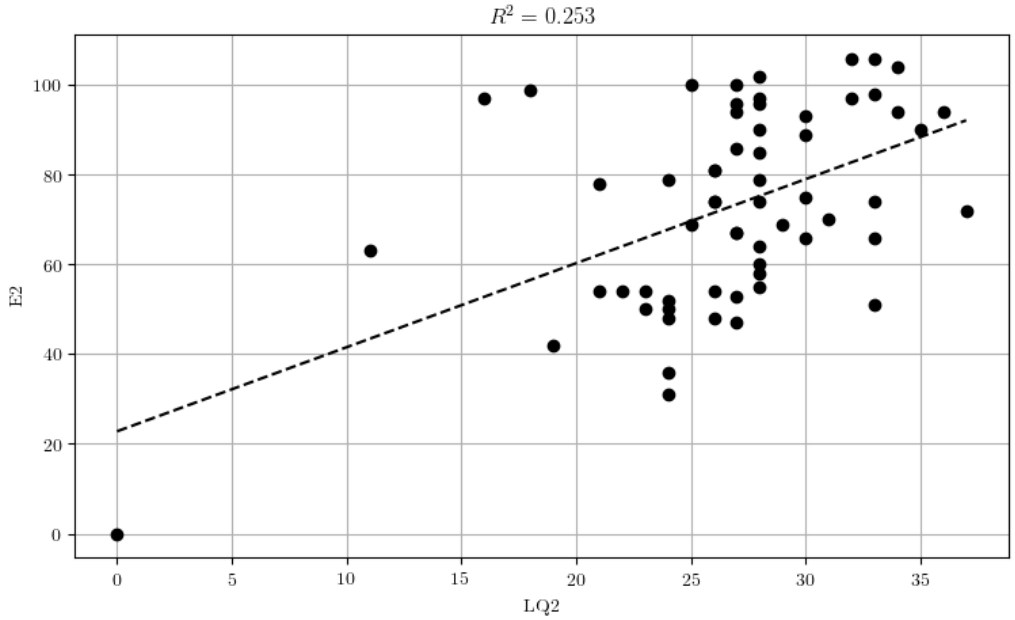

Figure 2

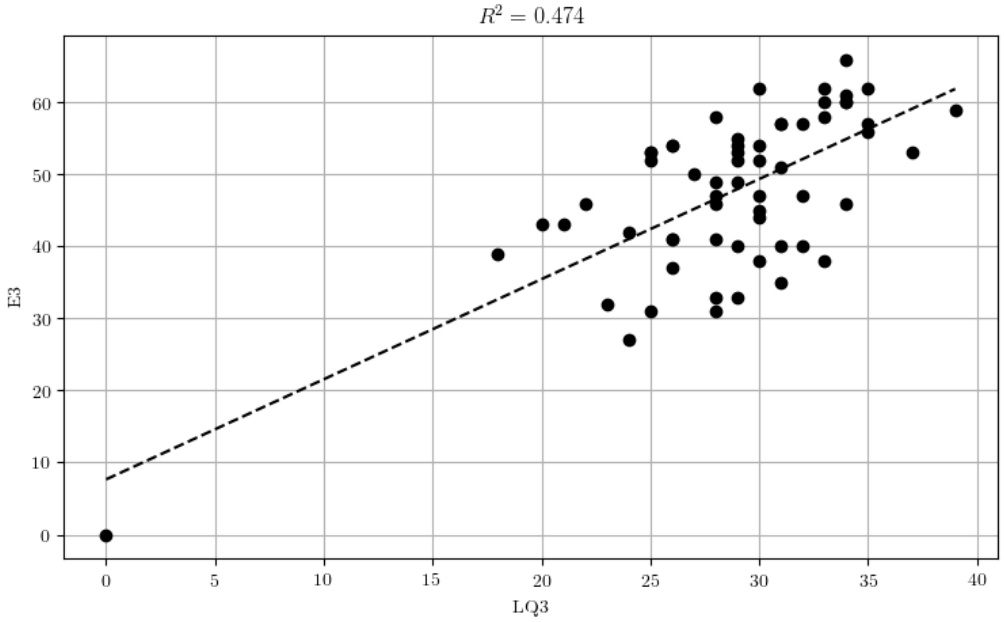

Figure 3

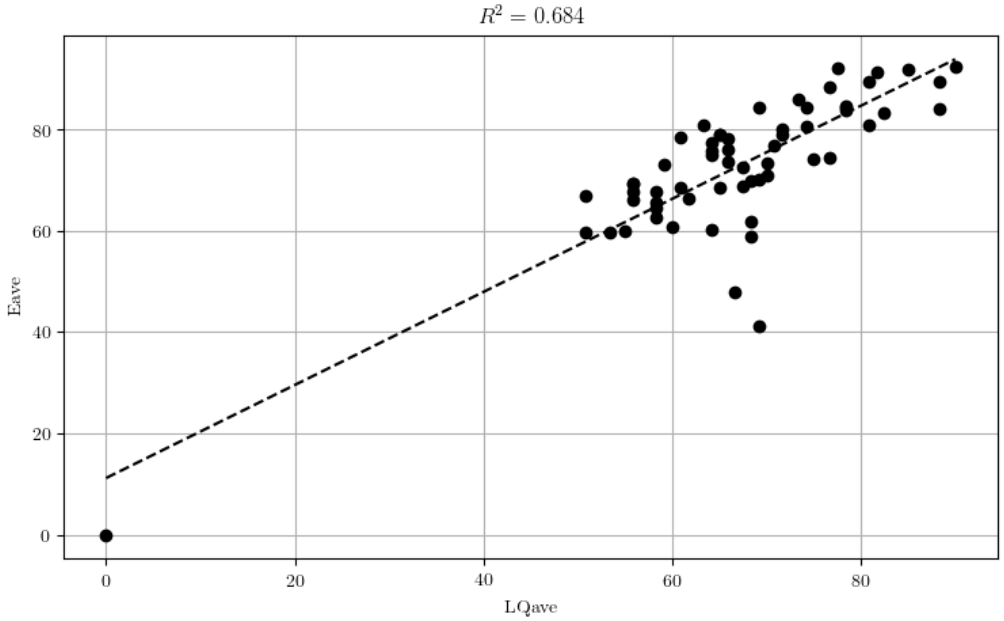

Figure 4

