2020

Clear-Sighted Statistics: Module 18: Linear Correlation and Regression

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 18: Linear Correlation and Regression

> "Correlation is *not* causation but it sure is a hint."[1]
>
> -- Edward Tufte

> The term "regression" is not a particularly happy one from the etymological point of view, but it is so firmly embedded in statistical literature that we make no attempt to replace it by an expression which would more suitably express its essential properties.[2]
>
> -- George Udny Yule and Maurice G. Kendall

I. Introduction

In this module we turn to simple linear correlation and regression, which focuses on the relationship between two interval or ratio variables. Correlation and regression (Ordinary Least Squares Regression or OLS) are a collection of some of the most widely used techniques in inferential statistics. After completing this module, you will be able to:

- Distinguish an independent variable (X) from a dependent variable (Y).
- Use Microsoft Excel to create scatter diagrams or X-Y charts, which chart the relationship between the independent and dependent variables.
- Calculate by hand and with Excel the coefficient of correlation, r, and interpret the result.
- Describe positive and negative correlations.
- Use scatter diagrams to visualize the relationship between the independent and dependent variables as well as understand the least squares line.
- Calculate by hand and with Excel the coefficient of determination, r^2, and interpret the result.
- Run a significance test of a correlation to determine whether there is a correlation in the population, and determine the probability of a Type II error and statistical power using G*Power.

- Conduct simple linear regression analysis using a handheld calculator, Microsoft Excel's built-in functions, and Excel's Regression Analysis tool.
- Conduct significance tests and develop confidence intervals to determine whether the independent variable predicts the dependent variable.
- Distinguish correlation from causation.
- Discuss spurious correlations and confounding variables.

You should download several files that accompany this module:

- Module18_Examples.xlsx, which shows the data and analysis for the examples used in this module.
- Module18_Exercises.xlsx, which shows the data for the end-of-module problems that you can solve.
- Student-t_table.xlsx or Student-t_table.pdf, the critical values table for the student-t distribution. This file is available in Appendix 2: Statistical Tables.

II. Correlation and Regression: An Overview

Correlation and regression cover different aspects of the relationship between an independent variable and dependent variable. An independent variable predicts the dependent variable. It is, therefore, sometimes called the *predictor* or *explanatory* variable. It is usually symbolized by the letter X. The independent variable is also called the *regressor*, *stimulus*, and *exogenous variable*. The dependent variable responds to changes in the independent variable, which is why it is sometimes called the *response* variable. The dependent variable is also called the *criterion variable*, *predictand*, *regressand*, and *endogenous variable*. It is usually symbolized by the letter Y. We will focus on simple linear correlation and regression, which has only one independent variable and one dependent variable. We will speak of XY variables, which are determined by the value of each datum's X and Y value.

Correlation measures the strength of the association between the independent variable and dependent variable. Technically, correlation examines how variance in the dependent variable is associated with variance in the independent variable. For example, we could calculate the number of speeding tickets given on the New Jersey turnpike and the number of state troopers patrolling the highway. The independent variable—the predictor—would be the number of state troopers and the dependent, or response variable, would be the number of speeding tickets.

As suggested many decades ago in the George Udny Yule and Maurice Kendall quotation at the start of this module, the term regression is odd, especially for contemporary students of statistics. A better term would be predictive modelling because regression predicts the value of the dependent variable based on the independent variable. It models the relationship between the X and Y variables by using the regression equation and fitting these variables to the least square line. The goal of regression is to explain variation in the Y variable based on changes in the X variable. Simple linear regression is based on the *sum of the squares*, which is a the sum of the squared differences between the observed and predicted value of the dependent variable or $\widehat{\mathrm{Y}}$. The difference between the observed Y value and the predicted Y value are errors in the model, which are called *residuals*. The objective of the regression model is to draw a straight line that comes closest to the observed values. This line is called the *least squares line*, *regression line*, or *line of best fit*. This line is unique. All other lines between the XY variable have a higher sum of the squares. See Figure 1.

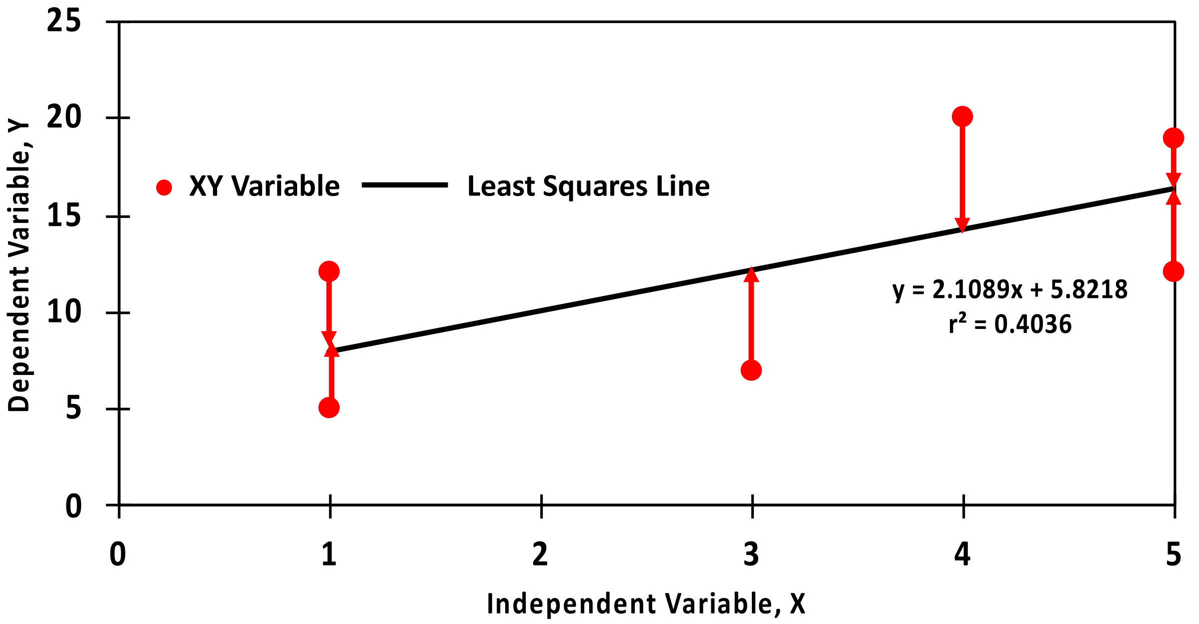


Figure 1: Scatter Diagram With the Least Squares Line

There are six requirements for linear correlation and regression:

1. The independent and dependent variables are quantitative and continuous.

2. The data to be correlated should *approximate* a normal distribution.

3. The relationship between the independent and dependent variables is linear.

4. If this relationship is not linear, the data in one or both of the variables may have to be transformed using a logarithmic scale. (Transforming data, however, is a sophisticated technique which we will not cover).

5. The variance around the regression line should be homoscedastic; which is to say, their variances should be roughly equal for all the Y variables predicted by X.

6. Outliers can distort the correlation: All data from the independent and dependent variables should be within plus or minus 3.29 standard deviations from its respective means.

No model explains the dependent variable completely. Regression models, therefore, also measure the random errors or residuals. This is a point that statisticians George E. P. Box and Norman R. Draper thought so important that they mentioned it twice in their book on statistical models:

- "Essentially, all models are wrong, but some are useful."[3]
- "(Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.)"[4]

It is less likely that the regression model will be useful when the correlation is weak, or when there are serious violations of the linear correlation and regression assumptions.

There is a more sophisticated type of linear regression called *multiple regression*. Multiple regression focuses on how one dependent variable responds to changes in two or more independent variables. Because multiple regression is typically not reviewed in introductory statistics classes, we will not cover it in *Clear-Sighted Statistics*.

III. The Basics

Francis Galton, the great Victorian era English statistician and eugenicist, first used the term correlation in the 1880s. Galton, a half-cousin of Charles Darwin, the founder of the theory of evolution, used statistics in his biometric studies of inherited traits. He also symbolized correlation with the letter r, a convention we still follow today. And, he pioneered the least square or regression line, which is still the basis of simple linear regression.[5]

In the 1890s, Galton's protégé and future biographer[6], Karl Pearson, developed the coefficient of correlation, which is also known as Pearson's product-moment coefficient of correlation, or PMCC. The coefficient of correlation, r for a sample or rho (ρ) for a population, measures the strength of the association between the independent and dependent variables. Scores range from -1.00 to +1.00. The more extreme the correlation coefficient, the stronger the correlation.

The correlation coefficient is a unit-free or “dimensionless” standardized measure of effect size, ES. Because r-scores are a unit-free, we can measure the strength of an association between two variables that are measured on different scales like:

- Income and incidence of smoking .
- Highest level of education and annual income.
- The relationship between demand and price in economics.

While the value of an r score can be positive or negative, when considering effect size for the purposes of estimating statistical power and the probability of a Type II error, we use the absolute value of r; which is to say, we drop the negative sign.

Table 1 shows for equations for calculating the coefficient of correlation for a population and sample.

Table 1: Equations for the Coefficient of Correlation

Population, ρ	**Sample, r**
$\rho = \dfrac{\Sigma(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}$	$r = \dfrac{\Sigma(X - \overline{X})(Y - \overline{Y}_Y)}{(n - 1)s_X s_Y}$

Where: r = sample correlation coefficient
ρ = population correlation coefficient
X = a random independent variable
Y = a random dependent variable
μ_X and μ_Y are the population means for the independent and dependent variables
σ_X and σ_Y are the population standard deviations for the independent and dependent variables
$\bar{X}$ and $\bar{Y}$ are the sample means for the independent and dependent variables
s_X and s_Y are the sample standard deviations for the independent and dependent variables
n = the number of paired variables

Table 2 shows how we interpret the coefficient of correlation. The closer the correlation is to zero, the weaker the correlation. The closer the r-value is to +1.00 or -1.00, the stronger the correlation.

Table 2: Interpretation of the Coefficient of Correlation

Negative	Positive	Meaning
0.00 to -0.099	0.00 to -.099	No Correlation
-0.10 to -0.299	0.10 to 0.299	Small Correlation
-0.30 to -0.499	0.30 to 0.499	Medium Correlation
-0.50 to -1.000	0.50 to 1.000	Large Correlation

The relationship between the independent and dependent variables can be illustrated using a scatter diagram, which is also called a scattergram or X-Y chart. Figure 2 shows a typical scatter diagram. Each dot represents the value of each XY variable based on the value of the independent variable (X) and dependent variable (Y).

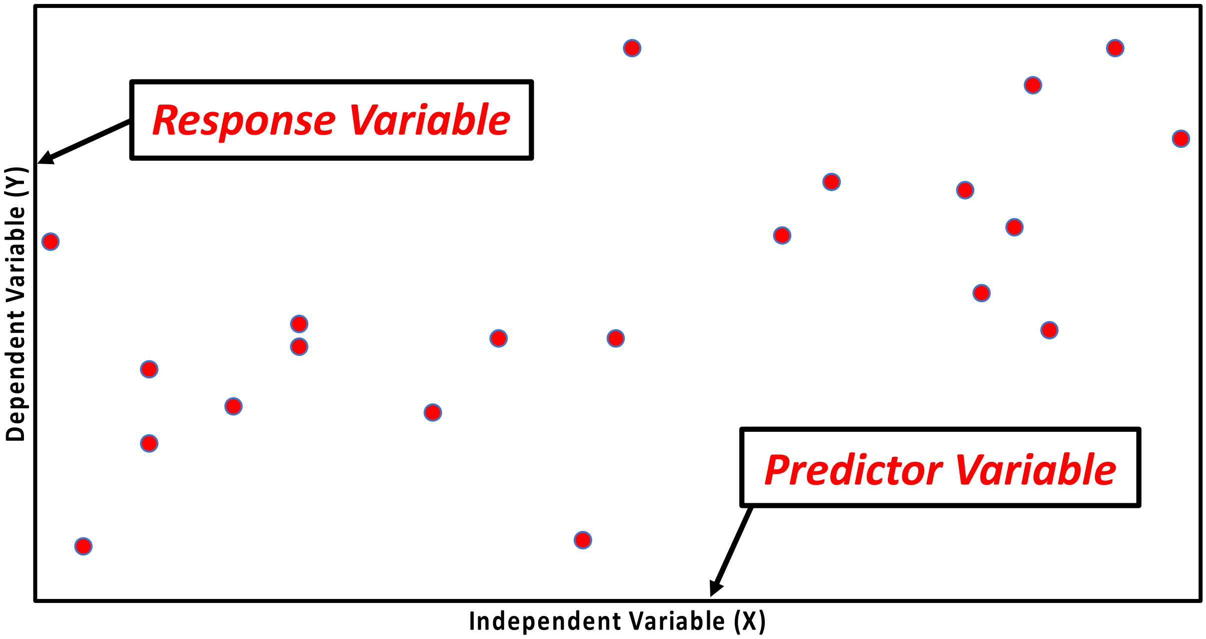
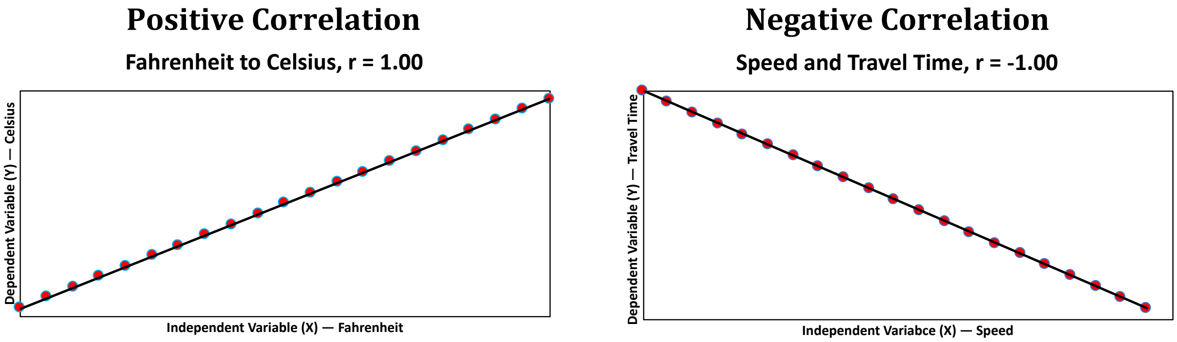


Figure 2: Scatter Diagram

Table 3 shows examples of a perfect positive and negative correlation. A perfect correlation has an r score of 1.00 or -1.00, which means that the independent variable predicts the changes in the dependent variable without and errors. An example of a perfect positive correlation is the mathematical relationship between temperature measured on the Fahrenheit and Celsius scales. An increase in temperature measured on the Fahrenheit scale is perfectly associated with increases on the Celsius scale. The r score is +1.00. Similarly there is a perfect negative or inverse correlation between speed and time travelled. As speed decreases, travel time increases. The r-score is -1.00. All XY values fall on the least squares or regression line. **Please note:** These are perfect correlations because the X and Y variables are not independent, which is a serious violation of the requirements for correlation.

Table 3: Perfect Positive and Negative Correlations



It also must be pointed out that perfect correlations are extremely rare. It is highly unlikely that we will ever see a perfect positive or negative correlation in the social sciences.

Contrast these perfect correlations with two values that have an r-score of 0.0001, which is to say, two values with no correlation. Figure 3 shows such a scatter diagram. Note the wide dispersion of X Y values from the flat least squares line.

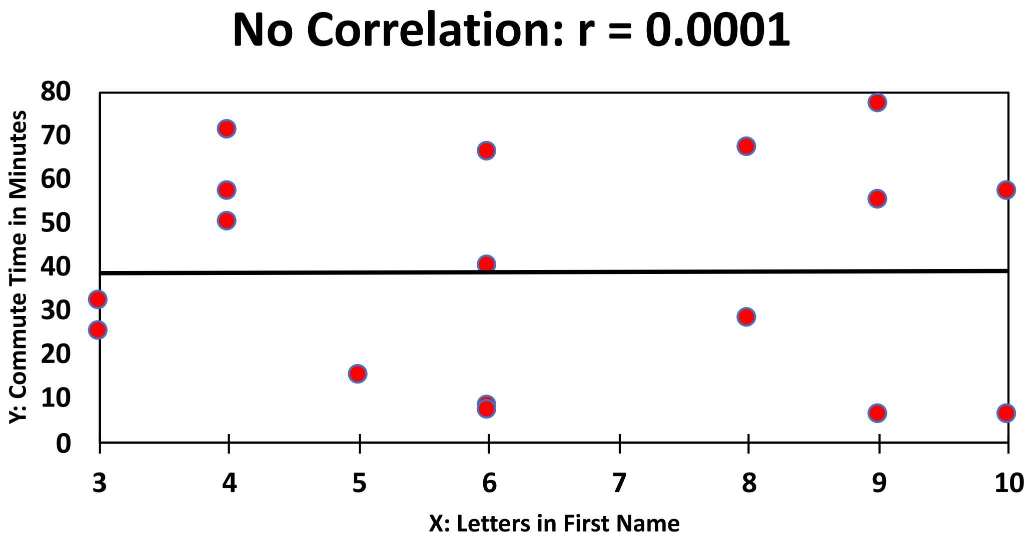
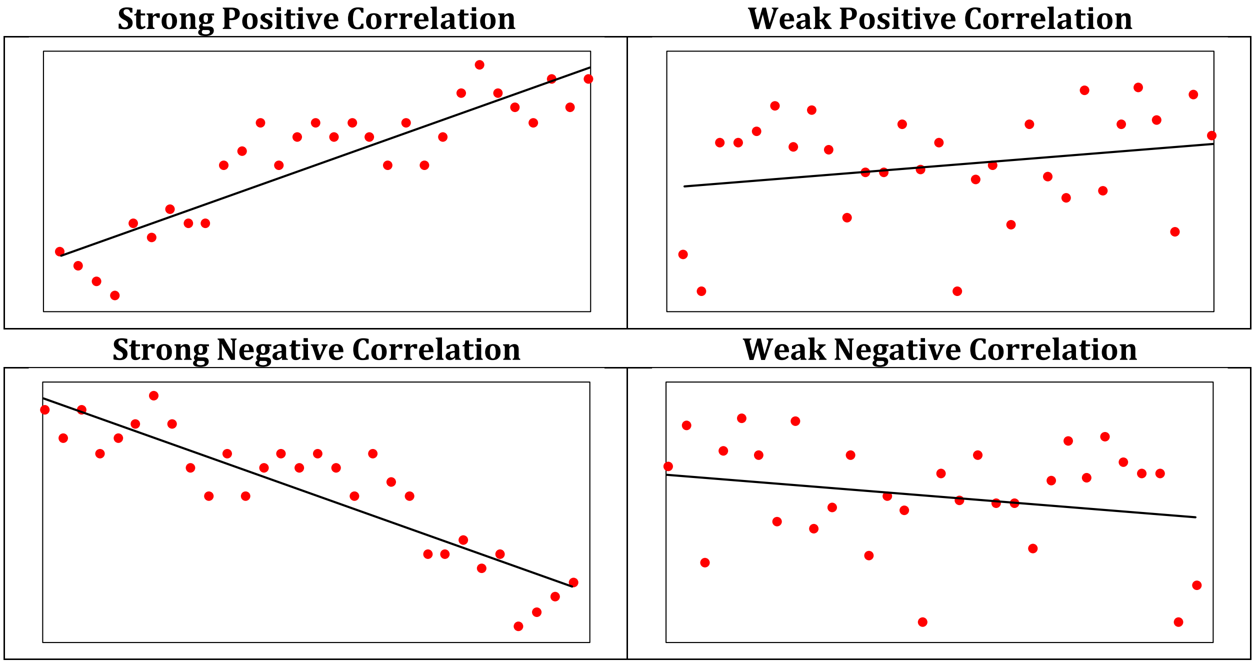

No Correlation: r = 0.0001

Figure 3: A scatter diagram with no correlation, r = 0.0001

As displayed in Table 4, the difference between a strong and weak correlation is shown by the dispersion of the XY values from the least squares line. The weaker the correlation, the greater the dispersion of the XY variables from the least squares line, or the looser the fit to this regression line.

