Correlation and Regression

* The purpose of these notes is to introduce the statistical techniques of correlational analysis and regression, and how to perform such analyses using Excel.
* We will cover the statistical technique of correlational analysis, which is a technique to determine the degree of association or relationship between two or more variables.
* We will cover basic linear regression, a useful technique for making predictions about one variable based on knowledge of another variable.

A quick example to illustrate correlational analysis:

* Imagine that you are a health professional studying exercise and dietary habits. You know that people differ quite dramatically in terms of how often they exercise and in terms of what they eat and how much they eat. Given this variability in exercise and eating behaviors, it could be important to identify if the two variables are related to one other. That is, do people who exercise more tend to also have more healthy eating habits? Furthermore, it would be useful to determine if there are other variables related to exercise and eating (such as age). Correlational analyses can be conducted to answer such questions by providing a quantifiable measure of degree of association between variables.
* When we use the term research design, we are referring to the issue of how the research was conducted.
* Correlational designs are often contrasted with experimental research designs.

**Correlational research designs**

* refer to empirical investigations in which the researcher records the values of two or more variables for the purpose of determining whether the variables related to one other.

**Experimental research**

* involves the manipulation or control of at least one variable under study.
* In correlational research designs, there is no such control. Rather, the researcher simply takes some measurements and then determines if there is a relationship between such measures. It is precisely because of the fact that there is no experiment manipulation that we can say that correlation does NOT indicate causation (or cause and effect).

**Correlation**

We have the following data from a salary analysis that shows how many years of education a person has after high school and their salary. This is a purely fictitious example.



* Education is our independent variable (it may impact salary) and salary is the dependent variable. To determine the correlation between education and salary we will find Pearsons R value. The value ranges from -1 to +1. The closer the value is to 1, the strong the positive relationship. The closer the value is to -1, the stronger the negative relationship.
* The syntax for the R value is =correl(array1, array2) as shown below:



* The result is an R value of 0.855, which indicates a relatively strong positive correlation, indicating that more education results in higher salary.

**Linear Regression**

Using the same data, we will perform a linear regression analysis. See the following steps to do this:

**Step 1:**

* Click on Data, then Data Analysis, then select Regression



**Step 2:**

* Input the salary values into the Y Range, and the education values into the X range as shown below.
* Ensure Labels is selected (since our first row is the label and not a number), and ensure Confidence Level is checked and at 95%.
* Also, so we get a graph, click Line Fit Plot, then click OK.



**Step 3:**

* The results are shown as below.
* Analyze the results



* Multiple R shows the correlation coefficient (strength of the relationship) which you already computed. In this case, a relatively strong positive relationship at 0.85.
* R Square shows the Coefficient of Determination, which is the amount of impact the independent variable has on the dependent variable when considering other causes. In this case, education accounts for around 73% of the impact on salary, while the remaining 27% could be from other factors.
* The Intercept Coefficient is 22199.86
* The Education Coefficient 2324.84.
* This means that in order to predict salary, we will use the following equation:

**Salary = 22199.86 + (2324.84 multiplied by # of years of education)**

* In other words, for a person with zero experience they can expect 22199.86 as salary and for every additional year of education their salary will increase by 2324.84
* For example, a person with 3 years of education could predict their salary to be: **22199.86+(2324.84\*3) = $29174.38**
* In review the Line Plot, we see that there is some sort of a straight line moving from left to right in relation to education and salary. Therefore, we can tell visually that there is a relationship between the two. If the dots were scattered across everywhere, we may question that relationship.
* We have identified that there is a relationship (there is always usually a relationship. It could be positive or negative. It could even be the 1% instead of 73% of the cause but there is still a relationship). Now, we need to determine if that relationship is significant enough to report. We want to be 95% or more confident that if we were to run this analysis on another set of data, that we would get a similar result.
* To determine significance, we look at the Significance F statistic.
* If the *significance f statistic* value was .05 or less, then we reject the null hypothesis.
* If the *significance* *f* value was .051 or higher, then we do not reject (in other words we accept it).
* In this case, our significant f statistic is 0.001. That is less than 0.05 so we REJECT the null hypothesis.
* Remember, our null hypothesis is that there is no significant relationship between number of years of education and salary. By rejecting this statement, we are accepting the alternative statement that there IS a significant relationship.