EE 381 Project 5

Due Date: 4-14-21

Instructions: The faculty member will facilitate completing this project.

Topic: Investigating the possibility of a linear relationship between two random variables.

Given two random variables (R.V.) that are related to the same subjects it can be of interests to determine whether or not there is a possible linear relationship. The linear relationship can be formally expressed as $Y = aX + b$ where a and b are real numbers and X and Y are the R.V. The information we obtain about the R.V. will be empirical. Further, we are working with R.V. thus we would be surprised if there was a perfect fit to of our data and linear relationship. You can make an initial conjecture about which of the two variables is the independent variable. (If in the end it turns out you are wrong you can start over with the variables swapped.) The first approach to studying a possible relationship between the variables is to obtain a visual representation.

Scatter Plot

Use a horizontal axis for the variable you decided is the independent variable. This is conventionally labeled x. The vertical axis will be the dependent variable and by convention will be labeled y. To make the scatter plot you will need pairs of samples (x, y) from the subjects under study. You then plot these ordered pairs on the scatter plot. If there is a general trend upward this may indicate what is termed *positive correlation*. If there is a general trend downward this may indicate what is termed *negative correlation*. If no trend is indicated there may not be a correlation of the linear type.

As always there is a desire to quantify or give a figure of merit to our perception. We will use some of the concepts from probability we have developed already. Remember the covariance between two variables?

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y.$$

It can be used to obtain a numerical value that will convey to us the likelihood of a linear relationship between the variables.

Correlation Coefficient

Let X and Y be any two R.V. The correlation coefficient of X and Y is denoted ρ (or $\rho(X, Y)$) and is given by

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = Cov(X^*, Y^*)$$

where $X^* = \frac{X - \mu_X}{\sigma_X}$ and $Y^* = \frac{Y - \mu_Y}{\sigma_Y}$.

Further,

$|\rho| \leq 1$ and $|\rho| = 1$ if and only if $Y = aX + b$.

The correlation coefficient ρ is a parameter and in practice we will not know it. The statistic related to it is r. The statistic r can be determined by the formula below. As before $-1 \leq r \leq 1$.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

A value for r is determine by entering the sample values into the formula for r. The heuristic employed is that if r is between eight-tenths and one then a positive correlation may exist and if r is between negative one and negative eight-tenths then a negative correlation may exist. To make the interpretation more accurate a process of quantizing the decision process is employed.

Hypothesis Test

We will use the traditional method for the hypothesis test. Further, if the samples are small and the two random variable are normally or approximately normally distributed that we can use the t-distribution. The usual form of the statement of the hypothesis is (though it can be one sided):

$$H_0: \rho = 0 \qquad\qquad H_1: \rho \neq 0$$

The critical value (C.V.) is determined using the t-table and degrees of freedom equal to n-2 where n is the sample size (The number of data pairs). The usual level of significance is 5%. The derivation of the test value (T.V.) is tangential to our present interest; hence, the formula is given below.

$$r\sqrt{\frac{n-2}{1-r^2}}$$

If the T.V. enters either of the two rejection regions determined by the critical values then the decision is to reject the null hypothesis and a correlations between the two variables is taken as the stepping off point for further studies of these two variables.

Correlation

There are five characterization of correlation.

1.) Cause and effect
2.) A reverse cause and effect: the variables are the reverse of that first assumed

3.) A third or lurking variable is involved
4.) There is a spectrum of variables with a complexity of interrelationships
5.) The relationship though it appears to exist is coincidental

If the researcher wishes to pursue the line of argument that there is a linear relationship then a straight line model can be made.

Least Square Fit

The formulas for the constants in the regression line, $y' = a + bx$, are

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \qquad b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

The name for this line is *the regression line*. The regression line can, potentially, be used to determine future values of the dependent variable.

The preceding has been a brief and limited discussion of *correlation regression*.

Exercise

For the following information provided draw a scatter plot, compute r, at a 5% level of significance perform a hypothesis test, determine the type of correlation, obtain the regression line, and use it to predict several new values of the dependent variable.

Listed in the table below are the number of grams of carbohydrates and the number of kilocalories for a 100-gram sample of various raw foods.

carbohydrates	15.25	16.55	11.10	13.01	14.13	15.11
kilocalories	59	72	43	55	56	59

You may want to consider using EXCEL to address this exercise.