Table 4: The difference between strong and weak correlations


Strong Positive Correlation

Weak Positive Correlation

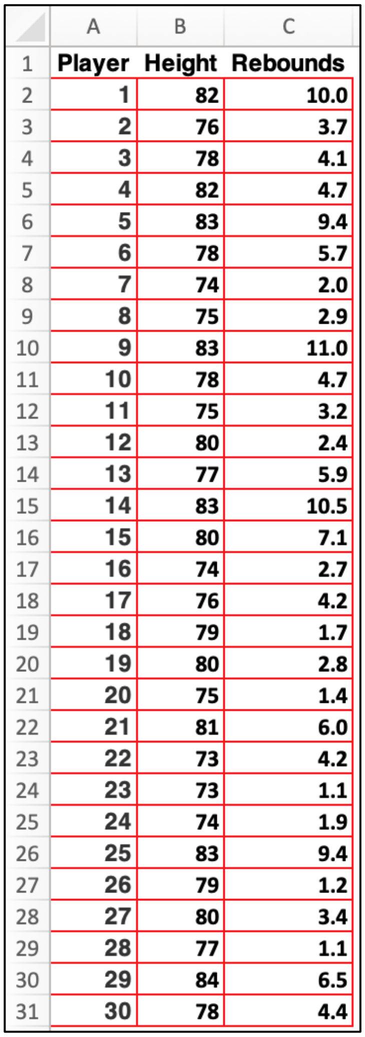
Strong Negative Correlation

Weak Negative Correlation

III. Correlation: The Relationship Between an NBA Player's Height and Average Rebounds per Game

Is the height of an athlete who plays in the National Basketball Association associated with his average number of rebounds per game? A rebound is a statistic awarded to a player who takes possession of a loose ball after a missed field goal or foul shot. It is considered a very important statistic because when a player gets a rebound, his team takes possession of the ball, and the time a team controls the ball is associated with winning games.

Figure 4 shows data on a random sample of thirty players from the 2019-20 season. The sample was taken on December 12, 2019 from data posted on http://stats.nba.com. Players' heights ranged from 73" (6'1") to 84" (7'0"). The mean height is 78.33" with a standard deviation of 3.36". There are no outliers because the height of all the players is within ±3.29 standard deviations from the mean. The average number of rebounds per game ranged from 1.1 to 11.0. The mean number of rebounds is 4.6 with a standard deviation of 2.96. There are no outliers.



	A	B	C
1	Player	Height	Rebounds
2	1	82	10.0
3	2	76	3.7
4	3	78	4.1
5	4	82	4.7
6	5	83	9.4
7	6	78	5.7
8	7	74	2.0
9	8	75	2.9
10	9	83	11.0
11	10	78	4.7
12	11	75	3.2
13	12	80	2.4
14	13	77	5.9
15	14	83	10.5
16	15	80	7.1
17	16	74	2.7
18	17	76	4.2
19	18	79	1.7
20	19	80	2.8
21	20	75	1.4
22	21	81	6.0
23	22	73	4.2
24	23	73	1.1
25	24	74	1.9
26	25	83	9.4
27	26	79	1.2
28	27	80	3.4
29	28	77	1.1
30	29	84	6.5
31	30	78	4.4

Figure 4: NBA players' heights and average rebound per game.

Question: Which of the two variables—a player's height or his average rebounds per game—is the independent or predictor variable? **Answer:** A player's height predicts the average number of rebounds per game because the player attained his height before the game began. Height is the independent variable and rebounds are the dependent variable.

A scatter diagram helps to visualize the relationship between the X and Y variables as shown in Figure 5:

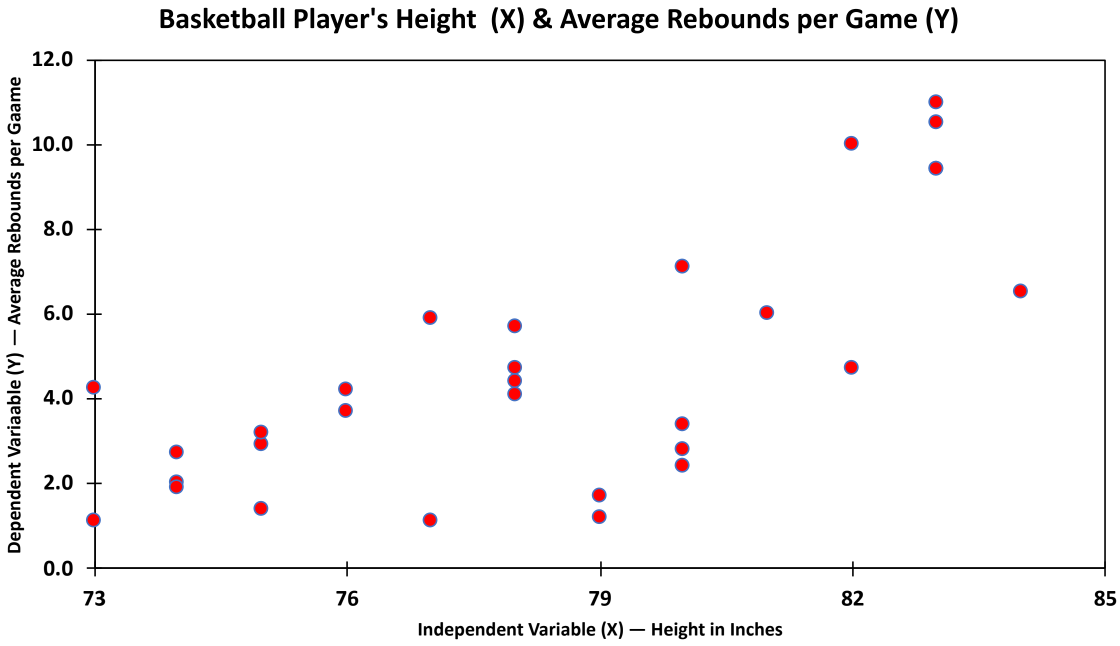


Figure 5: Scatter diagram showing the relationship between height and average rebounds per game

The values for the independent variable, height, are shown on the horizontal or X-Axis and the average rebounds per game, or dependent variable, are shown on the vertical or Y-Axis. Even though the least squares line is not shown, we can see that there is a positive correlation between a player's height and the average rebounds per game, given the fact that as players' heights increase the per game rebound average goes up. In addition, this is not a perfect correlation as the XY variables do not line up in a straight line.

We calculate the coefficient of correlation, r, to determine the strength of the association or correlation. Here, again, is the formula for the coefficient of correlation:

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{(n - 1)s_X s_Y}$$

Equation 1: Coefficient of Correlation Equation

Where: r = Coefficient of Correlation
X = Independent variable
$\bar{X}$ = Mean of the independent variable
$\bar{Y}$ = Mean of the dependent variable
Y = Dependent variable
s_X = Standard deviation for the independent variable

s_Y = Standard deviation for the dependent variable
n = Sample size (number of matched pairs)

There are seven steps to complete this calculation:

1. Count the number of paired observations, n.

2. Calculate the means: $\bar{X}$ and $\bar{Y}$.

3. Calculate the standard deviations: s_X and s_Y.

4. Subtract $\bar{X}$ from each X variable and $\bar{Y}$ from each Y variable.

5. Multiply $(X - \bar{X})$ and $(Y - \bar{Y})$ for each pair of random variables.

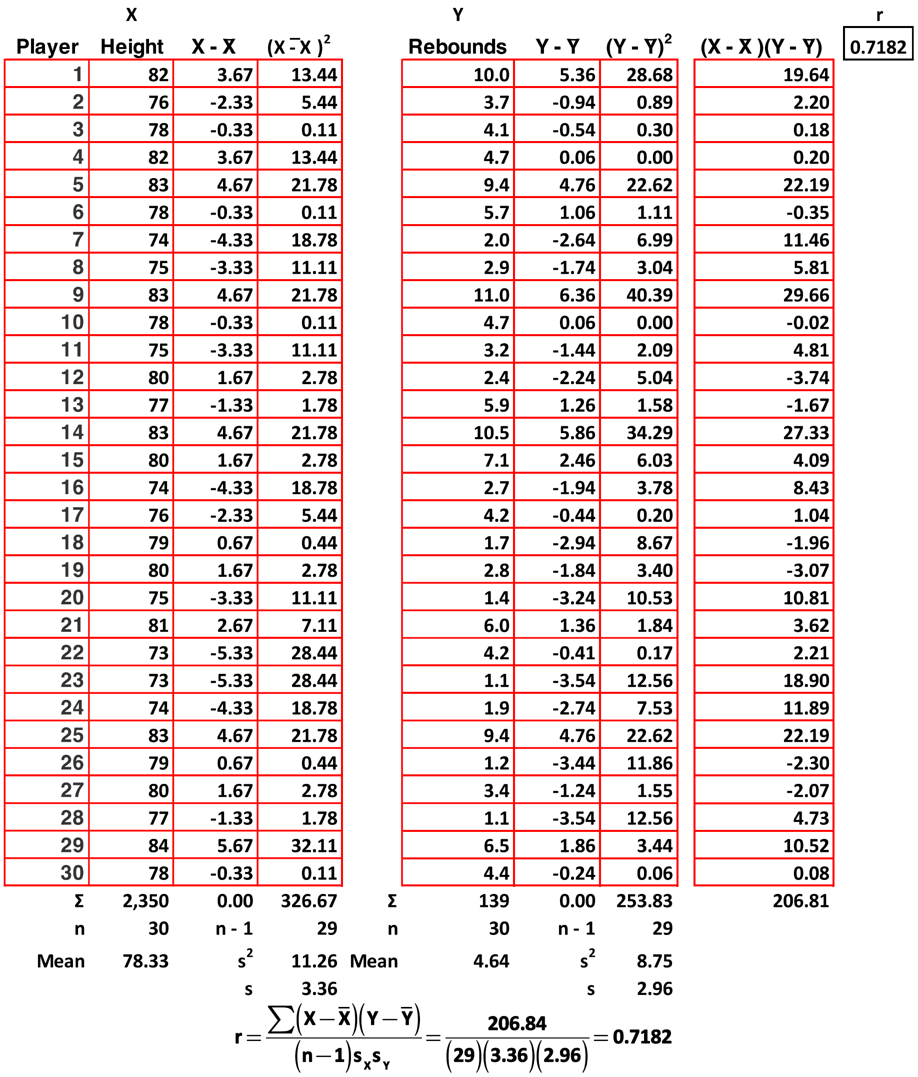
6. Sum $(X - \bar{X})(Y - \bar{Y})$.

7. Complete the formula be dividing $\Sigma X - \bar{X})(Y - \bar{Y})$ by $(n – 1)s_Xs_Y$.

Equation 2 shows the completed calculation. Figure 6 shows all the steps to complete this calculation:

$$\mathrm{r} = \frac{\Sigma(\mathrm{X} - \overline{\mathrm{X}})(\mathrm{Y} - \overline{\mathrm{Y}})}{(\mathrm{n} - 1)\mathrm{s}_{\mathrm{X}}\mathrm{s}_{\mathrm{Y}}} = \frac{206.81}{(29)(3.36)(2.96)} = 0.7182$$

Equation 2: The coefficient of correlation calculation



	X				Y				r
Player	Height	$X - \bar{X}$	$(X - \bar{X})^2$		Rebounds	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$	0.7182
1	82	3.67	13.44		10.0	5.36	28.68	19.64	
2	76	-2.33	5.44		3.7	-0.94	0.89	2.20	
3	78	-0.33	0.11		4.1	-0.54	0.30	0.18	
4	82	3.67	13.44		4.7	0.06	0.00	0.20	
5	83	4.67	21.78		9.4	4.76	22.62	22.19	
6	78	-0.33	0.11		5.7	1.06	1.11	-0.35	
7	74	-4.33	18.78		2.0	-2.64	6.99	11.46	
8	75	-3.33	11.11		2.9	-1.74	3.04	5.81	
9	83	4.67	21.78		11.0	6.36	40.39	29.66	
10	78	-0.33	0.11		4.7	0.06	0.00	-0.02	
11	75	-3.33	11.11		3.2	-1.44	2.09	4.81	
12	80	1.67	2.78		2.4	-2.24	5.04	-3.74	
13	77	-1.33	1.78		5.9	1.26	1.58	-1.67	
14	83	4.67	21.78		10.5	5.86	34.29	27.33	
15	80	1.67	2.78		7.1	2.46	6.03	4.09	
16	74	-4.33	18.78		2.7	-1.94	3.78	8.43	
17	76	-2.33	5.44		4.2	-0.44	0.20	1.04	
18	79	0.67	0.44		1.7	-2.94	8.67	-1.96	
19	80	1.67	2.78		2.8	-1.84	3.40	-3.07	
20	75	-3.33	11.11		1.4	-3.24	10.53	10.81	
21	81	2.67	7.11		6.0	1.36	1.84	3.62	
22	73	-5.33	28.44		4.2	-0.41	0.17	2.21	
23	73	-5.33	28.44		1.1	-3.54	12.56	18.90	
24	74	-4.33	18.78		1.9	-2.74	7.53	11.89	
25	83	4.67	21.78		9.4	4.76	22.62	22.19	
26	79	0.67	0.44		1.2	-3.44	11.86	-2.30	
27	80	1.67	2.78		3.4	-1.24	1.55	-2.07	
28	77	-1.33	1.78		1.1	-3.54	12.56	4.73	
29	84	5.67	32.11		6.5	1.86	3.44	10.52	
30	78	-0.33	0.11		4.4	-0.24	0.06	0.08	
Σ	2,350	0.00	326.67	Σ	139	0.00	253.83	206.81	
n	30	n - 1	29	n	30	n - 1	29		
Mean	78.33	s^2	11.26	Mean	4.64	s^2	8.75		
		s	3.36			s	2.96		

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{(n-1)s_X s_Y} = \frac{206.84}{(29)(3.36)(2.96)} = 0.7182$$

Figure 6: Calculating the coefficient of correlation

A correlation coefficient of 0.7182, which indicates a large positive correlation is based on the interpretations listed Table 2.

Performing these calculations by hand requires 129 mundane calculations that could take at least 20 minutes to complete. Given the inevitable boredom that will ensue making these repetitive calculations, you may make a few careless errors. You should know that Karl Pearson did not calculate his Pearson Product Moment coefficient by hand using paper and pencil. He employed computers. But his computers were not the digital computers we use today, they were intelligent and highly focused women:

> In the history of computing, the humbler levels of scientific work were open, even welcoming, to women. Indeed, by the early twentieth century computing was thought of as women's work and computers were assumed to be female. Respected mathematicians would blithely approximate the problem-solving horsepower of computing machines in 'girl-years' and describe a unit of machine labor as equal to one 'kilo-girl.'[7]

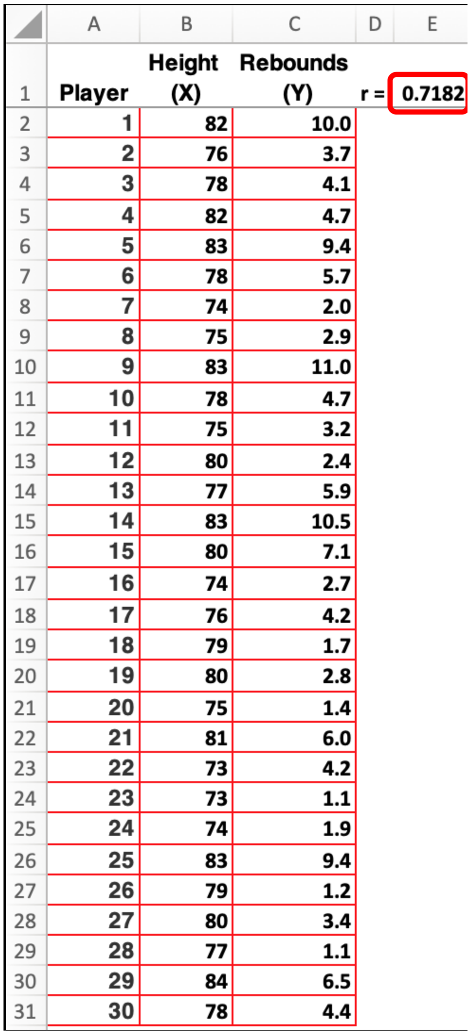
Today, fortunately, women are no longer restricted to the lower levels of scientific work. As a result, you cannot hire teams of female "computers" to perform this grunt-work. You can, however, harness the power of Excel to get some non-exploitative "kilo-girls."

Excel has two built-in correlation functions that calculate the coefficient of correlation in a few seconds. The older function is PEARSON, named after Karl Pearson. The newer function is CORREL, which is said to round off numbers more accurately than the PEARSON function. Both functions require the same number of observations for the independent and dependent variables and ignore text and logical values in the cell addresses. Equation 3 shows the syntax for these functions:

=CORREL(IndependentVariableArray,DependentVariableArray)
=PEARSON(IndependentVariableArray,DependentVariableArray)

Equation 3: CORREL and PEARSON Functions

Figure 7 shows the CORREL function in Cell E1: =CORREL(B2:B31,C2:C31) with the answer 0.7182.



	A	B	C	D	E
1	Player	Height (X)	Rebounds (Y)	r =	0.7182
2	1	82	10.0		
3	2	76	3.7		
4	3	78	4.1		
5	4	82	4.7		
6	5	83	9.4		
7	6	78	5.7		
8	7	74	2.0		
9	8	75	2.9		
10	9	83	11.0		
11	10	78	4.7		
12	11	75	3.2		
13	12	80	2.4		
14	13	77	5.9		
15	14	83	10.5		
16	15	80	7.1		
17	16	74	2.7		
18	17	76	4.2		
19	18	79	1.7		
20	19	80	2.8		
21	20	75	1.4		
22	21	81	6.0		
23	22	73	4.2		
24	23	73	1.1		
25	24	74	1.9		
26	25	83	9.4		
27	26	79	1.2		
28	27	80	3.4		
29	28	77	1.1		
30	29	84	6.5		
31	30	78	4.4		

Figure 7: CORREL Function

A correlation of 0.7182 (71.82 percent) is a large, positive correlation that indicates that as players' heights increase, rebound averages go up.

Excel's Data Analysis ToolPak will also calculate the correlation coefficient. On the Data ribbon, click on the Data Analysis icon. Depending on the version of Excel you are using, this icon looks like the ones shown in Figure 8. The icon on the left is from Excel 2019 on the Macintosh, and on the right is the icon from the Windows version of Excel 365.

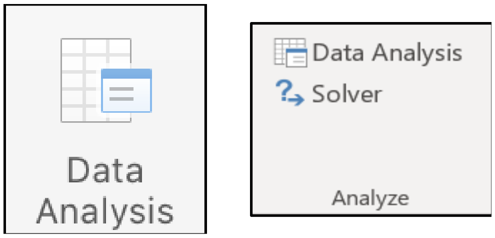


Figure 8: Data Analysis Icon

Once you click on this icon, the Analysis Tools window pops up as shown in Figure 9. Select Correlation and click OK.