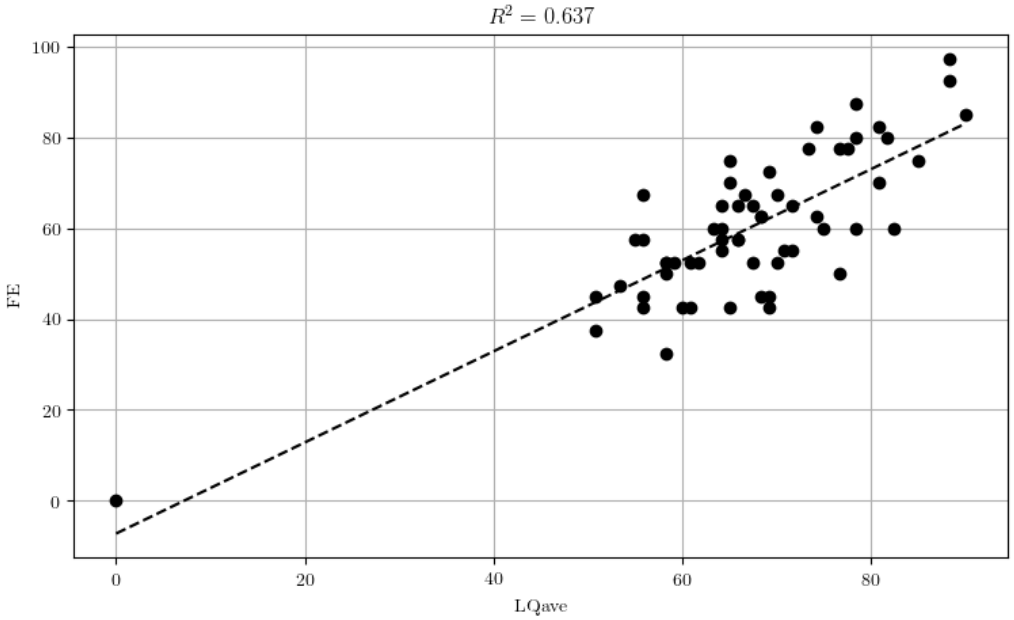

Figure 5

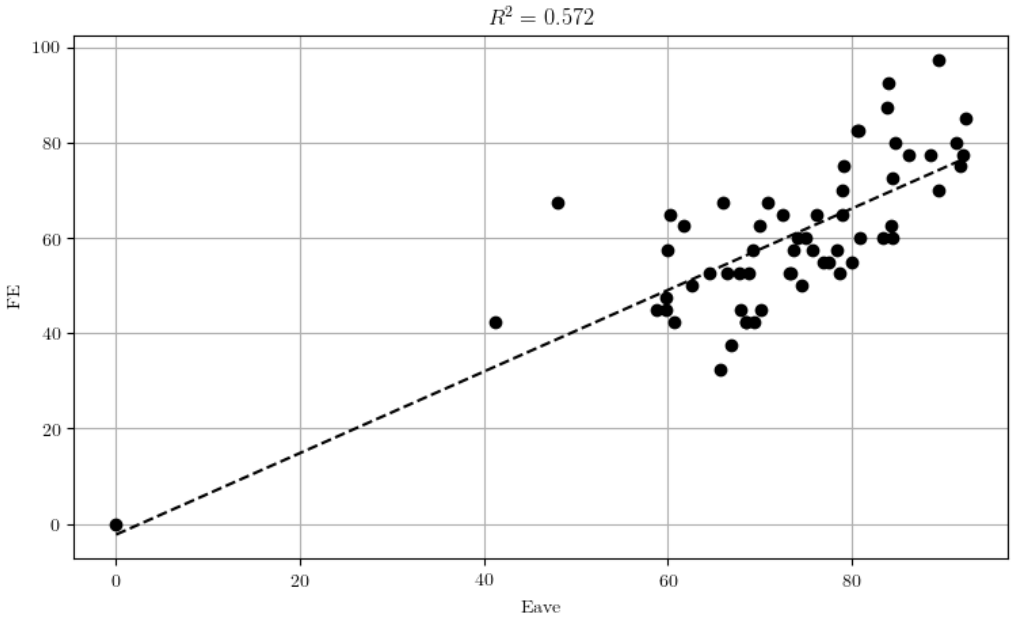

Figure 6

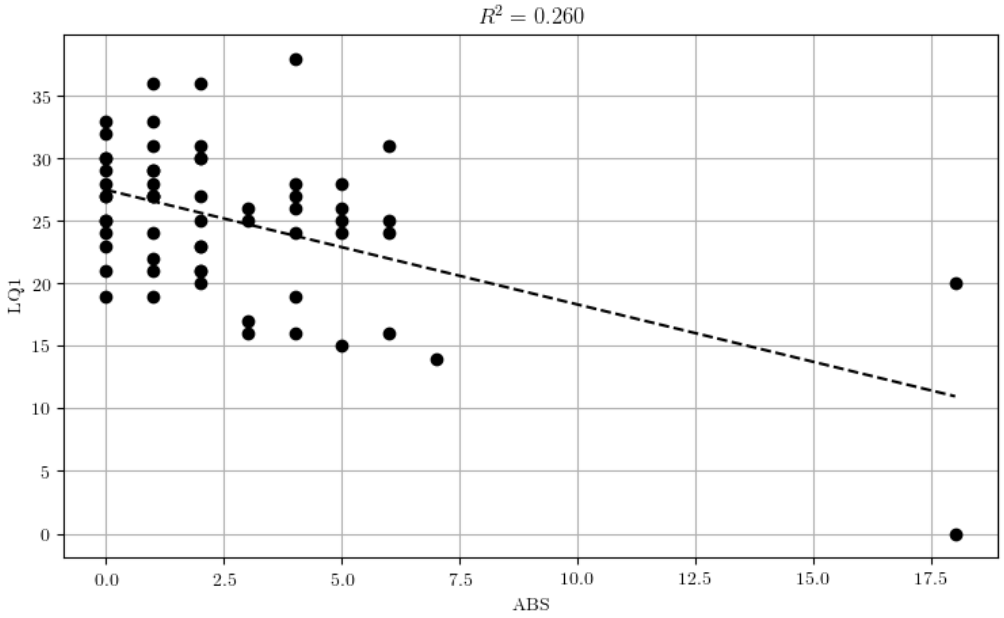

Figure 7

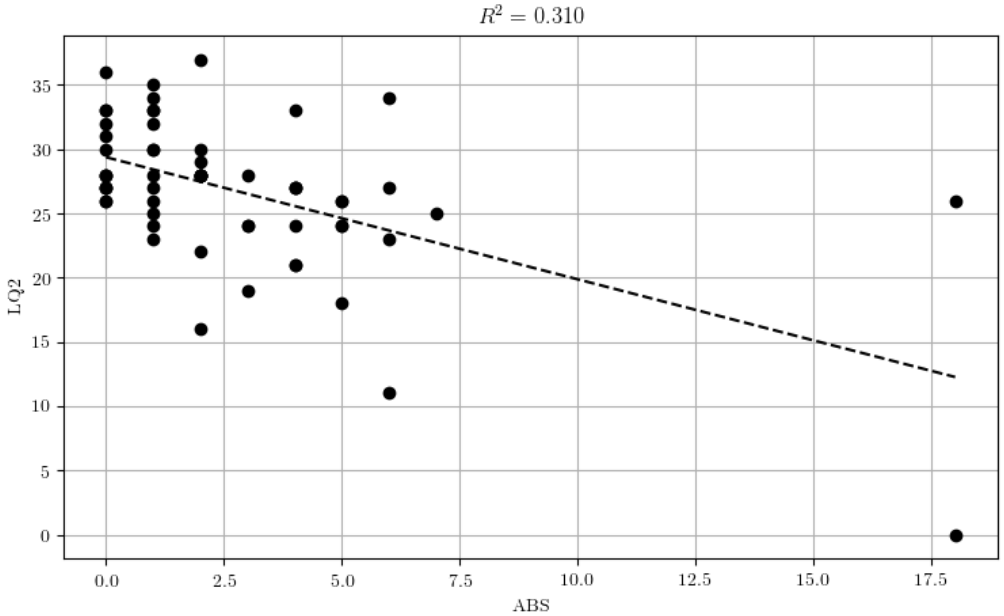