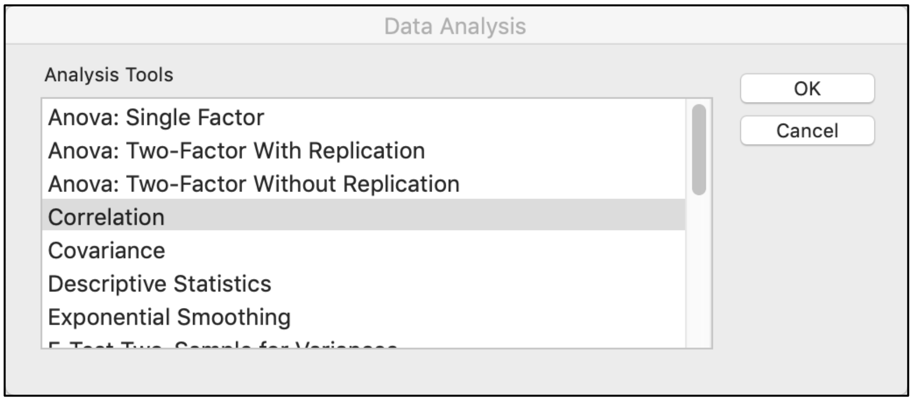
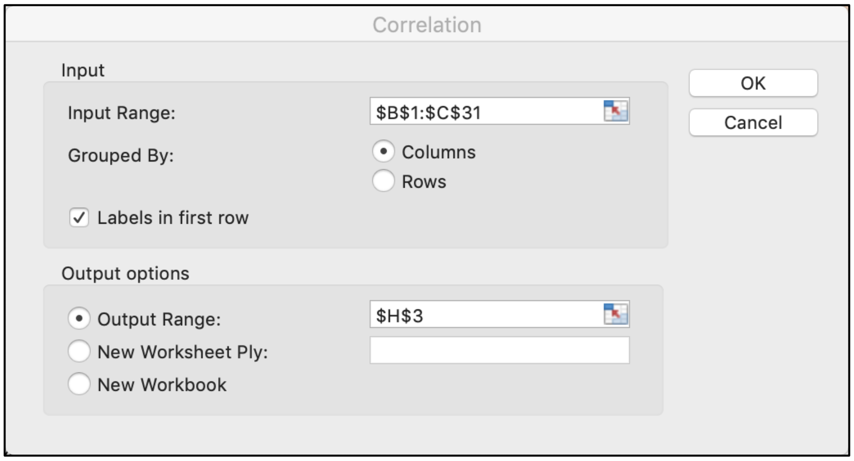


Figure 9: Data Analysis window

After selecting Correlation, a new Correlation window pops-up. This window allows you to select the input range and output options. The input range is Cells $B$1:$C$1. **Please note:** The independent variable should be in column B and the dependent variable is in Column C. "Labels" in the first row are checked because cells B1 and C1 have the names of the independent and dependent variables. The selected output range is set as $H$1 on the same worksheet as the data. See Figure 10:



Correlation

Input

Input Range: $B$1:$C$31

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $H$3

○ New Worksheet Ply:

○ New Workbook

Figure 10: Correlation Input window

After you enter this information, click OK and Excel will complete the analysis as shown in Figure 11.

	Height (X)	*Rebounds (Y)*
Height (X)	1	
Rebounds (Y)	0.71819506	1

Figure 11: Correlation Analysis output

Excel's DataAnalysis Tool will often save you time compared to using this program's built-in functions, which are far faster than performing these calculations by hand. Unfortunately, Excel's Correlation tool is not as robust as its Regression tool. It fails to report the coefficient of determination. The regression tool, on the other hand, will calculate the coefficient of correlation and the coefficient of determination, as well as conduct a regression analysis.

The coefficient of determination, or r^2: This statistic indicates how much the variation in the dependent variable is explained by the independent variable. It is very easy to calculate once you have the coefficient of correlation; just square the correlation coefficient, r. Here is the formula for r^2 along with the calculation for our example:

$$r^2 = \left[\frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)s_X s_Y}\right]^2 = \left[\frac{206.81}{(29)(3.36)(2.96)}\right]^2 = 0.7182^2 = 0.5158$$

Equation 4: Equation for the Coefficient of Determination

The coefficient of determination is more precise than the coefficient of correlation because it does not use “tee-shirt” effect sizes—small, medium, and large that were shown in Table 2—to describe the strength of the correlation. A coefficient of determination of 0.5158 means that a player’s height explains 51.58 percent of his average rebounds per game. Height, while the most important predictor of rebounds, is not a perfect predictor because it fails to explain 48.42 percent of a player’s rebounds, found by 1.0000 – 0.5158 = 0.4842.

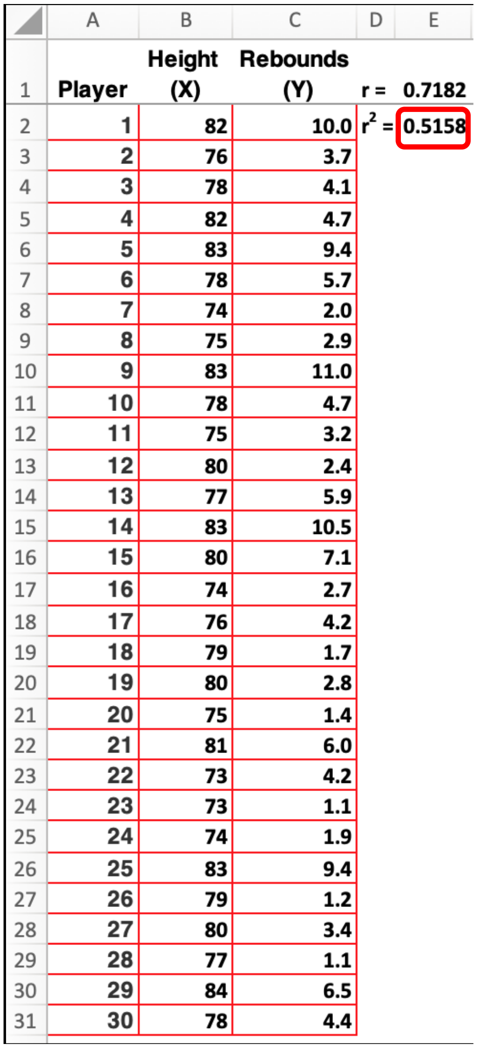
You can use Excel’s RSQ function to calculate the coefficient of correlation, but you may wonder why you should bother. After all, squaring 0.7182 on a handheld calculator is very easy. The reason for using RSQ is that Excel is more precise than a handheld calculator because it calculates values to fifteen digits past the decimal point. Our r score of 0.7182 is actually 0.7181950614491000. When Excel calculates the value of r^2, it uses all 15 digits even when it displays the result rounded off to only two or four digits past the decimal point.

The syntax for RSQ is shown in Equation 5:

=RSQ(DependentVariableArray,IndependentArray)

Equation 5: RSQ Syntax

For our example, Figure 12 shows the results of the RSQ function in cell E2.



	A	B	C	D	E
1	Player	Height (X)	Rebounds (Y)	r =	0.7182
2	1	82	10.0	r^2 =	0.5158
3	2	76	3.7		
4	3	78	4.1		
5	4	82	4.7		
6	5	83	9.4		
7	6	78	5.7		
8	7	74	2.0		
9	8	75	2.9		
10	9	83	11.0		
11	10	78	4.7		
12	11	75	3.2		
13	12	80	2.4		
14	13	77	5.9		
15	14	83	10.5		
16	15	80	7.1		
17	16	74	2.7		
18	17	76	4.2		
19	18	79	1.7		
20	19	80	2.8		
21	20	75	1.4		
22	21	81	6.0		
23	22	73	4.2		
24	23	73	1.1		
25	24	74	1.9		
26	25	83	9.4		
27	26	79	1.2		
28	27	80	3.4		
29	28	77	1.1		
30	29	84	6.5		
31	30	78	4.4		

Figure 12: Coefficient of Determination = 0.5158

The formula is cell E3 is:

$$= \mathrm{RSQ}(B2{:}\ B31{:}\ A2{:}\ A31)$$

Equation 6: =RSQ Function for Figure 12, Cell E3

V. Testing the Significance of the Correlation With a t-test

In our example, we have a large positive correlation. But, a question arises: Could there actually be no correlation in the population? Is rho, ρ, actually zero? We can answer this question using a Null Hypothesis test.

Here are the Null and Alternate Hypotheses: H_0: ρ = 0; H_1: ρ ≠ 0.

This is a two-tailed test as evidenced by the not equal sign, ≠, in the Alternate Hypothesis. If the test statistic falls in the left-tail, we have a negative or inverse correlation. If the test statistic falls in the right-tail, we have a positive correlation. When the test statistic does not fall in either rejection region, we fail to reject the null hypothesis and conclude that there is no correlation in the population. This conclusion, of course, assumes that our test has sufficient statistical power, which we can estimate *a priori* or *post hoc*.

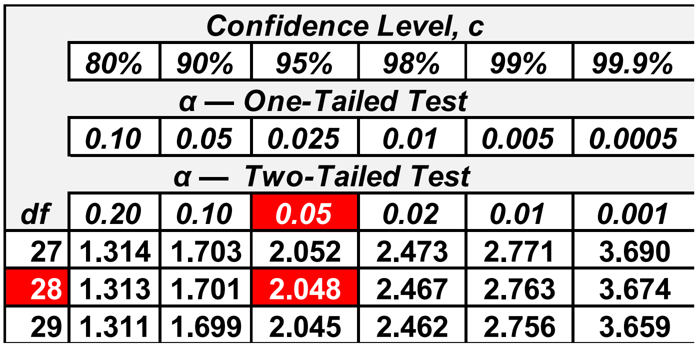
As always, the second step in the Null Hypothesis test is to set the significance level, which we will set at 0.05. Recall that this is our tolerance for committing a Type I error. A Type I error in this context rejects the Null Hypothesis when there really is no correlation in the population.

Then we select the test statistic. Equation 7 shows the test statistic for this t-test:

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

Equation 7: t-test for the correlation coefficient with n - 2 degrees of freedom

Here is how to write the decision rule for this two-tailed test with 28 degrees of freedom (30 – 2). We can find the critical values for t using the t-test table or Excel's TINV function, =TINV(alpha,df). Figure 13 shows the table of critical values for a t-distribution.



	Confidence Level, c					
	80%	90%	95%	98%	99%	99.9%
	α — One-Tailed Test					
	0.10	0.05	0.025	0.01	0.005	0.0005
	α — Two-Tailed Test					
df	0.20	0.10	0.05	0.02	0.01	0.001
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659

Figure 13: t-distribution table, critical value for t is -2.048 and +2.048

The formula for finding the critical value in Excel is: =TINV(0.05,28). The critical values for this two-tailed test are -2.048 and +2.048.

The decision rule: Reject the null hypothesis if t is less than -2.048 or greater than +2.048. Figure 14 shows the chart of the t-distribution with the two rejection regions.

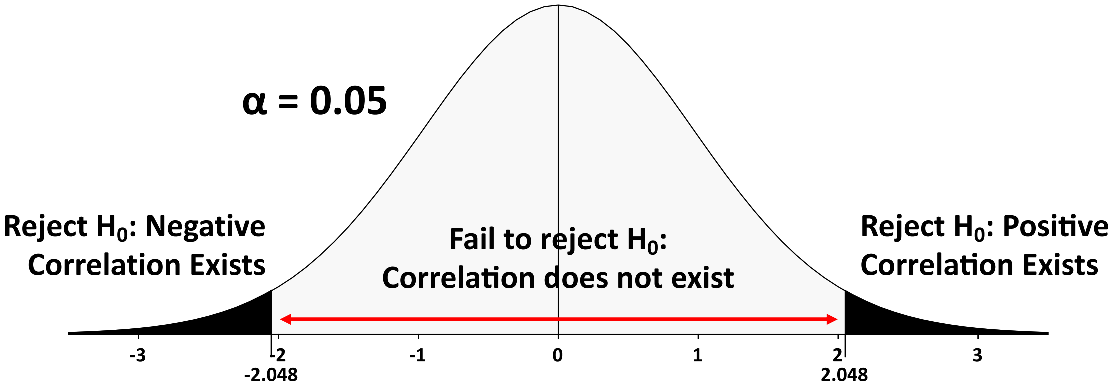


Figure 14: t-distribution with 28 degrees of freedom and a 5% significance level

Our decision to reject or fail to reject the null hypothesis can also be made using the p-value. We should reject the Null Hypothesis if the p-value is less than or equal to the significance level.

Next, we calculate the value of the test statistic, p-value, statistical power, and make a decision regarding the null hypothesis. The value of the test statistic is very high, 5.461, found by:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7182\sqrt{30-2}}{\sqrt{1-0.5158}} = 5.461$$

Equation 8: The equation for the test statistic

This is a very large t-value. It is clearly in the rejection region.

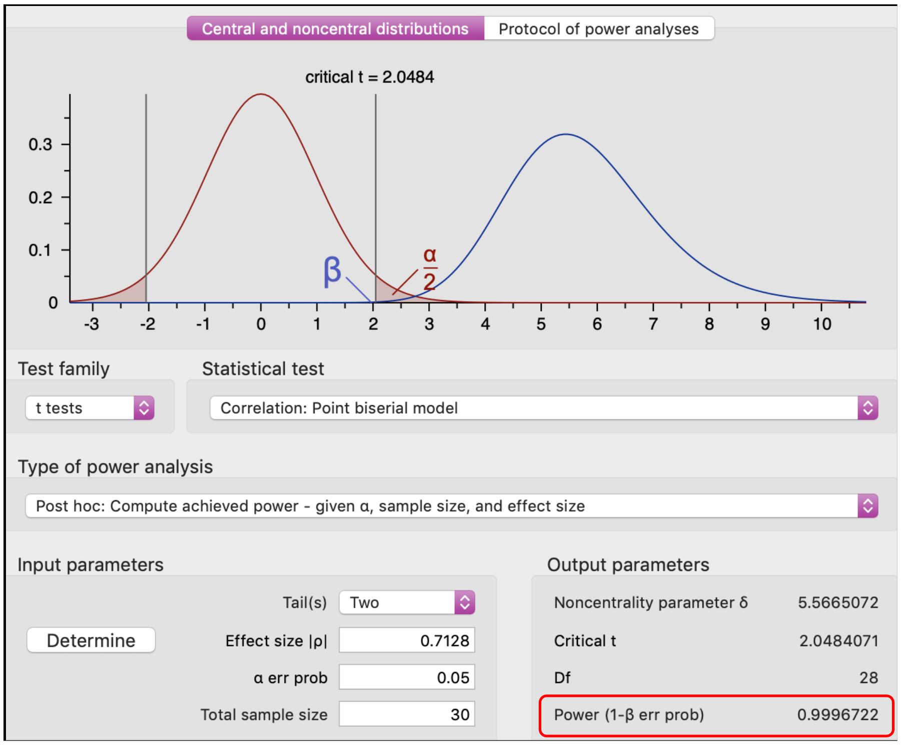
The p-value is 0.00001, which can be found using the Excel function shown in Equation 9:

```
=TDIST(ABS(teststatistic),df,number_of_tails)
=TDIST(ABS(5.461),28,2)
```

Equation 9: Excel's formula for p-value

This tiny p-value is an indication that we should reject the null hypothesis. There is only a 0.00001 chance that the results are merely due to sampling error or 1 in a 100,000. We conclude that there is a correlation in the population. For tiny p-values like this one, we would report the value as <0.001.

We can conduct a *post hoc* power analysis using G*Power; although strictly speaking, this step is not necessary because we are rejecting the Null Hypothesis. As shown in Figure 15, the test family is t tests, The statistical test is Correlation: Point biserial module. The type of power analysis is Post hoc: Compute achieve power – given α, sample size, and effect size. The input parameters are Tail(s) = two, Effect size (ρ) = 0.7128, which is our r-score, α err prob is the level of significance, 0.05, and Total sample size is 30. Statistical power is greater than 99.9 percent and the probability of a Type II error (failing to reject a wrong Null Hypothesis is less than 0.1 percent.


Central and noncentral distributions
Protocol of power analyses
critical t = 2.0484
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
β
α
2
Determine

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

Tail(s): Two
Effect size |ρ|: 0.7128
α err prob: 0.05
Total sample size: 30

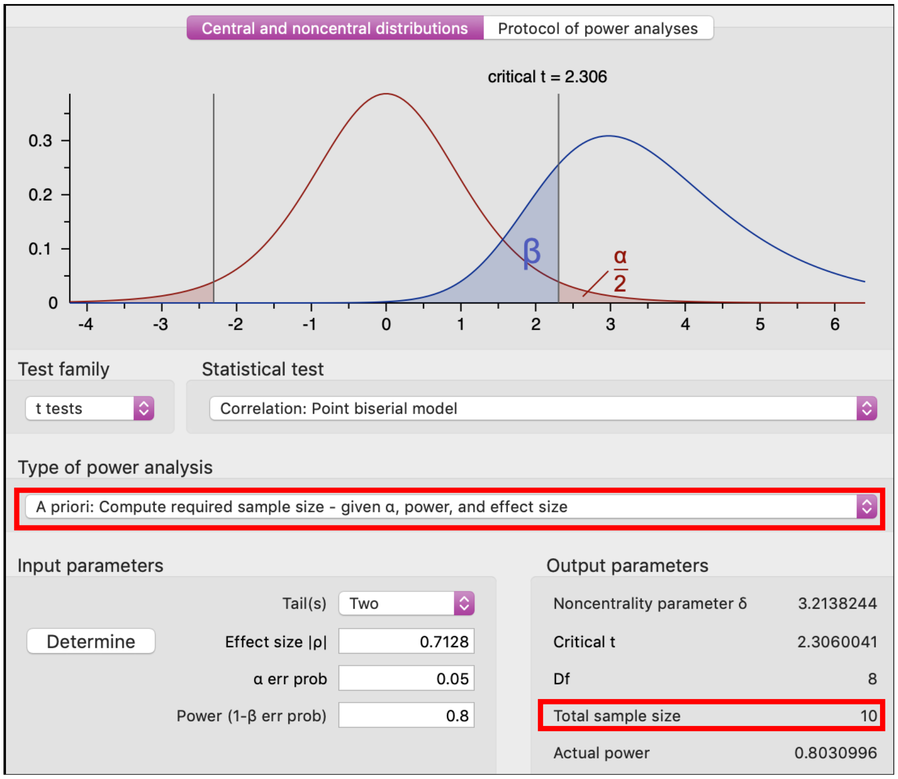
Output parameters

Noncentrality parameter δ: 5.5665072
Critical t: 2.0484071
Df: 28
Power (1-β err prob): 0.9996722

Figure 15: Post Hoc Power Analysis

Given the large effect size and very high statistical power, we could have achieved sufficient statistical power with a smaller sample size. The *a priori* power analysis shown in

Figure 16 indicates sufficient statistical power of 80 percent could have been achieved with a sample of only 10 matched pairs.


Central and noncentral distributions
Protocol of power analyses
critical t = 2.306
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
5
6
β
α/2
Determine

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s): Two

Effect size |ρ|: 0.7128

α err prob: 0.05

Power (1-β err prob): 0.8

Output parameters

Noncentrality parameter δ: 3.2138244

Critical t: 2.3060041

Df: 8

Total sample size: 10

Actual power: 0.8030996

Figure 16: A Priori Power Analysis

Conclusion: There is a large positive correlation in the population between an NBA player's height and his average rebounds per game.

VI. Some Caveats About Linear Correlation

1. The coefficients of correlation and determination describe the strength of the linear relationship between the X and Y variables.

2. r and r^2 are measures of association, ***not*** causation.

3. High r and r^2 values do ***not*** necessarily mean that X is a useful predictor of Y.

4. To determine how well the independent variable, X, predicts the dependent variable, Y, we must conduct a regression analysis.

VII. Simple of Ordinary Least Square (OLS) Regression

We are now going to conduct a regression analysis three ways: 1) by hand, 2) using Excel's built-in functions, and 3) Using Excel's Regression Analysis Tool. We will discover that Excel lets us perform the necessary calculations far faster than doing them with pencil and paper or a handheld calculator. We will also see that Excel's Regression Analysis tool performs most of the important regression calculations very quickly.

Regression models estimate values of the dependent variable Y based on a selected value of the independent variable X. Regression is based on the least square line. This line is sometimes called the regression line or the best fit line. Excel calls this line the trendline, but Excel has a variety of trendlines, so you must select the *linear* trend line to add a least squares line to an XY chart. The least squares regression line is the straight line between the observed X variable and predicted Y variables that make the vertical distance, or the residuals, from each observed Y variable as small as possible. The least squares line is often called the best-fit line because it is the only straight line through all the XY variables that minimizes the sum of the squares of the vertical distance between the observed Y value and the predicted Y value. The predicted Y value is symbolized by the Y-hat symbol, Ŷ. The differences between Y and Ŷ are called residuals, errors, or derivatives.

A useful first step in any simple linear regression analysis is to draw a scatter diagram. Doing so will help determine whether the relationship between X and Y is linear. Figure 17 shows a scatter diagram with a non-linear relationship.

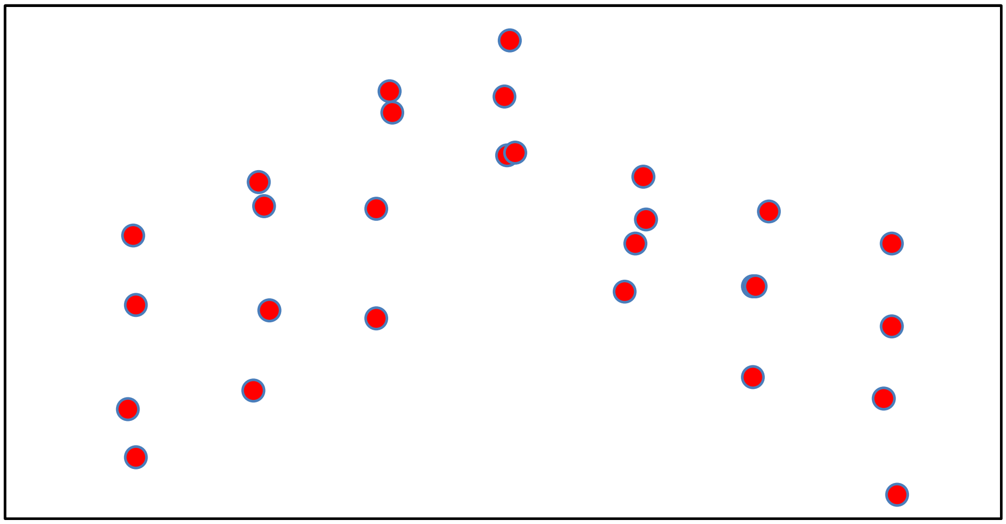


Figure 17: Scatter Diagram for a Nonlinear Relationship

Figure 18 shows that the line that minimizes the distance between each XY variable is curved. The relationship between X and Y, therefore, is not linear. We would either have to transform one or both of the variables to create a linear relationship or use a more sophisticated non-linear regression analysis.

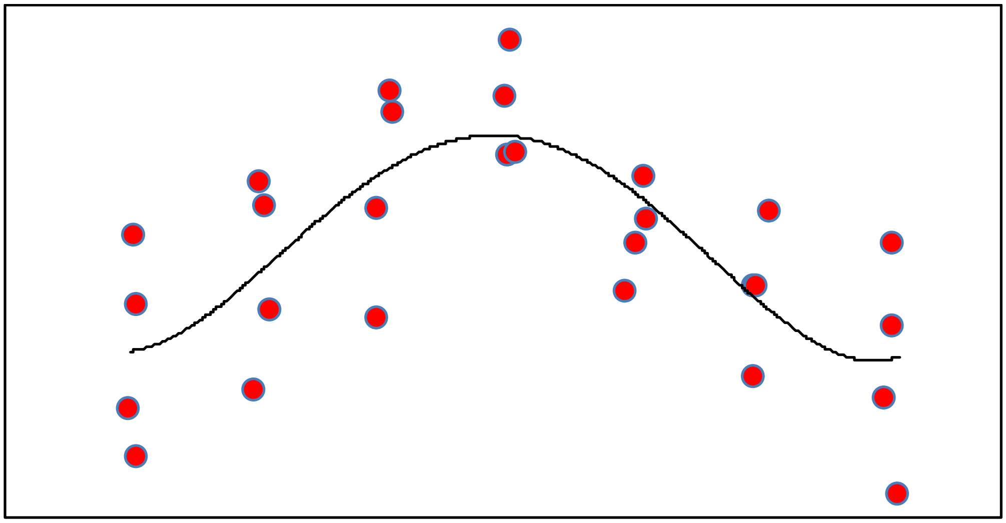


Figure 18: Scatter Diagram for a Nonlinear Relationship With a Nonlinear Trendline

Let's create a scatter diagram for the relationship between a basketball player's height and average rebounds per game with a least squares line using Excel. The first step is to select the X and Y variables on the Excel workbook. To repeat, the X and Y arrays should be in adjacent columns with the X variables to the left of the Y variables. The second step is to select "Insert Chart" and select a "Scatter" chart. Select the one without lines between the XY variables as shown in Figure 19.

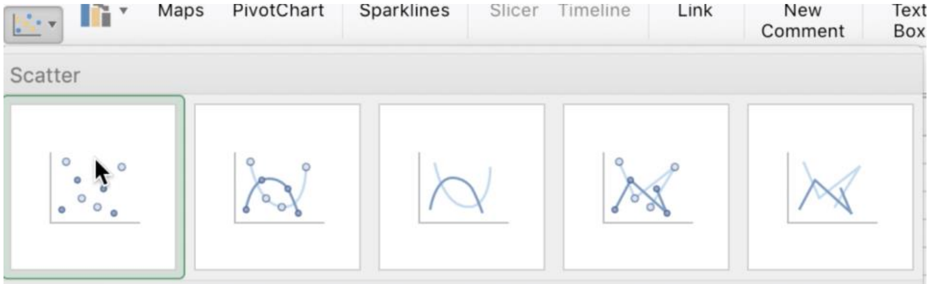


Figure 19: XY Scatter Chart

Excel produces the chart shown in Figure 20. This chart needs to have the least squares trendline added. It also needs proper formatting.

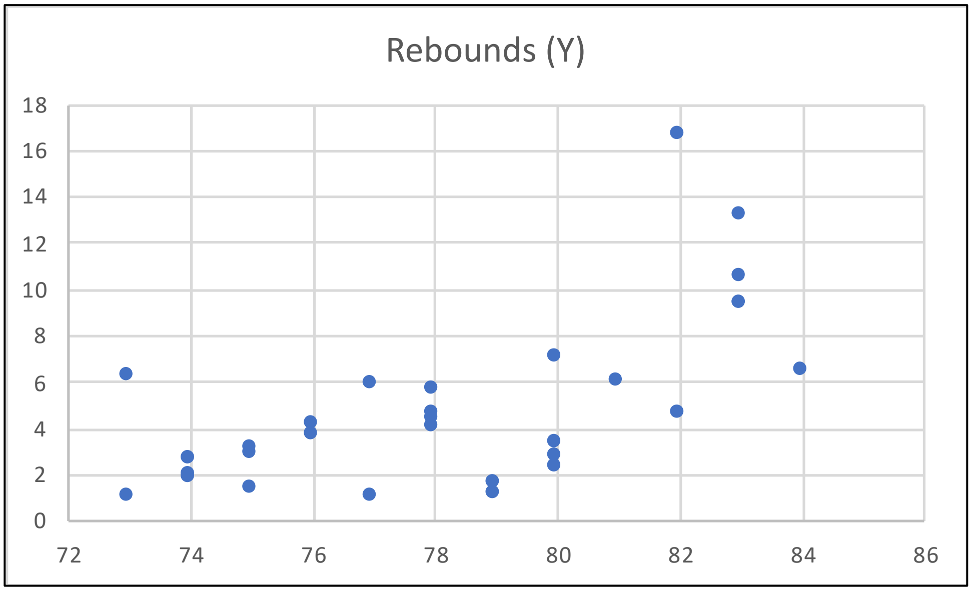


Figure 20: Excel's Scatter Chart Output

Adding a least squares line is very easy in Excel. Highlight the chart. Click on the "Add Chart Elements" icon shown in Figure 21:

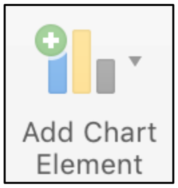
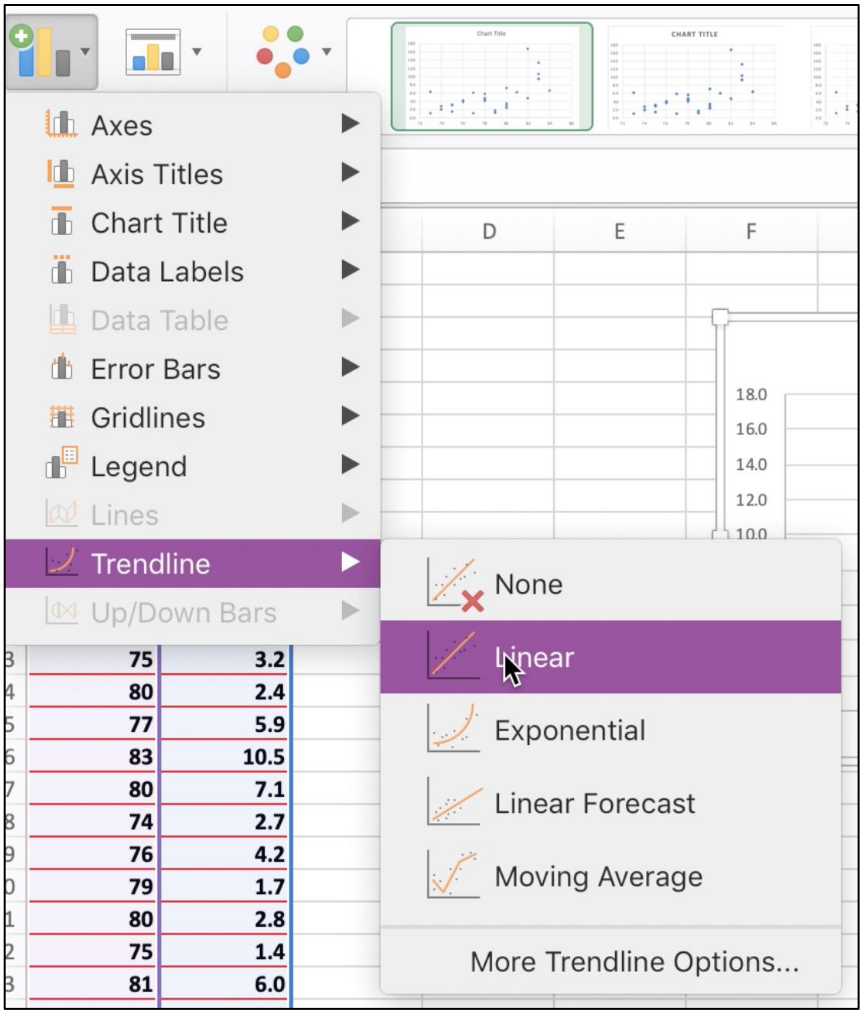


Figure 21: Add Chart Elements Icon

The Chart Elements Options window appears. Select Trendline and Linear as shown in Figure 22:



3	75	3.2
4	80	2.4
5	77	5.9
6	83	10.5
7	80	7.1
8	74	2.7
9	76	4.2
0	79	1.7
1	80	2.8
2	75	1.4
3	81	6.0

Figure 22: Chart Elements Options Window

Excel adds the least squares line to the chart shown in Figure 23:

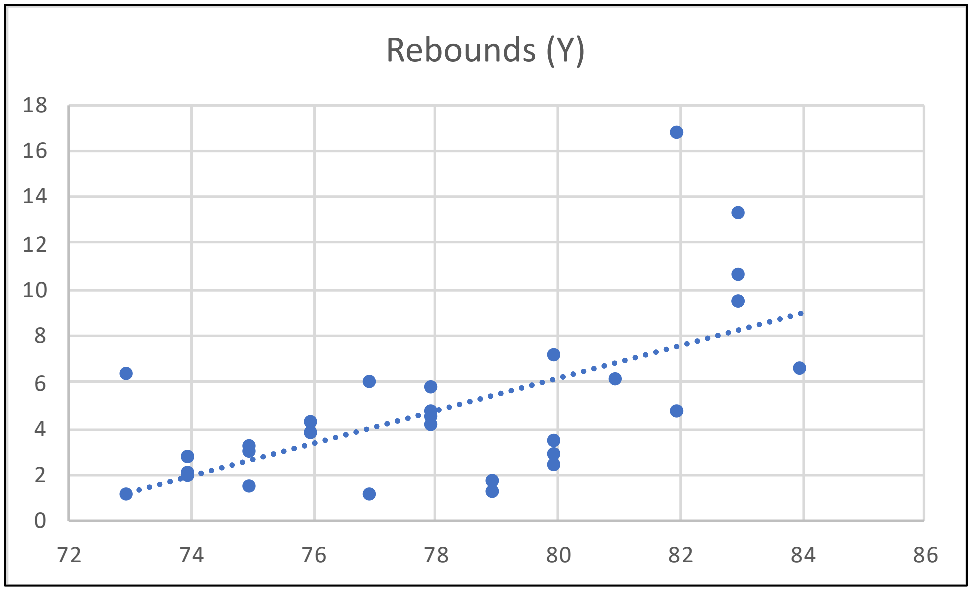


Figure 23: Excel's Scatter Chart Output With Least Squares Line

The chart is now ready for further formatting. Figure 24 shows a properly formatted scatter diagram with a least squares line. Part of the formatting is to add a chart title as well as titles for the X-Axis and Y-Axis. This chart also shows the equation for the least squares line, which is called the regression equation, and the coefficient of determination, r^2.

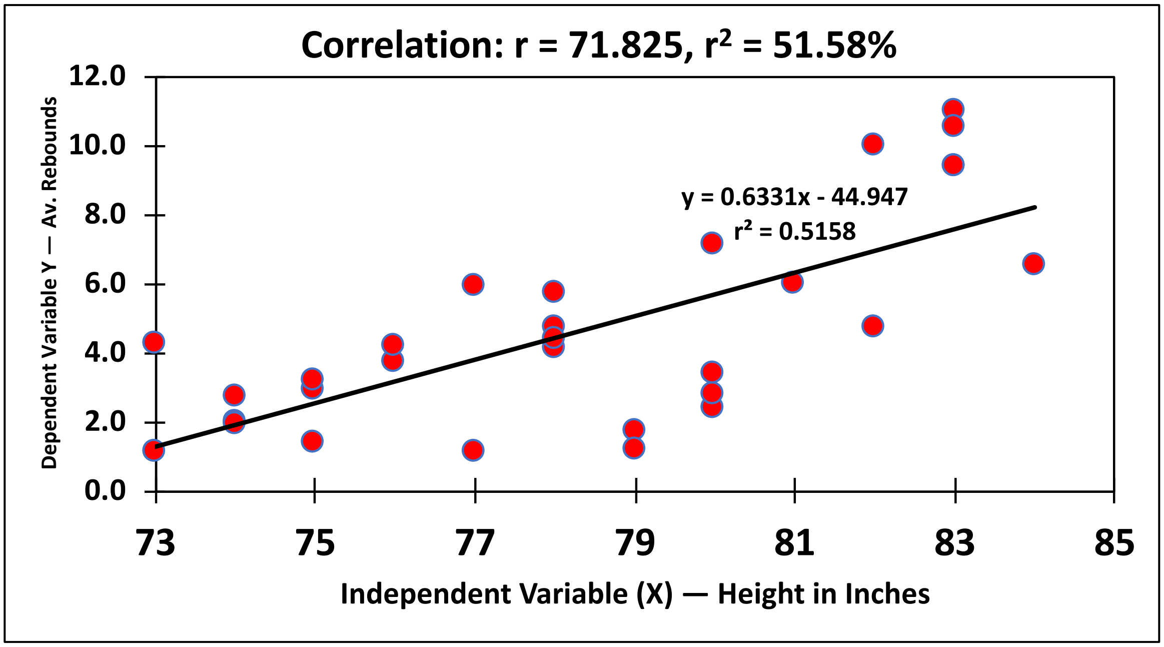


Figure 24: Formatted Scatter Diagram with Least Squares Line

From this scatter diagram we can see that the X and Y variables have a strong positive linear relationship. We can now turn to performing the regression analysis. **Please note:** You can also construct the least squares line using Excel's LINEST function. Equation 10 shows the syntax for this function:

=LINEST(known_y's,known_x's,constant,stats)

Equation 10: LINEST Function Syntax

Where: known_y's is the cell address of the range of dependent variables
known_x's is the cell address of the range of independent variable
constant is an optional argument, it TRUE or omitted, the Y-Intercept is calculated normally, if FALSE, the Y-Intercept is forced to be zero
Stats is an optional argument, if TRUE, the LINEST function returns an array with additional regression statistics, if FALSE or omitted, LINEST returns the Y-Intercept constant and the slope coefficient

As shown in Figure 25, there are three critical equations used in regression: 1) The slope of the least squares line or b; 2) the Y-Intercept or a; and 3) the regression equation, which predicts the Y variable, Ŷ for a given X variable:

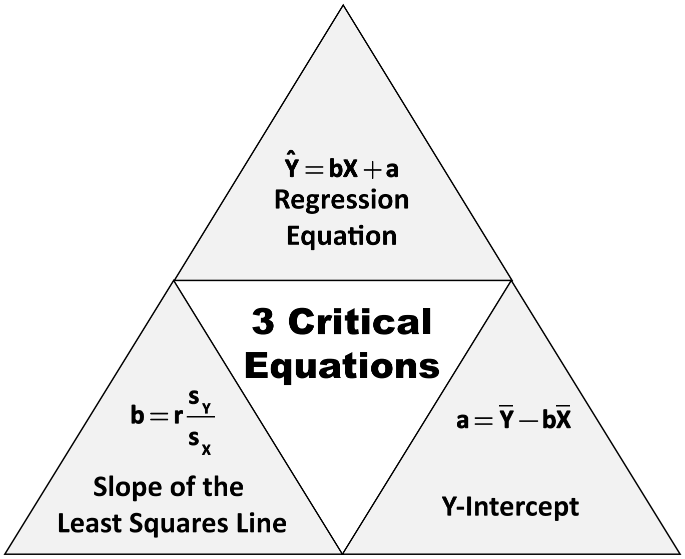


Figure 25: The Three Critical Regression Equations

1) The Slope of the Least Squares Line, b

The first equation calculates the slope of the least squares line or b. The slope of the line indicates the rate the line rises or falls over the horizontal distance, which is the distances the independent variance increases. Equation 11 shows this equation and the calculation for our example:

$$b - r\frac{s_Y}{s_X} = 0.7182\frac{2.96}{3.36} = 0.6331$$

Equation 11: Slope of the Least Squares Line

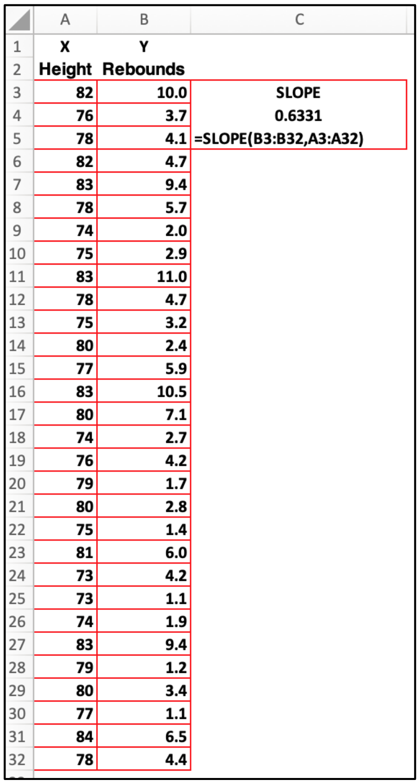
Where: r = Coefficient of Correlation
s_Y = Standard Deviation of the Dependent Variable
s_X = Standard Deviation of the Independent Variable
b = Slope of the line (average change in the predicted value of Y, $\hat{Y}$, for each change in X)

The slope of the least squares line for our example is 0.6331. This means that when a player's height increases by an inch, his average rebounds per game are expected to go up by 0.6331. The problem with this equation is that we must already know the values of the standard deviations of X and Y, as well has the coefficient of correlation, r.

Microsoft Excel's SLOPE function can quickly calculate the slope of the least squares line ***without*** having to calculate the standard deviations for X and Y or the coefficient of correlation. Equation 12 shows the syntax for this function and Figure 26 shows the calculation for our example.