Figure 8

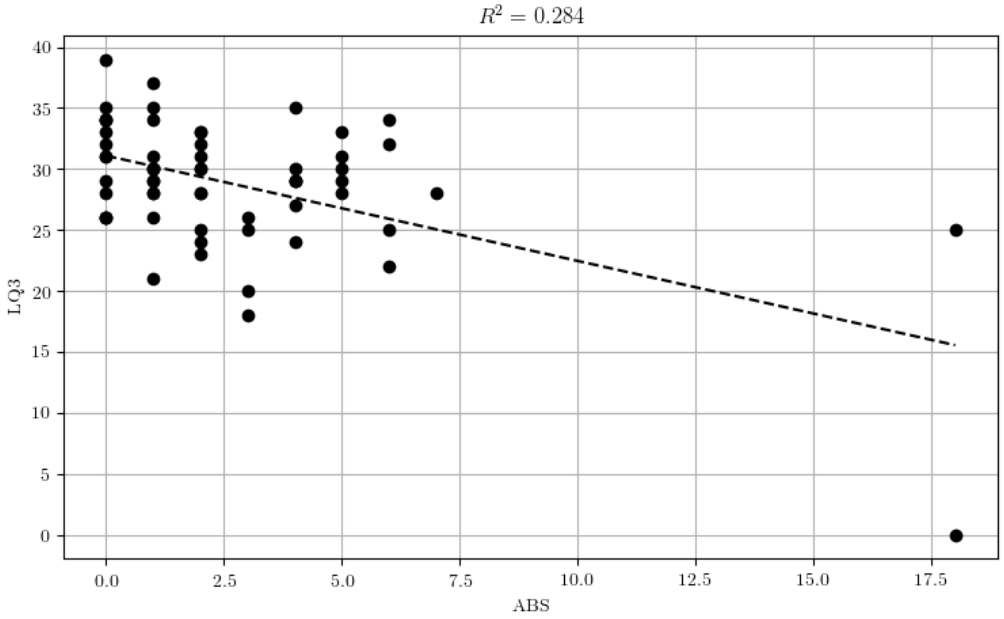

Figure 9

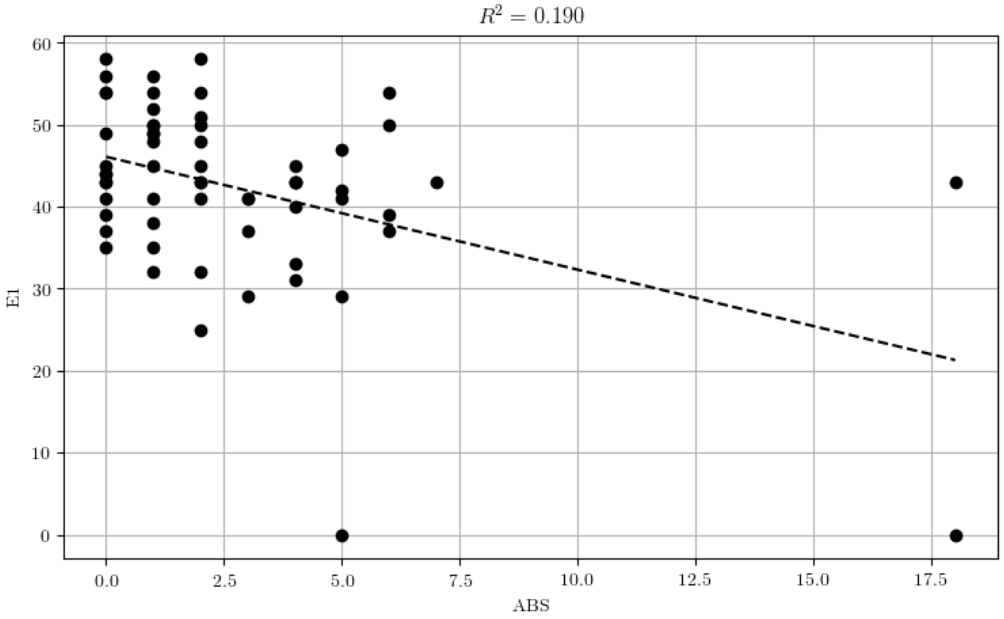

Figure 10

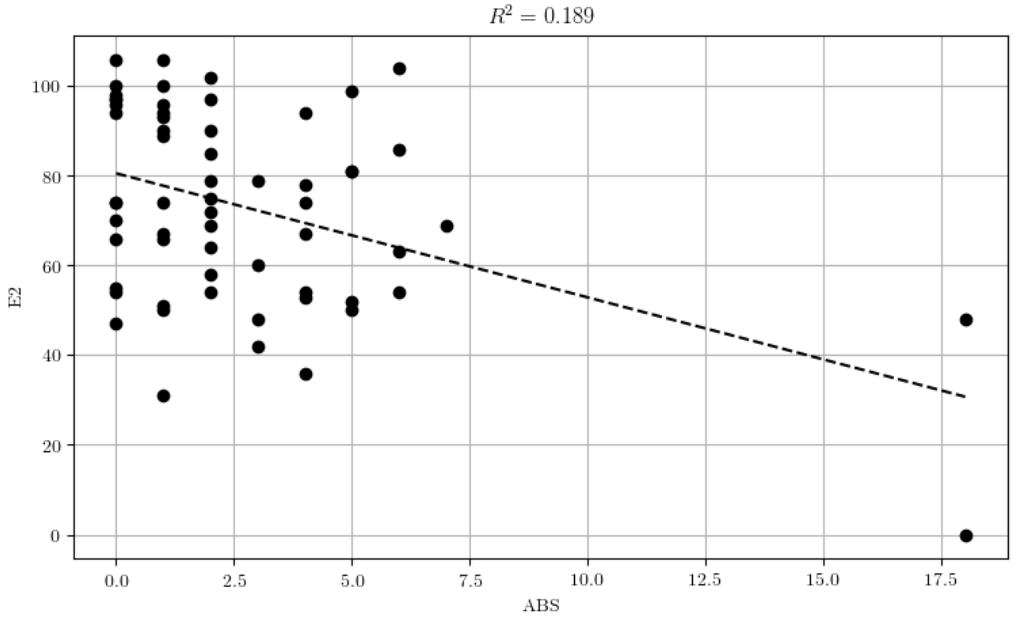