=SLOPE(Known-Y-Values,Known-X-Values)

Equation 12: The SLOPE Function Syntax



	A	B	C
1	X	Y	
2	Height	Rebounds	
3	82	10.0	SLOPE
4	76	3.7	0.6331
5	78	4.1	=SLOPE(B3:B32,A3:A32)
6	82	4.7	
7	83	9.4	
8	78	5.7	
9	74	2.0	
10	75	2.9	
11	83	11.0	
12	78	4.7	
13	75	3.2	
14	80	2.4	
15	77	5.9	
16	83	10.5	
17	80	7.1	
18	74	2.7	
19	76	4.2	
20	79	1.7	
21	80	2.8	
22	75	1.4	
23	81	6.0	
24	73	4.2	
25	73	1.1	
26	74	1.9	
27	83	9.4	
28	79	1.2	
29	80	3.4	
30	77	1.1	
31	84	6.5	
32	78	4.4	

Figure 26: Slope of the Least Squares Line Calculation in Excel

2) The Y-Intercept, a

The second equation is for the Y-Intercept (a). This is the point where the least squares line intersects with, or crosses, the vertical or Y-Axis. A positive Y-Intercept means that the least squares line crosses the Y-Axis above its origin and a negative Y-Intercept indicates that the regression line crosses the Y-Axis below its origin. In practice, a negative Y-Intercept has no real meaning other than that it is needed for the regression equation. In our example, for instance, it is impossible for a player to have negative rebounds.

Equation 13 shows the formula for the Y-Intercept and the calculation for our example. The intercept is -44.95.

$$\mathrm{a} = \overline{\mathrm{Y}} - \mathrm{b}\overline{\mathrm{X}} = 4.64 - (0.6311 * 78.33) = -44.95$$

Equation 13: Y-Intercept

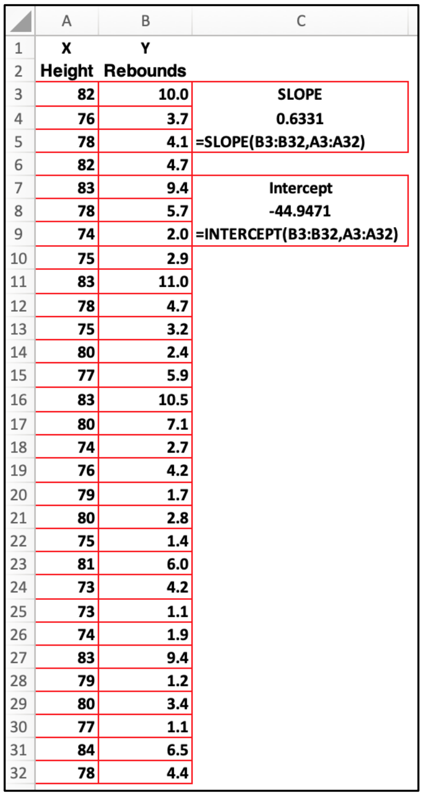
Where: a = Y-Intercept

b = Slope of the line (average change in the predicted value of Y, $\hat{Y}$, for each change in X)
$\bar{X}$ = Mean of the independent variable
$\bar{Y}$ = Mean of the dependent variable

Excel's INTERCEPT function quickly calculates the Y-Intercept without entering the slope of the line, $\bar{X}$, and $\bar{Y}$. Equation 14 shows the syntax for this function and Figure 27 shows the calculation for our example.

=INTERCEPT(Known-Y-Values,Known-X-Values)

Equation 14: The Intercept Function Syntax



	A	B	C
1	X	Y	
2	Height	Rebounds	
3	82	10.0	SLOPE
4	76	3.7	0.6331
5	78	4.1	=SLOPE(B3:B32,A3:A32)
6	82	4.7	
7	83	9.4	Intercept
8	78	5.7	-44.9471
9	74	2.0	=INTERCEPT(B3:B32,A3:A32)
10	75	2.9	
11	83	11.0	
12	78	4.7	
13	75	3.2	
14	80	2.4	
15	77	5.9	
16	83	10.5	
17	80	7.1	
18	74	2.7	
19	76	4.2	
20	79	1.7	
21	80	2.8	
22	75	1.4	
23	81	6.0	
24	73	4.2	
25	73	1.1	
26	74	1.9	
27	83	9.4	
28	79	1.2	
29	80	3.4	
30	77	1.1	
31	84	6.5	
32	78	4.4	

Figure 27: Intercept of the Least Squares Line Calculation in Excel

3) The Regression Equation to Calculate the Predicted value of Y, $\hat{Y}$

The regression equation calculates the estimated or predicted value of Y, or $\hat{Y}$, for any selected value of X. X and $\hat{Y}$ provide the coordinates for the least squares line. We would need these coordinates if we were to draw it with the least squares line without taking advantage of adding this line using Excel's charting elements tool. Given that we have 30 X-values, we would have to repeat this calculation thirty times. Needless to say, this is a time-consuming chore. Fortunately, we can use Excel's built-in function to make fast work of these calculations. In addition, Excel's Regression tool will calculate Ŷ. Equation 15 shows the formula for the regression equation and the predicted values for Y when X is 80" and 74".

$$= \mathrm{bX} + 1 = (0.6331 * 80) \pm 44.9471 = 5.7$$

A player who is 80" tall is expected to average 5.7 rebounds per game

$$= \mathrm{bX} + 1 = (0.6331 * 74) \pm 44.9471 = 1.9$$

A player who is 74 tall is expected to average 1.9 rebounds per game

Equation 15: The Regression Equation

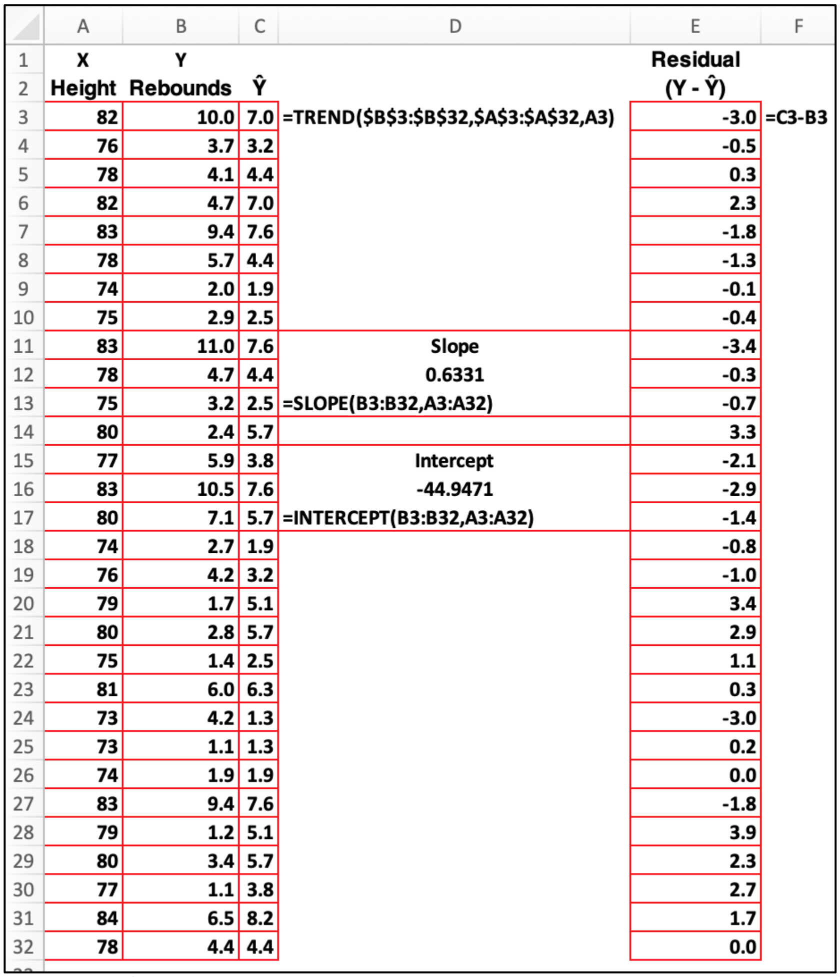
Where:
Ŷ = Estimated or predicted value of Y for a selected X value
A = The Y-Intercept, the value of Y when X is zero, -44.9471 in our example. Yes, no basketball player can be -44.94 inches tall
b = Slope of the line (average change in the predicted value of Y, Ŷ, for each change in X)
X = Any selected value of X

Excel's TREND function will quickly calculate Ŷ. Equation 16 shows the syntax for this function and Figure 28 shows what our workbook looks like when we calculate Ŷ and the residuals.

=TREND(Y-Array,X-Array,new-X,Const)

Equation 16: Syntax for the TREND Function

Where:
Y-Array is the cell array with the Y variables
X-Array is the cell array with the X variables
new-X is the X variable to predict Y (this argument is optional)
Const (Constant) is optional, if blank or TRUE, b is normally distributed, if FALSE, b is set at zero



	A	B	C	D	E	F
1	X	Y			Residual	
2	Height	Rebounds	Ŷ		(Y - Ŷ)	
3	82	10.0	7.0	=TREND($B$3:$B$32,$A$3:$A$32,A3)	-3.0	=C3-B3
4	76	3.7	3.2		-0.5	
5	78	4.1	4.4		0.3	
6	82	4.7	7.0		2.3	
7	83	9.4	7.6		-1.8	
8	78	5.7	4.4		-1.3	
9	74	2.0	1.9		-0.1	
10	75	2.9	2.5		-0.4	
11	83	11.0	7.6	Slope	-3.4	
12	78	4.7	4.4	0.6331	-0.3	
13	75	3.2	2.5	=SLOPE(B3:B32,A3:A32)	-0.7	
14	80	2.4	5.7		3.3	
15	77	5.9	3.8	Intercept	-2.1	
16	83	10.5	7.6	-44.9471	-2.9	
17	80	7.1	5.7	=INTERCEPT(B3:B32,A3:A32)	-1.4	
18	74	2.7	1.9		-0.8	
19	76	4.2	3.2		-1.0	
20	79	1.7	5.1		3.4	
21	80	2.8	5.7		2.9	
22	75	1.4	2.5		1.1	
23	81	6.0	6.3		0.3	
24	73	4.2	1.3		-3.0	
25	73	1.1	1.3		0.2	
26	74	1.9	1.9		0.0	
27	83	9.4	7.6		-1.8	
28	79	1.2	5.1		3.9	
29	80	3.4	5.7		2.3	
30	77	1.1	3.8		2.7	
31	84	6.5	8.2		1.7	
32	78	4.4	4.4		0.0	

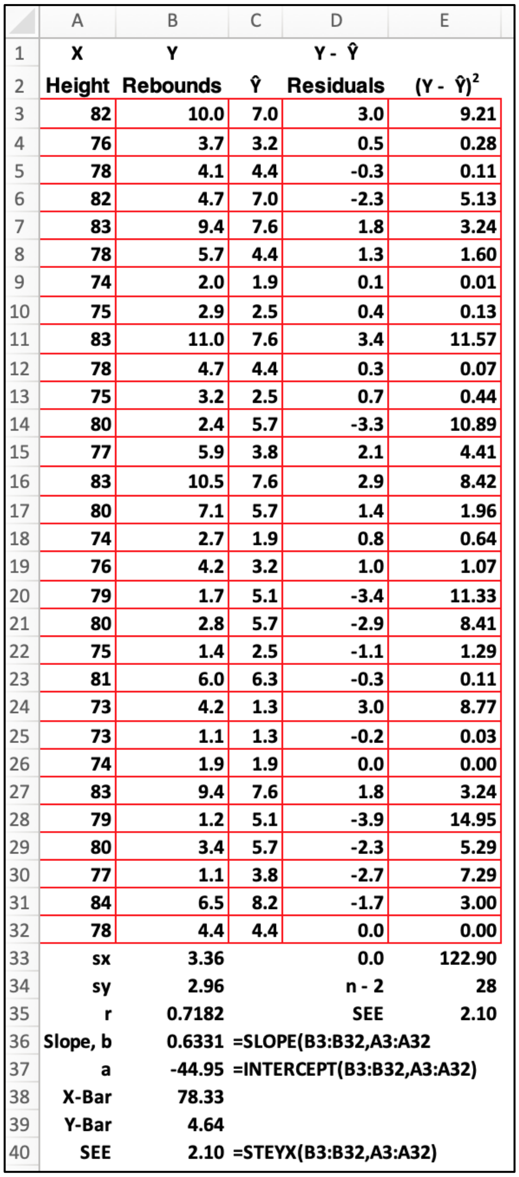
Figure 28: Ŷ and Residuals Calculated Using Excel

The residuals, as previously noted, are the errors in the prediction of the dependent variable. Essentially, they are the portion of the dependent variable that is ***not*** explained by the independent variable. The residuals are found by Y minus Ŷ. On the scatter diagram, the residuals represent the vertical distance between Y and Ŷ.

A critical part of the regression analysis is the Standard Error of the Estimate (SEE). The SEE is essentially the standard deviation of the residuals. It measures the accuracy of the predicted Y values, Ŷ. The smaller the SEE, the higher the correlation coefficient and the stronger the regression model. Equation 17 shows the formula for SEE and the calculation for this problem. Figure 29 shows the data.

$$\mathrm{SEE} = \sqrt{\frac{\Sigma(\mathrm{Y} - \widehat{\mathrm{Y}})^2}{\mathrm{n} - 2}} = \sqrt{\frac{122.90}{30 - 2}} = 2.10$$

Equation 17: Formula and Calculations for SEE



	A	B	C	D	E
1	X	Y		Y - Ŷ	
2	Height	Rebounds	Ŷ	Residuals	(Y - Ŷ)^2
3	82	10.0	7.0	3.0	9.21
4	76	3.7	3.2	0.5	0.28
5	78	4.1	4.4	-0.3	0.11
6	82	4.7	7.0	-2.3	5.13
7	83	9.4	7.6	1.8	3.24
8	78	5.7	4.4	1.3	1.60
9	74	2.0	1.9	0.1	0.01
10	75	2.9	2.5	0.4	0.13
11	83	11.0	7.6	3.4	11.57
12	78	4.7	4.4	0.3	0.07
13	75	3.2	2.5	0.7	0.44
14	80	2.4	5.7	-3.3	10.89
15	77	5.9	3.8	2.1	4.41
16	83	10.5	7.6	2.9	8.42
17	80	7.1	5.7	1.4	1.96
18	74	2.7	1.9	0.8	0.64
19	76	4.2	3.2	1.0	1.07
20	79	1.7	5.1	-3.4	11.33
21	80	2.8	5.7	-2.9	8.41
22	75	1.4	2.5	-1.1	1.29
23	81	6.0	6.3	-0.3	0.11
24	73	4.2	1.3	3.0	8.77
25	73	1.1	1.3	-0.2	0.03
26	74	1.9	1.9	0.0	0.00
27	83	9.4	7.6	1.8	3.24
28	79	1.2	5.1	-3.9	14.95
29	80	3.4	5.7	-2.3	5.29
30	77	1.1	3.8	-2.7	7.29
31	84	6.5	8.2	-1.7	3.00
32	78	4.4	4.4	0.0	0.00
33	sx	3.36		0.0	122.90
34	sy	2.96		n - 2	28
35	r	0.7182		SEE	2.10
36	Slope, b	0.6331	=SLOPE(B3:B32,A3:A32		
37	a	-44.95	=INTERCEPT(B3:B32,A3:A32)		
38	X-Bar	78.33			
39	Y-Bar	4.64			
40	SEE	2.10	=STEYX(B3:B32,A3:A32)		

Figure 29: Calculation of the SEE

Excel's STEYX function calculates the SEE in a few seconds. All it requires is the dependent variable range and the independent variable range. The bottom of Figure 28 shows the use of this function for our example in Cell B40 and the formula is displayed in Cell C40 and in Equation 18 shown below.

=STEYX(Y_variable_range,X_variable_range)

Equation 18: STEYX Syntax

VIII. Testing the Significance of the Slope of the Least Squares Line

There are other important significance tests that we should run. This is a test to determine whether the slope of the least squares line in the population is equal to zero. **Remember**: the slope in the sample is symbolized by the letter b. The slope in the population is symbolized by the Greek letter beta, β. Unfortunately, this symbol can be confused with a Type II error, which is also symbolized with β.

Here are the null and alternate hypotheses for this test: H_0: β = 0; H_1: β ≠ 0. When we fail to reject the null hypothesis, we are saying that the regression equation does not adequately predict Y. When we reject the null hypothesis, we conclude that the regression equation does predict Y.

We will use a 5 percent significance level. Our test statistic is a t-test with n – 2 degrees of freedom. We have 28 degrees of freedom. Just as in our test of the correlation, the critical values for our two-tailed test are -2.048 and +2.048. The rejection rule is: Reject the null hypothesis if t is less than -2.048 or greater than 2.048.

Unfortunately Excel does not have a built-in function for this test. But, as we shall see shortly, Excel's Regression tool will conduct this test in the blink of an eye. Without using the Regression tool, we can calculate the values of these test statistic in two steps: 1) Calculate the value of s_b, or the standard error of the estimated slope based on sample statistics (Equation 19), 2) Calculate the test statistic (Equation 20). Figure 30 shows how these calculations were performed in Excel along with the calculation of the p-value for this problem.

$$s_b = \frac{SEE}{s_X/\sqrt{n-1}} = \frac{2.10}{3.36/\sqrt{30-1}} = 0.12$$

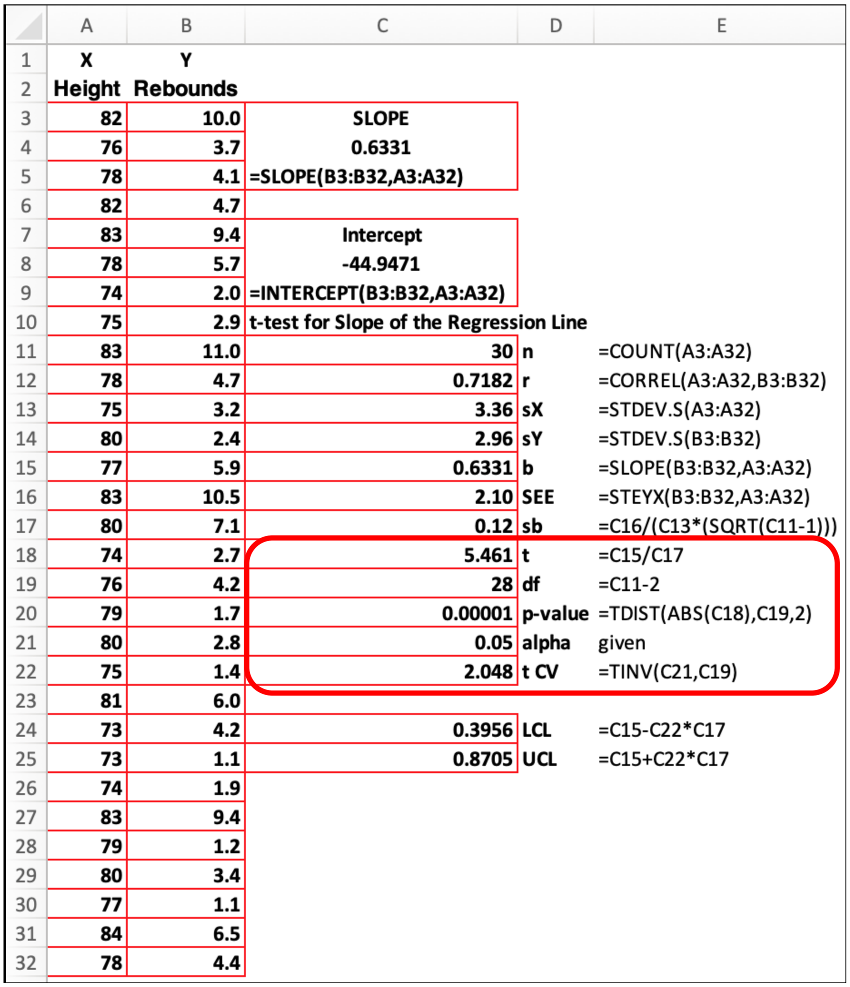
Equation 19: Standard Error of the Estimated Slope

Where: s_b = the standard error of the estimated slope based on sample statistics
s_X = is standard deviation of the independent variable
SEE = the standard error of the estimate
n = the number of paired observations

$$t = \frac{b}{s_b} = \frac{0.6331}{0.12} = 5.461$$

Equation 20: The Test Statistic for the Slope of the Least Squares Line

Where: s_b = the standard error of the estimated slope based on sample statistics
b = the slope of the least squares line based on sample statistics



	A	B	C	D	E
1	X	Y			
2	Height	Rebounds			
3	82	10.0	SLOPE		
4	76	3.7	0.6331		
5	78	4.1	=SLOPE(B3:B32,A3:A32)		
6	82	4.7			
7	83	9.4	Intercept		
8	78	5.7	-44.9471		
9	74	2.0	=INTERCEPT(B3:B32,A3:A32)		
10	75	2.9	t-test for Slope of the Regression Line		
11	83	11.0	30	n	=COUNT(A3:A32)
12	78	4.7	0.7182	r	=CORREL(A3:A32,B3:B32)
13	75	3.2	3.36	sX	=STDEV.S(A3:A32)
14	80	2.4	2.96	sY	=STDEV.S(B3:B32)
15	77	5.9	0.6331	b	=SLOPE(B3:B32,A3:A32)
16	83	10.5	2.10	SEE	=STEYX(B3:B32,A3:A32)
17	80	7.1	0.12	sb	=C16/(C13*(SQRT(C11-1)))
18	74	2.7	5.461	t	=C15/C17
19	76	4.2	28	df	=C11-2
20	79	1.7	0.00001	p-value	=TDIST(ABS(C18),C19,2)
21	80	2.8	0.05	alpha	given
22	75	1.4	2.048	t CV	=TINV(C21,C19)
23	81	6.0			
24	73	4.2	0.3956	LCL	=C15-C22*C17
25	73	1.1	0.8705	UCL	=C15+C22*C17
26	74	1.9			
27	83	9.4			
28	79	1.2			
29	80	3.4			
30	77	1.1			
31	84	6.5			
32	78	4.4			

Figure 30: Calculation of the Test Statistic and p-value for the Slope of the Regression Line

The value of our test statistic, 5.461, is very large while the p-value, 0.00001, is tiny. **Conclusion:** We have sufficient evidence to reject the Null Hypothesis. We conclude that our regression model predicts the value of the dependent variable.

IX. Confidence Intervals and Prediction Intervals

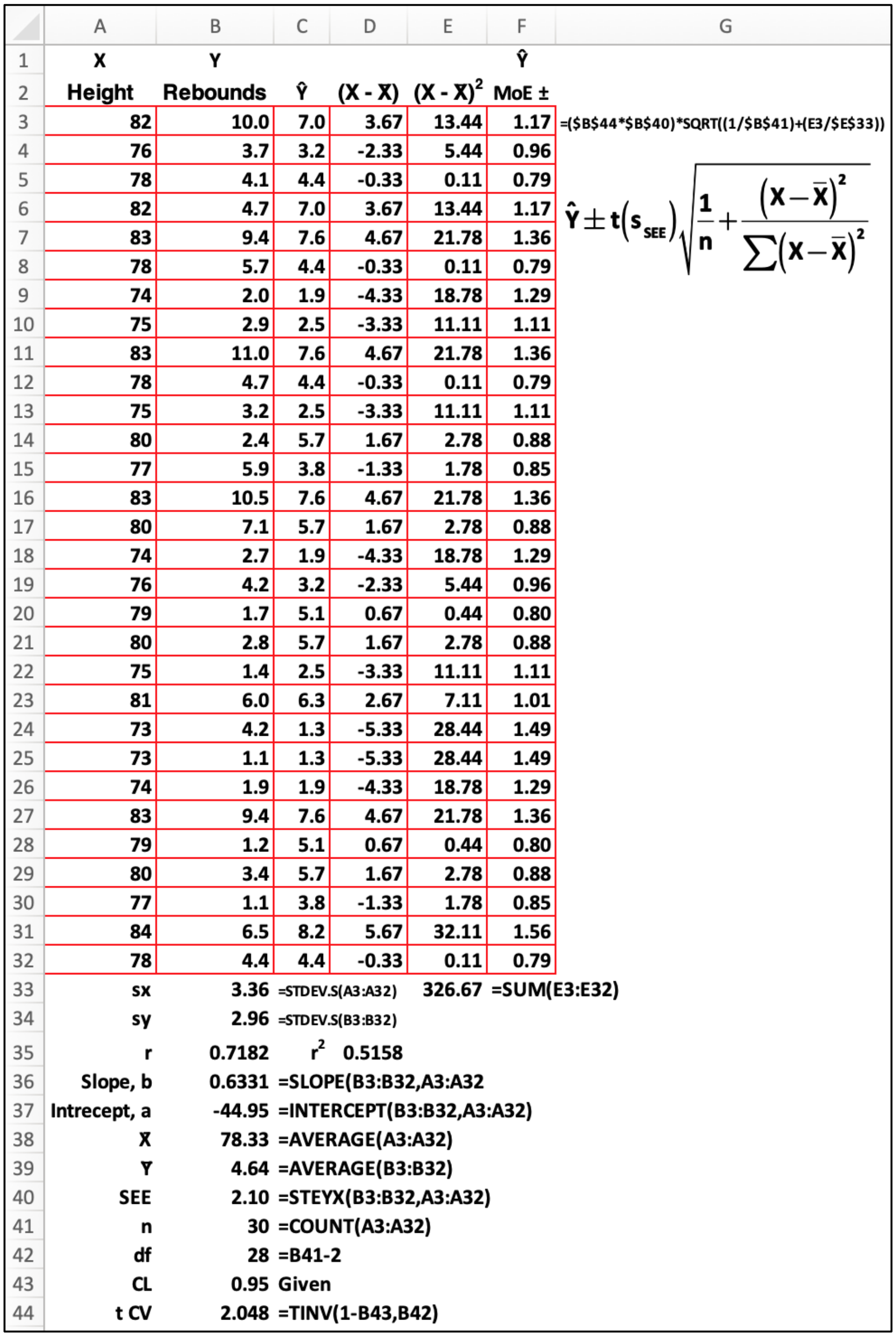
Because the regression equation is based on sample statistics that do not perfectly predict the value of the dependent variables, we set up intervals for the predicted value of the dependent variable, Ŷ. In fact, we need to create two types of intervals: 1) Confidence Intervals and 2) Prediction Intervals. Unfortunately, Excel does not have built-in functions for these calculations nor are they included in Excel's Regression Analysis. Even so, Excel can make short work of these calculations.

Confidence Intervals for Ŷ: Confidence intervals provide the margin of error (MoE) for Ŷ. The formula for constructing these confidence intervals is shown in Equation 21 and the calculations performed in Excel are shown in Figure 31.

$$\hat{Y} \pm t(\mathrm{SEE})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{(n - \bar{X})^2}}$$

Equation 21: Equation for the Confidence Interval of Ŷ

Where: Ŷ = Predicted value of Y
t = Critical value of t with n – 2 degrees of freedom
SEE = Standard Error of the Estimate



	A	B	C	D	E	F	G
1	X	Y				Ŷ	
2	Height	Rebounds	Ŷ	$(X - \bar{X})$	$(X - \bar{X})^2$	MoE ±	
3	82	10.0	7.0	3.67	13.44	1.17	=($B$44*$B$40)*SQRT((1/$B$41)+(E3/$E$33))
4	76	3.7	3.2	-2.33	5.44	0.96	
5	78	4.1	4.4	-0.33	0.11	0.79	
6	82	4.7	7.0	3.67	13.44	1.17	
7	83	9.4	7.6	4.67	21.78	1.36	
8	78	5.7	4.4	-0.33	0.11	0.79	
9	74	2.0	1.9	-4.33	18.78	1.29	
10	75	2.9	2.5	-3.33	11.11	1.11	
11	83	11.0	7.6	4.67	21.78	1.36	
12	78	4.7	4.4	-0.33	0.11	0.79	
13	75	3.2	2.5	-3.33	11.11	1.11	
14	80	2.4	5.7	1.67	2.78	0.88	
15	77	5.9	3.8	-1.33	1.78	0.85	
16	83	10.5	7.6	4.67	21.78	1.36	
17	80	7.1	5.7	1.67	2.78	0.88	
18	74	2.7	1.9	-4.33	18.78	1.29	
19	76	4.2	3.2	-2.33	5.44	0.96	
20	79	1.7	5.1	0.67	0.44	0.80	
21	80	2.8	5.7	1.67	2.78	0.88	
22	75	1.4	2.5	-3.33	11.11	1.11	
23	81	6.0	6.3	2.67	7.11	1.01	
24	73	4.2	1.3	-5.33	28.44	1.49	
25	73	1.1	1.3	-5.33	28.44	1.49	
26	74	1.9	1.9	-4.33	18.78	1.29	
27	83	9.4	7.6	4.67	21.78	1.36	
28	79	1.2	5.1	0.67	0.44	0.80	
29	80	3.4	5.7	1.67	2.78	0.88	
30	77	1.1	3.8	-1.33	1.78	0.85	
31	84	6.5	8.2	5.67	32.11	1.56	
32	78	4.4	4.4	-0.33	0.11	0.79	
33	sx	3.36	=STDEV.S(A3:A32)		326.67	=SUM(E3:E32)	
34	sy	2.96	=STDEV.S(B3:B32)				
35	r	0.7182	r^2	0.5158			
36	Slope, b	0.6331	=SLOPE(B3:B32,A3:A32				
37	Intrecept, a	-44.95	=INTERCEPT(B3:B32,A3:A32)				
38	$\bar{X}$	78.33	=AVERAGE(A3:A32)				
39	$\bar{Y}$	4.64	=AVERAGE(B3:B32)				
40	SEE	2.10	=STEYX(B3:B32,A3:A32)				
41	n	30	=COUNT(A3:A32)				
42	df	28	=B41-2				
43	CL	0.95	Given				
44	t CV	2.048	=TINV(1-B43,B42)				

$$\hat{Y} \pm t(s_{SEE})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

Figure 31: Confidence Interval for Ŷ

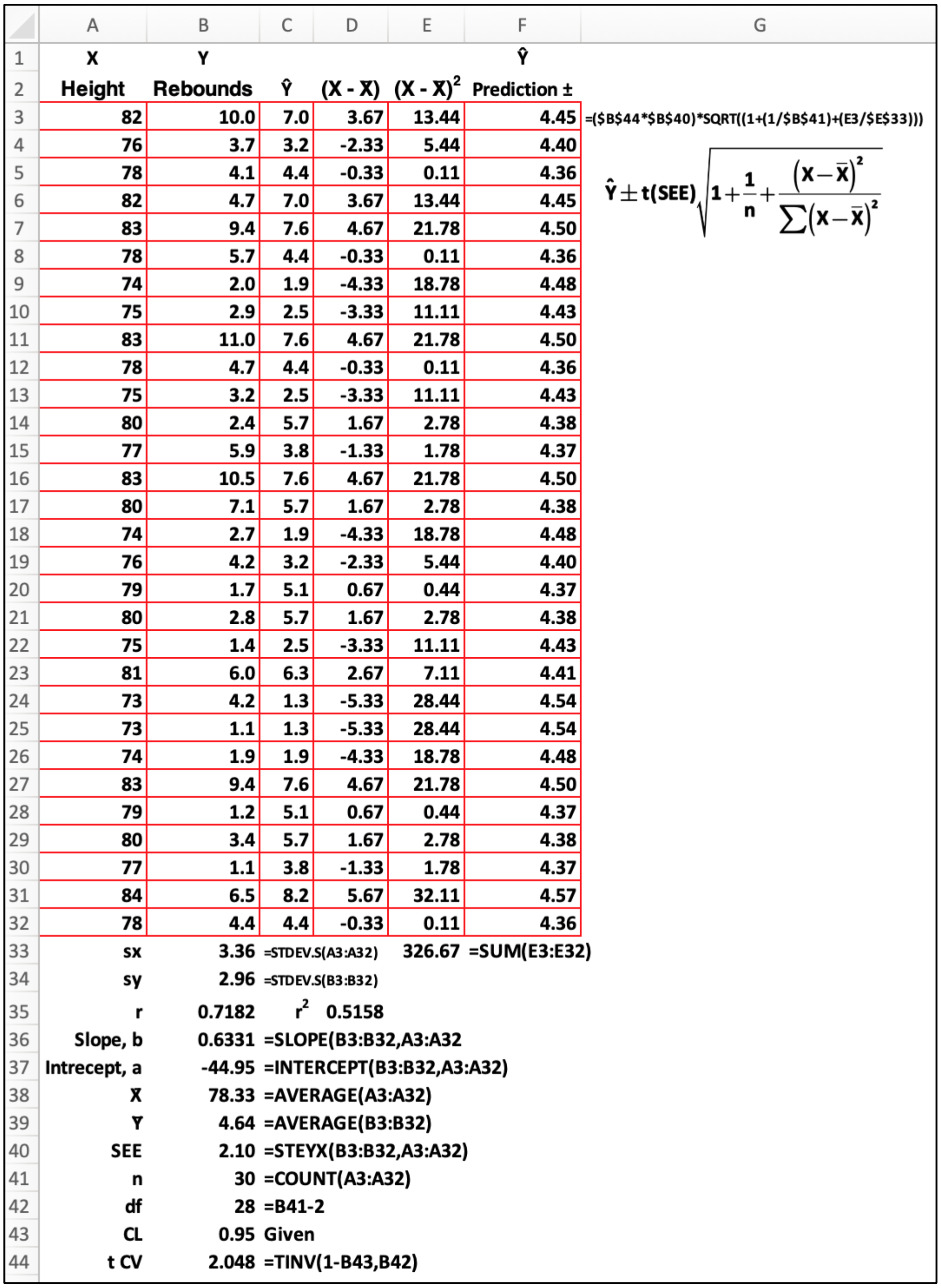
The first player on our list is 82” tall, there is a 95 percent probability of 10.0 average rebounds per game, plus or minus 1.17 rebounds. The lower confidence limit or LCL is 8.23 rebounds found by 10.00 – 1.17. The upper confidence limit is 11.17 rebounds found by 10.0 0 + 1.17.

Prediction Intervals for Ŷ: Prediction intervals provide the range of values that allows for random errors for a *future* observation. **Prediction intervals are wider than confidence intervals** because there is more variation for an individual element of the group than in the entire group. Equation 22 provides the formula for calculating prediction intervals for Ŷ. Figure 32 shows these calculations conducted in Excel.

$$\hat{Y} \pm t(\mathrm{SEE})\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{(n-\bar{X})^2}}$$

Equation 22: Equation for the Prediction Interval of $\hat{Y}$

Where: Ŷ = Predicted value of Y
t = Critical value of t with n – 2 degrees of freedom
SEE = Standard Error of the Estimate



	A	B	C	D	E	F	G
1	X	Y				$\hat{Y}$	
2	Height	Rebounds	$\hat{Y}$	$(X - \bar{X})$	$(X - \bar{X})^2$	Prediction ±	
3	82	10.0	7.0	3.67	13.44	4.45	=($B$44*$B$40)*SQRT((1+(1/$B$41)+(E3/$E$33)))
4	76	3.7	3.2	-2.33	5.44	4.40	
5	78	4.1	4.4	-0.33	0.11	4.36	
6	82	4.7	7.0	3.67	13.44	4.45	
7	83	9.4	7.6	4.67	21.78	4.50	
8	78	5.7	4.4	-0.33	0.11	4.36	
9	74	2.0	1.9	-4.33	18.78	4.48	
10	75	2.9	2.5	-3.33	11.11	4.43	
11	83	11.0	7.6	4.67	21.78	4.50	
12	78	4.7	4.4	-0.33	0.11	4.36	
13	75	3.2	2.5	-3.33	11.11	4.43	
14	80	2.4	5.7	1.67	2.78	4.38	
15	77	5.9	3.8	-1.33	1.78	4.37	
16	83	10.5	7.6	4.67	21.78	4.50	
17	80	7.1	5.7	1.67	2.78	4.38	
18	74	2.7	1.9	-4.33	18.78	4.48	
19	76	4.2	3.2	-2.33	5.44	4.40	
20	79	1.7	5.1	0.67	0.44	4.37	
21	80	2.8	5.7	1.67	2.78	4.38	
22	75	1.4	2.5	-3.33	11.11	4.43	
23	81	6.0	6.3	2.67	7.11	4.41	
24	73	4.2	1.3	-5.33	28.44	4.54	
25	73	1.1	1.3	-5.33	28.44	4.54	
26	74	1.9	1.9	-4.33	18.78	4.48	
27	83	9.4	7.6	4.67	21.78	4.50	
28	79	1.2	5.1	0.67	0.44	4.37	
29	80	3.4	5.7	1.67	2.78	4.38	
30	77	1.1	3.8	-1.33	1.78	4.37	
31	84	6.5	8.2	5.67	32.11	4.57	
32	78	4.4	4.4	-0.33	0.11	4.36	
33	sx	3.36	=STDEV.S(A3:A32)		326.67	=SUM(E3:E32)	
34	sy	2.96	=STDEV.S(B3:B32)				
35	r	0.7182	r^2	0.5158			
36	Slope, b	0.6331	=SLOPE(B3:B32,A3:A32				
37	Intrecept, a	-44.95	=INTERCEPT(B3:B32,A3:A32)				
38	$\bar{X}$	78.33	=AVERAGE(A3:A32)				
39	$\bar{Y}$	4.64	=AVERAGE(B3:B32)				
40	SEE	2.10	=STEYX(B3:B32,A3:A32)				
41	n	30	=COUNT(A3:A32)				
42	df	28	=B41-2				
43	CL	0.95	Given				
44	t CV	2.048	=TINV(1-B43,B42)				

$$\hat{Y} \pm t(SEE)\sqrt{1+\frac{1}{n}+\frac{\left(X-\bar{X}\right)^2}{\sum\left(X-\bar{X}\right)^2}}$$

Figure 32: Prediction Intervals for $\hat{Y}$

The prediction interval is in Column F in Figure 32. If we had a 82” tall new player we would expect him to have a 95 percent probability of averaging 7 rebounds per game, plus or minus 4.55 rebounds. The lower limit of the prediction would be 2.45 rebounds, found by 7.0 - 4.55. The upper limit would be 11.55 rebounds, found by 7.0 + 4.55.

X. Using Excel’s Regression Tool

Excel’s regression tools can save time performing regression analysis by:

- Calculating the coefficient of correlation.
- Calculating the coefficient of determination.
- Determining the slope of the least squares line.
- Testing the significance of the slope of this line and the Y-Intercept.
- Finding the Ŷ and the residuals.
- Determining whether the residuals approximate a normal distribution.

To launch Excel’s Regression tool, click on the Data Analysis icon located on the Data ribbon. Figure 33 shows what this icon looks like.

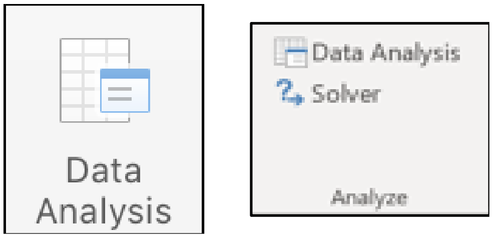


Figure 33: Data Analysis Icons

Once you click on this icon, the Analysis Tools selection window appears. Scroll down the list and select Regression and click OK. See Figure 34.