Figure 11

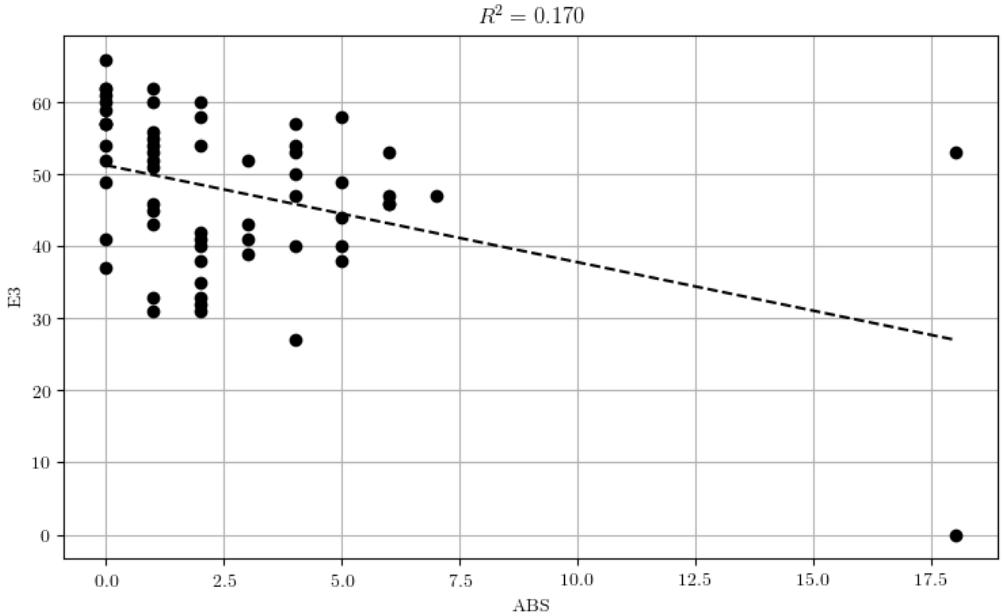

Figure 12

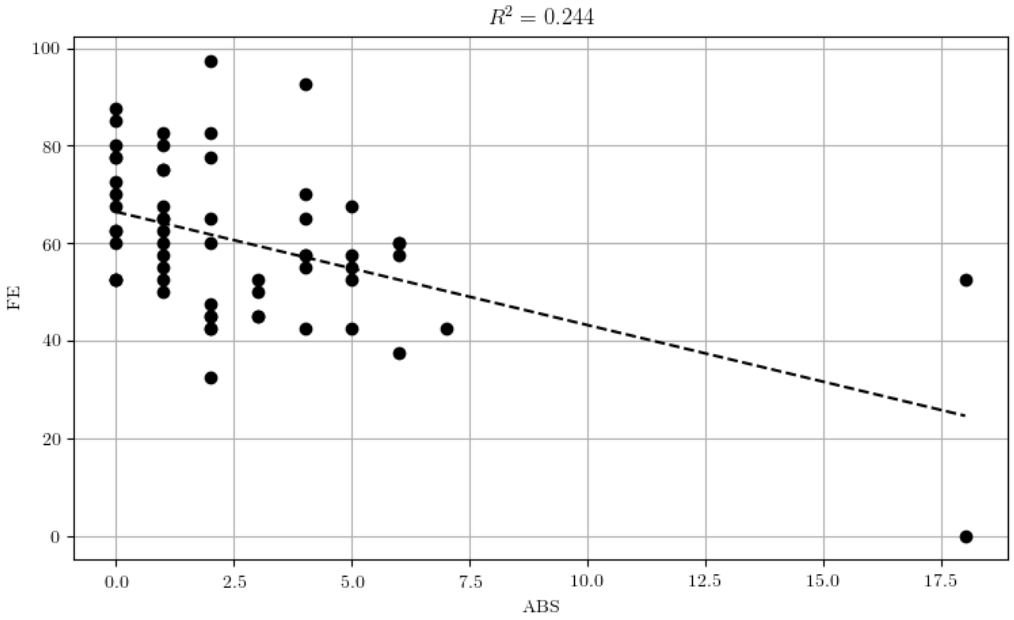

Figure 13

Code for calculating the probabilty of the hypothesis that rejects a correlation between two variables

```
from scipy.special import hyp2f1, gamma
import numpy as np

def F(r,N):
    a = (2*gamma((N-1)/2))/(np.sqrt(np.pi)*gamma((N-2)/2))
    b = hyp2f1(1/2, 2-(N/2),3/2,r)
    return (1-np.sqrt(r)*a*b)*100

r1 = float(input("r1 = "))
print("Probability1 = ", F(r1,61))
```