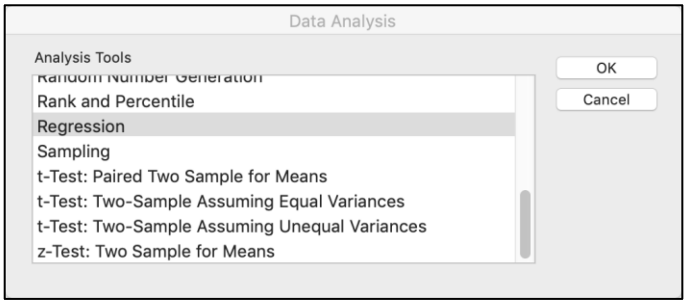
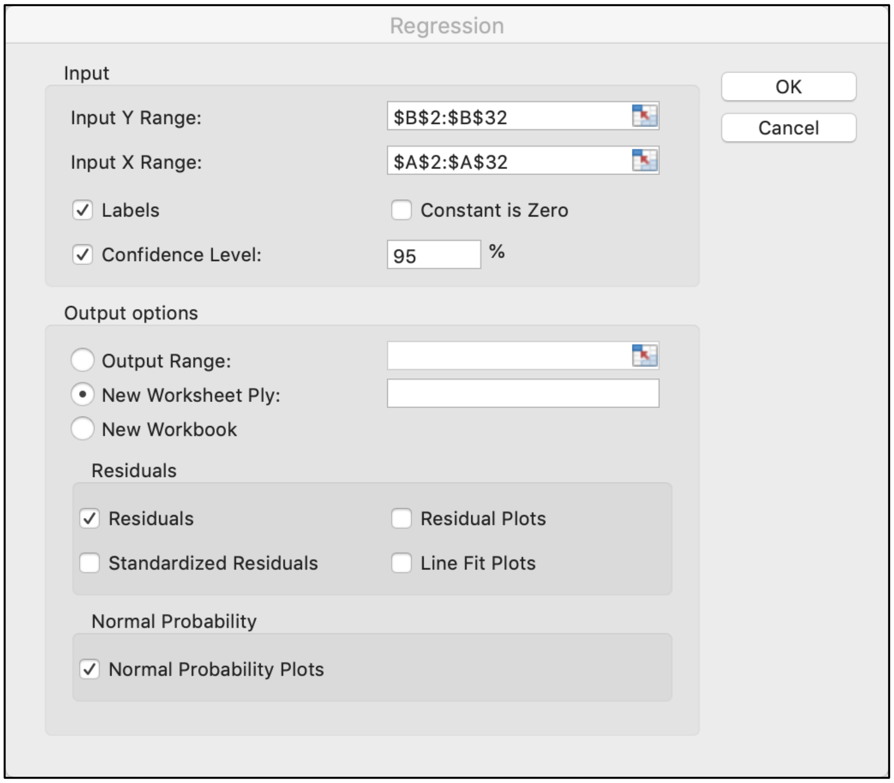

Data Analysis
Analysis Tools
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel

Figure 34: Analysis Tools Selection Window

Excel will now present the Regression option screen. See Figure 35. Under Input, enter the Y Range, $B$2:$B$32, and the X Range, $A$2:$A$32. You need not type these cell ranges. You can highlight the first cell in the workbook and drag the cursor to the last cell. Check the labels box because the names of the X and Y values are in the first cell of their respective ranges. Check the Confidence Level box. Excel defaults to a 95 percent confidence level, but you can enter different confidence levels. **Please note**: These confidence levels are for the slope of the regression line and the Y-Intercept, not the residuals.

Under Output options, you have three choices: 1) Output range, which places the output in a place on the workbook that you designate. Just enter one cell address and Excel will use that as the starting place for entering the analysis. Make certain that Excel will not write over cells with existing data. 2) Excel will place the analysis in a new worksheet. 3) Excel will place the analysis in a new workbook. Under Residuals check *residuals*, and Excel will calculate the residuals. Check Normal Probability Plots and Excel will create an XY chart of the Ŷ, which you can use to determine whether the residuals approximate a normal distribution. There are more rigorous tests for testing the normality of the residuals, like the Shapiro-Wilks test. The Shapiro-Wilks test is not easily conducted in Excel. More sophisticated statistical software, like SPSS, can run this and similar tests merely by checking a box.



Input

Input Y Range: $B$2:$B$32

Input X Range: $A$2:$A$32

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

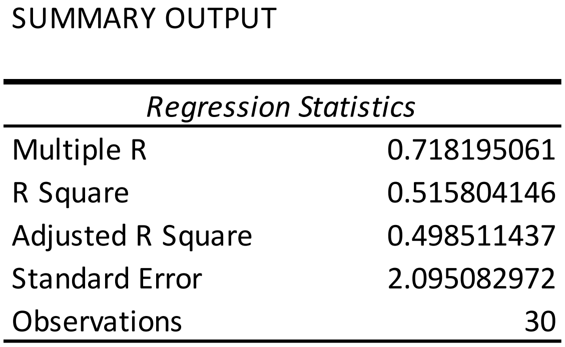
Normal Probability

☑ Normal Probability Plots

Figure 35: Regression Option Window

Excel quickly calculates the regression and correlation. statistics, an ANOVA table for the regression, t-tests for the Y-Intercept, which we will ignore, the Slope, the residuals for Ŷ, and the normal probability plot. Excel will also calculate confidence intervals for the Y-Intercept and slope of the regression line. Let's review the outputs of these analyses.

1) Summary Output - Regression Statistics: See Figure 36.



SUMMARY OUTPUT

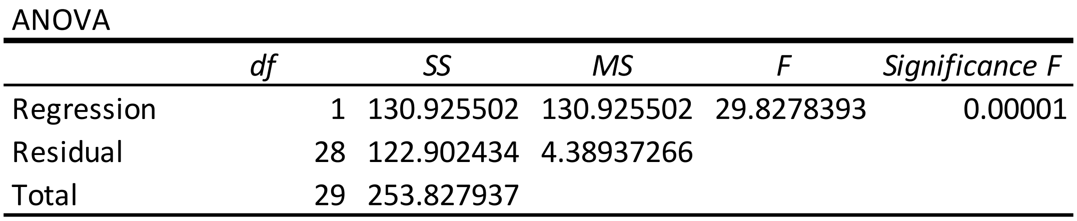
Regression Statistics	
Multiple R	0.718195061
R Square	0.515804146
Adjusted R Square	0.498511437
Standard Error	2.095082972
Observations	30

Figure 36: Summary Output - Regression Statistics

The labelling of these statistics is more appropriate for multiple regression than linear regression. The first line item is labelled Multiple R, but it is the coefficient of correlation, r. The second item is R Square. This is the coefficient of determination, r^2. Ignore Adjusted R Square because this item does not apply to linear regression. Standard Error is SEE. Observations is the number of paired observations. The analysis tool, despite its inappropriate labelling, provides more useful information than the Correlation tool. The Regression tool also saves us from entering the Excel functions: COUNT, CORREL, RSQ, STEYX, and TREND.

2) Excel's Regression ANOVA Table: See Figure 37.

This ANOVA table tests whether the slope of the regression line, β, equals zero. It also shows the relationship among the values on the ANOVA table and, the coefficient of determination and the Standard Error of the Estimate. The null and alternate hypotheses for this test are: H_0: β = 0; H_1: β ≠ 0.



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	130.925502	130.925502	29.8278393	0.00001
Residual	28	122.902434	4.38937266		
Total	29	253.827937			

Figure 37: ANOVA Table for Regression

Significance F is the p-value, which is very low. We have sufficient evidence to reject the Null Hypothesis that β equals zero. **Conclusion:** The independent variable, players' heights, predicts the dependent variable, average rebounds per game. See Equations 23 and 24.

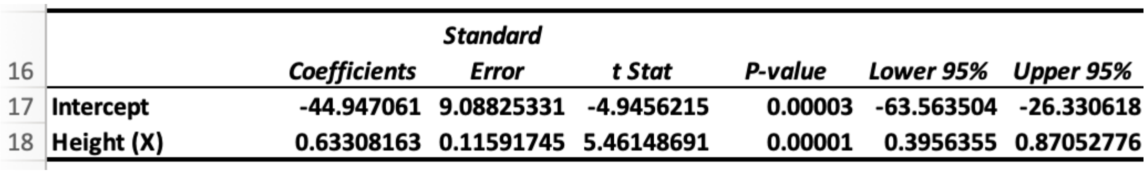
$$S\text{EE} = \sqrt{\text{Residual Mean Square}} = \sqrt{4.38937266} = 2.10$$

Equation 23: SEE = Square Took of Residual Mean Square

$$\text{r}^2 = \frac{\text{SS Regression}}{\text{SS Total}} = 1 - \frac{\text{SS Residual}}{\text{SS Total}} = \frac{112.9024345}{253.827937} = 0.5180$$

Equation 24: r^2 and the Regression ANOVA Table

3) Excel's t-tests: See Figure 38. Excel also provides two t-tests: One for the Y-Intercept the other for the Slope of the least squares line. Both of these tests have significant results. See Figure 38. The test for the slope of the least squares line is labelled "Height."

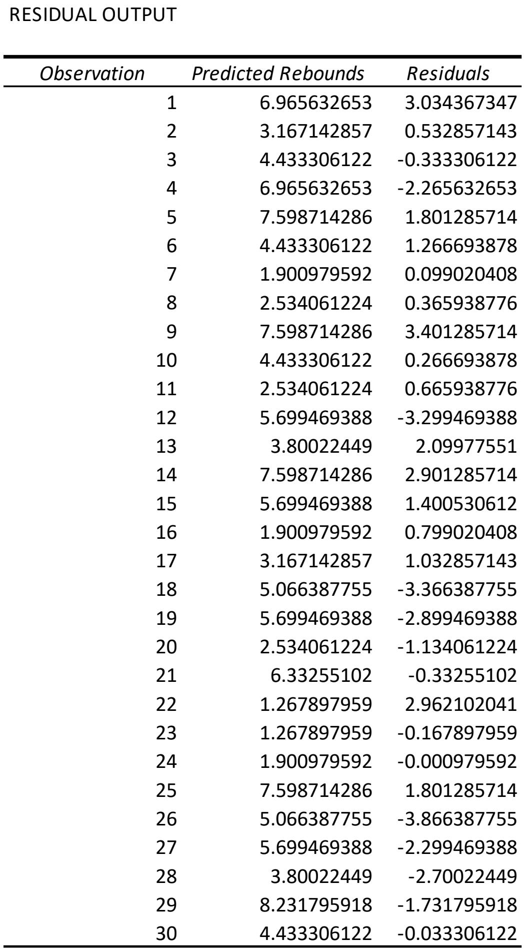


		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
16							
17	**Intercept**	**-44.947061**	**9.08825331**	**-4.9456215**	**0.00003**	**-63.563504**	**-26.330618**
18	**Height (X)**	**0.63308163**	**0.11591745**	**5.46148691**	**0.00001**	**0.3956355**	**0.87052776**

Figure 38: t-tests for the Y-Intercept and Slope of the Least Squares Line

Line 17, provides the Y-Intercept, -44.947061. Line 18, provides the slope of the line, 0.63308163. The value of the t-statistic is 5.46148691, and the p-value is 0.00001. **Please note:** The p-value cells had to be reformatted because the values are so small. This test provides evidence that the independent variable predicts the dependent variable.**:** In addition, Excel also reports the lower and upper limits for a 95 percent confidence interval for the Y-Intercept and slope of the regression line. The LCL is labelled Lower 95% and the UCL is labelled Upper 95%. You will recalled that confidence intervals are an inverse hypothesis test. Because zero is not included between the upper and lower confidence limits, we reject the Null Hypothesis. In the case of the slope of the regression line, height is a useful predictor of rebounds. Excel, for some unexplained reason, reports the Lower 95% and Upper 95% twice. The repetition of these values is not shown in Figure 38.

4) Excel Calculation of Ŷ and the Residual Output: See Figure 39. The Regression tool calculates estimated values of Y and the residuals. This saves us from using the TREND function:



RESIDUAL OUTPUT

Observation	*Predicted Rebounds*	*Residuals*
1	6.965632653	3.034367347
2	3.167142857	0.532857143
3	4.433306122	-0.333306122
4	6.965632653	-2.265632653
5	7.598714286	1.801285714
6	4.433306122	1.266693878
7	1.900979592	0.099020408
8	2.534061224	0.365938776
9	7.598714286	3.401285714
10	4.433306122	0.266693878
11	2.534061224	0.665938776
12	5.699469388	-3.299469388
13	3.80022449	2.09977551
14	7.598714286	2.901285714
15	5.699469388	1.400530612
16	1.900979592	0.799020408
17	3.167142857	1.032857143
18	5.066387755	-3.366387755
19	5.699469388	-2.899469388
20	2.534061224	-1.134061224
21	6.33255102	-0.33255102
22	1.267897959	2.962102041
23	1.267897959	-0.167897959
24	1.900979592	-0.000979592
25	7.598714286	1.801285714
26	5.066387755	-3.866387755
27	5.699469388	-2.299469388
28	3.80022449	-2.70022449
29	8.231795918	-1.731795918
30	4.433306122	-0.033306122

Figure 39: Residual Output

5) Normal Probability Plot: Excel plots the distribution of the residuals as an XY chart. This plot can be used to see if this distribution approximates a normal distribution. The straighter the line of dots, the more closely the distribution approximates a normal distribution. Based on the plot shown on Figure 40, we can conclude that this distribution approximates a normal distribution.

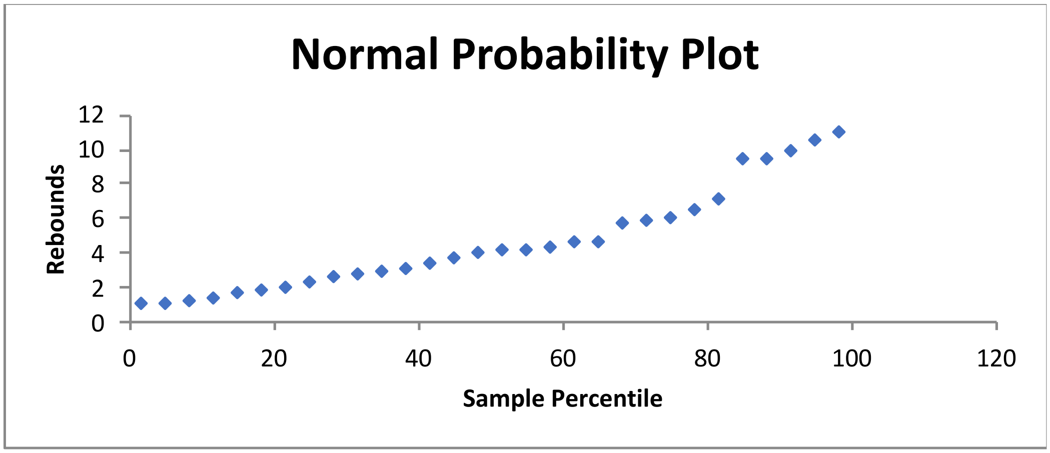


Figure 40: Normal Probability Plot

XI. Caveat Regarding Extrapolation

Extrapolation is using the regression model to predict Y-values from X-values that are *not* in the model. **This is a risky venture.** We can only predict the average number of rebounds per game for players between 73” and 82” tall. We cannot predict the rebounds for a player whose height is 72” (6’0”) or less, or 83” (7’1”) or more because we cannot make any inferences for values outside the range of our X-values.

XII. Correlation and Causation

For 120 years statisticians have warned that correlation is not synonymous with causation. Correlation is merely a measure of association. It does not prove a causal relationship even though correlation is a necessary but not sufficient condition to prove causation.

Karl Pearson, in his biography of his mentor Francis Galton, wrote, “Up to 1889 men of science had thought only in terms of causation, in the future they were to admit to another working category, that of correlation, and thus open to quantitative analysis wide fields of medical psychological, and sociological research.”[8]

The adoption of correlation was a major revolution in the history of science. Yet, causation has always been and remains a very important topic for both the physical and social sciences. As Judea Pearl, the artificial intelligence expert and recipient of the prestigious A. M. Turing Award, and his colleagues point out, "We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures…we also need to know *how* and *why* causes influence their effect."[9] Or, as Francis Bacon, who helped develop the scientific method over 400 years ago ,wrote, "Human knowledge and human power meet in one; *for where the cause is not known the effect cannot be produced* (italics added)."[10]

Despite Pearson and generations of statisticians who followed him, the physical and social sciences are focused on issues of causation and questions about causal inference remain a major concern. Questions about causation abound:

- Do statins lower LDL cholesterol and reduce the risk of heart disease?
- Does drinking sugar-sweetened carbonated soft drinks like Coca-Cola cause people to gain weight?
- Will doubling a brand's advertising spending lead to higher sales?
- Do mosquitoes spread West Nile virus?
- Does reducing the taxes on corporations and billionaires contribute to increased national debt?
- Does social distancing cause the spread of COVID-19 to be reduced?
- Will taking a new analgesic relieve a person's headache?
- Does a medical procedure have an unacceptable high risk of causing terrible side effects?
- Does the MMR vaccine cause autism?

Yet statisticians remain reluctant to mention causation. In fact, the typical introductory Statistics textbook mentions causation once, and this mention is the mantra that **correlation does not imply causation**.

How scientists establish causation, is something rarely mentioned in the standard introductory textbook. The gold standard for establishing causation are random controlled trials or tests (RCT). In such experiments, subjects are randomly assigned to either a control or treatment group. These subjects do not know whether they have been assigned to the treatment group, which is exposed to the stimuli, or to the control group, which has no exposure. In "double-blind" studies, the researchers and analysts also do not know to which group a subject has been assigned. The researchers then analyze the data to see if the effect is more prevalent in one of the studied groups. If it is, the researchers conclude that there is a causal effect. Sometimes RCTs are not possible, as the great Ronald Fisher wrote in his attack on researchers who argue that there is a causal link between smoking cigarettes and lung cancer.

> …randomization is totally impossible…in an inquiry of this kind. It is not the fault of [the researchers]… they cannot produce evidence in which a thousand children of teen age have been laid under a ban that they shall never smoke, and a thousand more chosen at random from the same age group have been under compulsion to smoke at least thirty cigarettes a day. If that type of experiment could have been done, there would be no difficulty [in establishing a causal link between smoking cigarettes and lung cancer]….No one feels—and especially a medical man could not feel—that it is right to do something to a human being which probably will do him harm.[11]

Medical ethics—the Hippocratic oath of "do no harm"—limits the ability of medical researchers to conduct RCT.

Historically, statisticians have been reluctant to discuss causation. Fisher a notable exception when he attacked the notion that smoking causes lung cancer. Fisher's argument merits discussion. In the 1950s, the accumulated evidence suggested that smoking cigarettes causes lung cancer. The *British Medical Journal* published a series of articles on the link between smoking cigarettes and lung cancer. In the summer of 1958, this prestigious journal published an editorial titled "Dangers of Cigarette-Smoking," which called for using "all the modern devices of publicity" to alert the public to the serious health risks of smoking.[12] This editorial prompted a harsh response from the acerbic Fisher. In his letter to the editor, Fisher wrote:

> Your annotation on "Dangers of Cigarette-smoking" leads up to the demand that these hazards "must be brought home to the public by all the modern devices of publicity". That is just what some of us with research interests are afraid of. In recent wars, for example, we have seen how unscrupulously the "modern devices of publicity" are liable to be used under the impulsion of fear; and surely the "yellow peril" of modern times is not the mild and soothing weed [tobacco] but the original creation of states of frantic alarm.

An important part of the mounting evidence of the causal link between smoking and cancer was established by Austin Bradford Hill and Richard Doll in their *observational* study published in the 1950s. By 1958, studies of patients in Scandinavia, the United States, Canada, Japan, and France, corroborated Hill and Doll's results: Cancer patients were more likely to be smokers than non-smokers.[13] These studies created a big stir, and were quickly accepted, as Fisher noted in his critique of Hill and Doll's work published in *The Centennial Review of Arts & Sciences*, an American peer-reviewed journal.. Nineteen investigations around the world have concurred with the findings of Hill and Doll, Fisher noted.[14] Yet Fisher argued that these studies "… were merely repetitions of evidence of the same kind, and it is necessary to try to examine whether that kind is sufficient for any scientific conclusion."[15] He points out that, among statisticians, skepticism abounds about the causal link between cigarettes and lung cancer. He even mentions that in his conversations with Hill, a fellow statistician, Hill was uncomfortable with the claim of causation. Fisher goes so far as to remind readers that correlation is not causation and to subtly suggest that what

Hill and Doll found may be nothing more than a spurious correlation; which is to say, mere coincidence without any real meaning.[16]

Fisher launched into a full-scale critique of Hill and Doll's research and the notion that there is a causal link between smoking cigarettes and lung cancer. Here is a summary of a few of his points:

1. The researchers failed to investigate properly the question of whether the lung cancer patients inhaled their cigarettes. Fisher argued that this is an important issue as pipe smokers and cigar smokers typically do not inhale and they seem to have fewer cases of lung cancer than cigarette smokers.

2. The fact that cigarette smokers have higher incidences of lung cancer than pipe and cigar smokers suggests that tobacco itself does not cause lung cancer. Fisher suggested that the cause might be cigarette paper or the fact that cigarette tobacco is not fermented to the same extent as the tobacco used in pipes and cigars.

3. Fisher flips the dependent and independent variables when he suggests that maybe lung cancer causes cigarette smoking. Fisher writes: "Is it possible then that lung cancer—that is to say, the pre-cancerous condition which must exist and is known to exist for years in those who are going to show overt lung cancer—is one of the causes of smoking cigarettes? I don't think it can be excluded."[17] He then added: "It [smoking] is the kind of comfort that might be a real solace to anyone in the fifteen years of approaching lung cancer. And to take that poor chap's cigarettes away from him would be rather like taking away his white stick from a blind man. It would make an already unhappy person a little more unhappy than he needs to be."[18]

Fisher argued that more research was needed into: 1) The role of inhaling cigarettes, and the "genotype" of lung cancer patients as well as that of smokers of cigarettes, pipes, and cigars.[19] Fisher adds that proving causation is complicated, which it is. Causation as a philosophical concept has been hotly debated issue in the work of philosophers from Aristotle in the fourth century BCE, David Hume in the seventeenth century and Bertrand Russel in the twentieth. Fisher's argument against the causal link between lung cancer and cigarettes is best summarized in his comment in *Nature*:

> The curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer.[20]

Today, virtually every scientist and statistician accepts the idea that the use of tobacco causes deadly illnesses and that observational studies like Hill and Doll's can prove causation.

In 1965, Austin Bradford Hill published nine criteria to prove causation.[21] Here they are:

1. **Strength of the Effect:** Small effects do not mean that there is no causal effect, though the larger the effect, the more likely there is a causal effect.

2. **Reproducibility:** Consistent findings observed by different researchers in different places with different samples strengthen the probability of a causal effect.

3. **Specificity:** Causation is likely when there is a very specific population at a specific site stricken with the disease with no other likely explanation.

4. **Temporality:** The effect must occur after the cause.

5. **Dose Responsiveness:** Greater exposure should generally lead to greater incidence of the effect and the effect reduces upon a reduction of exposure.

6. **Plausibility:** Biological, chemical, or mechanical evidence for a "causal chain."

7. **Coherence:** The effect fits with established knowledge.

8. **Experiment:** The effect can be replicated with experiments.

9. **Analogy:** Similarities between observed associations.

Establishing causation is not easy. A causal link must meet Hill's nine criteria. In particular, the cause must precede the effect. There must also be a concomitant variation between the cause and effect. A change in the cause must lead to a change in the effect. Causality, however, is not deterministic. Not every smoker gets lung cancer. Nonsmokers also get lung cancer, but at a much lower rate. **Causation, therefore, is probabilistic, which means that the effect is more likely to happen when the cause is present.**

David Spiegelhalter, a past president of the Royal Statistical Society, writes that scholars like Judea Pearl and others are making progress in laying out the principles of causal regression models based on observational data.[22] Pearl and his co-writer, Dana MacKenzie, lay out the basis for this causal revolution for a popular audience in the *Book of Why: The New Science of Cause and Effect*. Pearl argues that the calculus of causation is based on two languages: 1) Causal diagrams to express what we know, and 2) Symbolic language to express what we want to know.[23] Pearl's innovations in causal inference, however, are beyond the scope of an introductory textbook.

Spurious Correlations: While the great statisticians of the late nineteenth and twentieth centuries avoided the issue of causation, they did speak of spurious correlations usually in the context of warning that correlation does not mean causation. Fisher, in fact, did this in "Cigarettes, Cancer, and Statistics" when he mentioned statistician George Udny Yules' favorite example of a spurious correlation: The positive correlation between increases in imported apples in the United Kingdom and the rise of the divorce rate.[24]

Karl Pearson was the first to mention spurious correlations in 1897.[25] It is a term that refers to a statistically significant correlation between two variables due to mere coincidence. They either have no genuine relationship, or a third unseen variable called a confounding or lurking variable is present and explains, if not causes, the two correlated variables. A spurious correlation may also arise when the X and Y variables are not independent. The perfect positive and negative correlations cited above may, in fact, be spurious correlations because they are essentially non-independent measures.

Here is a popular example of a spurious correlation commonly used by professors teaching introductory statistics classes: As ice cream sales rise so do the number of times lifeguards have to rescue swimmers from drowning. Do increased sales of Ben & Jerry's, Häagen-Dazs, and Breyers ice cream cause people to drown? Or, when the number of drownings goes up cause people to run out and buy ice cream? The professor typically starts the lesson by telling students that correlation is not synonymous with causation. This is because spurious correlations are usually infused with a presumption of causation. Then the professor asks, which of these two events is the independent variable? Most students answer that ice cream sales are the independent variable. Whatever students answer, however, the professor will say they are wrong. Why? Well, there is a *confounding variable* that explains, if not causes, both increased ice cream sales and more lifeguard rescues: Hot weather. Summer weather is associated with more people swimming (and a greater need for lifeguards) and higher ice cream sales. Except for the members of the Coney Island Polar Bear Club, nobody goes to the beach to swim during cold winter weather. Ice cream sales also have a positive correlation with forest fires and shark attacks. These are also spurious correlations: Hot summer weather is the confounding variable. It dries out forests and results in forest fires and also warms oceans causing sharks to migrate to cooler northern waters where they are more likely to encounter summer bathers.

There are thousands of delightfully foolish examples of spurious correlations that do not have a confounding variable. Tyler Vigen's book and website, *Spurious Correlations*, has many hilarious examples. One of my favorites is the 66.6 percent correlation between the number of films Nicolas Cage appears in during a year, and the number of people who fall into swimming pools and drown. I often wonder what a multiple correlation would show if a second independent variable—ice cream sales—were added.

Here are a few more famous spurious correlations:

1. **The hemline theory of the stock market:** As hemlines on women's skirts rise, the stock market goes up. When hemlines descend, the stock market goes down. This spurious correlation goes back to the 1920s and is said to have a very high coefficient of correlation.

2. **The winner of the Super Bowl and the stock market:** If the winning team is a member of the American Football Conference, the stock market will go down and if the winning team is in the National Football Conference, the stock market will go up.

3. **U.S. Murder Rate and Microsoft Internet Explorer Usage:** From 2006 to 2011, both the murder rate and the usage of Microsoft's internet browser declined rate.

4. **Storks and Dutch (German or Danish) babies:** The number of storks nesting on the rooftops of Dutch houses is said to be positively correlated with the number of children living in those houses. By the way, no reasonable person believes in the old folklore that storks deliver babies. There is a confounding variable: Large houses tend to attract both big families and storks seeking a nesting place. It is interesting to note that Darrell Huff (1913 – 2001), the author of the best-selling statistics book of the twentieth-century, *How to Lie With Statistics*, used this amusing spurious correlation when he testified before Congress in the 1950s and 1960s to ridicule the notion that there is a causal link between tobacco and cancer.[26] Was Mr. Duff being a fool or a knave when he worked for the tobacco industry as an expert witness who disputed the research that showed a causal link with their product and other deadly illnesses?

Not all spurious correlations are harmless tomfoolery. Some are the work of knaves and misguided fools. Here are examples from the website, Statistics How to[27].

1. Universal health care breeds terrorism (Fox News)

2. Living next to freeways causes autism (*L. A. Times*)

3. Junk food does not cause obesity (Global Energy Balance Network). Okay, this is a spurious non-correlation. The GEBN is a non-profit organization that claims to fund research into obesity. According to *The New York Times* this organization has received funding from Coca-Cola.[28]

4. Fox News makes viewers stupid, or at least misinformed (World Public Opinion, a project managed by the Program on International Policy Attitudes at the University of Maryland). I will let you decide whether this is a spurious correlation.

One of the more nefarious spurious correlations to take hold is the notion that the MMR vaccination causes autism. The MMR vaccine was first licensed for use in 1971. This vaccine protects against measles, mumps, and rubella, which is also called German measles.

According to the Infectious Disease Society of America (IDSA), a community of over 12,000 physicians, scientists, and public health researchers, prior to the MMR vaccine, measles, mumps, and rubella, sickened 3 to 4 million people a year and annually lead to around 500 deaths and 48,000 hospitalization. Nearly every American child got the measles before age sixteen.[29] Since the vaccine's introduction, over 575 million doses have been given worldwide. It has an excellent safety record.[30]

Autism Speaks, the largest autism advocacy group in the U.S., defines autism or ASD as "…a broad range of conditions characterized by challenges with social skills, repetitive behaviors, speech and nonverbal communication. According to the Centers for Disease Control, autism now affects an estimated 1 in 59 children in the United States today."[31] How did this life-saving vaccine come to be viewed as a causal link to autism?

The anti-vaccination movement has a long history. It had been dormant until 1998, when the highly respected British medical journal, *Lancet*, published an article by Dr. Andrew J. Wakefield, and twelve other medical researchers. This article was billed as an "early report" because it reported preliminary findings of a study of *only twelve children*. It raised the possibility that the MMR vaccine might cause autism, but the authors clearly stated that they did not prove an association, let alone a causal link, between the vaccine and autism.[32] The flaws in this research were readily apparent. There was no control group and the sample size was tiny. It is never a good sign when the number of authors of a study exceed the number of subjects.

After the publication of Wakefield et al.'s article, the media picked up this story. Vaccination panic erupted. Soon, celebrities like Jenny McCarthy, a mother of an autistic child, who used their access to mass media to spread fear and distrust of vaccinations through appearances on talk shows, books, and social media.[33] This anti-vaccination campaign contributed to a drop in the rate of vaccinations and a resurgence of measles. In 2019, the World Health Organization listed *vaccine hesitancy* as one of the top ten threats to global health.[34]

Numerous studies have failed to show a causal link between the MMR vaccine and autism. A study tracking 650,000 Danish children found that the MMR vaccine "…does not increase the risk for autism, does not trigger autism in susceptible children, and is not associated with clustering of autism cases after vaccination."[35]

In 2010, Lancet retracted Dr. Wakefield's article, and *The British Medical Journal* published a series of articles by journalist Brian Deer, who exposed the fact that Dr. Wakefield fraudulently manufactured the data used in the study and had financial conflicts of interest[36]. The British General Medical Council, the public body that licenses physicians in the United Kingdom, found Dr. Wakefield guilty of three-dozen charges, including dishonesty and abuse of children. The doctor lost his medical license in the United Kingdom.[37] While he maintains his innocence, his reputation is in tatters, as a 2011 article in the *Sunday New York Times Magazine* noted:

> Andrew Wakefield has become one of the most reviled doctors of his generation, blamed directly and indirectly, depending on the accuser, for

> irresponsibly starting a panic with tragic repercussions: vaccination rates so low that childhood diseases once all but eradicated here—whooping cough and measles, among them—have re-emerged, endangering young lives.[38]

Dr. Wakefield now lives in Texas and has become a prominent antivaxxer.[39]

Why the link between the MMR vaccination and autism is a spurious correlation. Fact: The rate of MMR vaccinations has increased since its introduction in the early 1970s. **Fact:** The incidence of autism has also grown sharply. On the surface this would seem to suggest an important link between vaccination and autism. But, the increased rate of autism may be due to changes in the diagnostic criteria for this malady. A further confounding factor is that autism becomes apparent around the time a child would have received his or her first MMR vaccination**. Please note:** A recently published meta-analysis of 54 studies with 13,784,284 participants shows that the male-to-female ratio in autism is 3 to 1.[40]

Autism Speaks says there are genetic and environmental factors for autism. The environmental risks include the advanced age of either parent, multiple births (twins and triplets), multiple pregnancies spaced less than one year apart, premature births, and low birth weights. The MMR vaccine, they say, does not cause autism.[41]

XIII. Summary

In this module we have:

- Defined correlation as the strength of the association between two quantitative variables: The independent or predictor variable, X, and the dependent or response variable, Y.

- Defined and interpreted the coefficient of correlation, r, and calculated it by hand and with Microsoft Excel.

- Defined and interpreted the coefficient of determination, r^2, and calculated it by hand and with Microsoft Excel.

- Tested the significance of the correlation.

- Defined regression as a model used to predict the value of Y based on a value of X.

- Conducted a regression analysis by hand and with Microsoft Excel.

- Created the least squares line.

- Generated confidence intervals and prediction intervals for $\hat{Y}$, the estimated value of the dependent variable.

- Tested the significance of the least squares line.

- Distinguished correlation from causation.
- Discussed ways researchers establish causal relationships.

- Described the problem of spurious correlations.

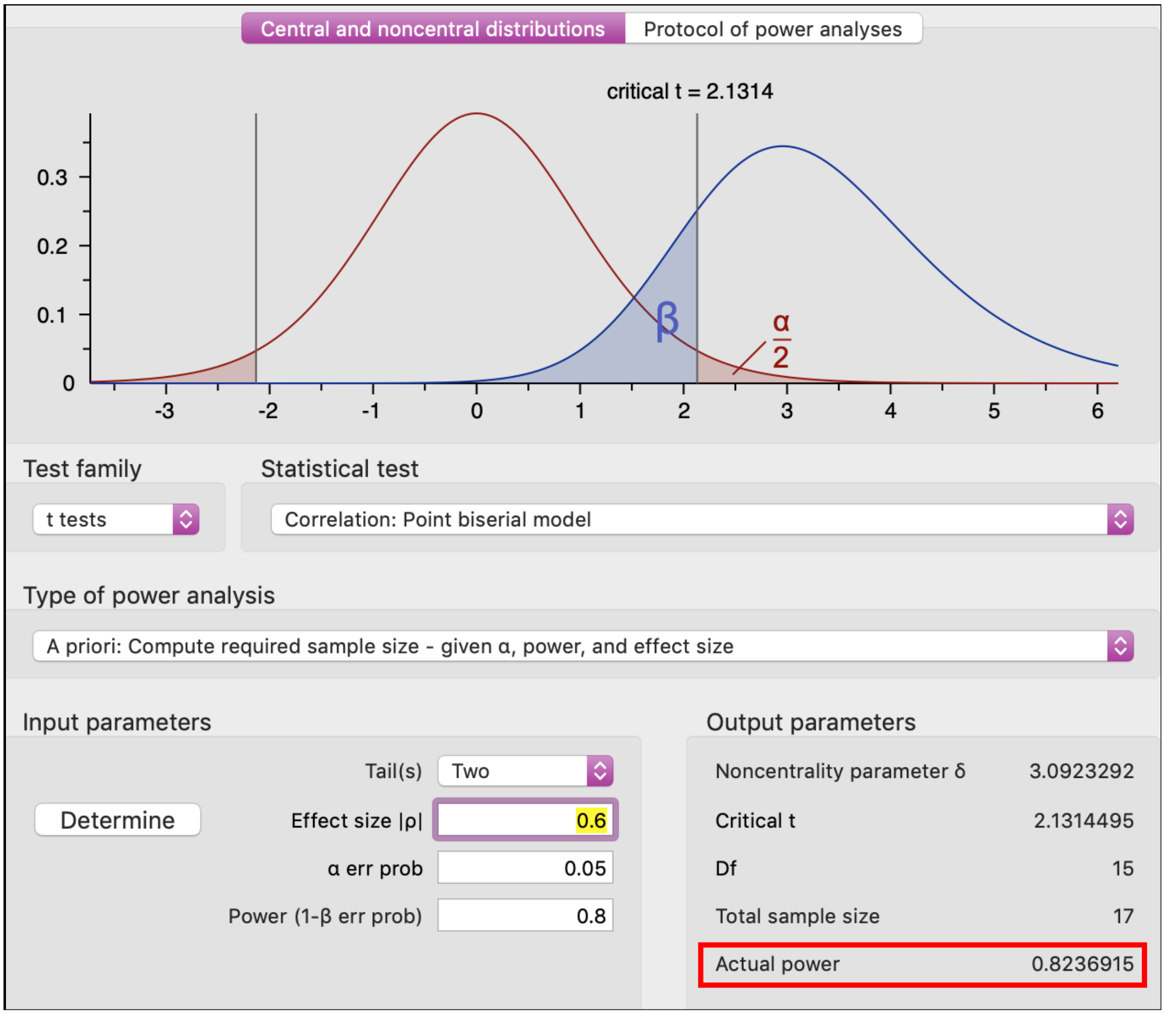
Correlation and regression are some of the most important techniques in inferential statistics. There are, however, other more sophisticated types of regression that are covered in advanced statistics courses. As we discussed, there is multiple regression, which uses two or more independent variables to predict the dependent variable. Most serious investigations based on linear regression models rely on multiple regression and not simple linear regression with only one independent variable. There are also nonparametric regression models like Spearman's Rho and Kendall's Tau. Logistic regression models analyze multiple independent variables and a categorical or qualitative dependent variable. For non-linear data there is polynomial regression. There is also structural equation modeling, which uses mathematical and computer algorithms to construct causal models.

XIV. Exercises

Data for these exercises is in 18_Exercises.xlsx. Conduct a priori power analyses using G*Power.

Exercise 1: Student Absences and Grades on the Final Exam

Dr. V. noticed that the more frequently a student is late or absent from class the worse he or she performs on the final exam. He decided to investigate. The first thing he did was to conduct an *a priori* power analysis using G*Power to determine the necessary sample size. Based on his experience, he estimated the correlation coefficient as -0.60. He entered 0.6 in effect size |ρ|. He set the α err prob at 0.05 and the Power (1 - β err prob) at 0.8. The analysis, depicted in Figure 41, shows that a sample size of 17 students would yield 82.24 percent power. What is the sample size needed to achieve 80% power.


Central and noncentral distributions
Protocol of power analyses
critical t = 2.1314
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
6
β
α
2
Determine

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s)	Two
Effect size |ρ|	0.6
α err prob	0.05
Power (1-β err prob)	0.8

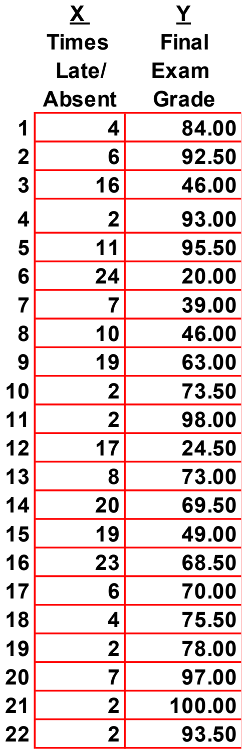
Output parameters

Noncentrality parameter δ	3.0923292
Critical t	2.1314495
Df	15
Total sample size	17
Actual power	0.8236915

Figure 41: A Priori Power Analysis

Not being certain of the exact effect size, Dr. V. collected a random sample of 22 students. This sample includes the number of times a student is absent and their grades on the final exam. The data can be found in the Excel file, 18_Exercises. Look for the worksheets with titles that start with “Exercise1_.”

Figure 42 shows the results of Dr. V.’s sample:



	X Times Late/ Absent	Y Final Exam Grade
1	4	84.00
2	6	92.50
3	16	46.00
4	2	93.00
5	11	95.50
6	24	20.00
7	7	39.00
8	10	46.00
9	19	63.00
10	2	73.50
11	2	98.00
12	17	24.50
13	8	73.00
14	20	69.50
15	19	49.00
16	23	68.50
17	6	70.00
18	4	75.50
19	2	78.00
20	7	97.00
21	2	100.00
22	2	93.50

Figure 42: Dr. V.'s Sample Data

Question 1: Which variable is the independent variable and which is the dependent variable?

Question 2: Using Microsoft Excel, construct a scatter diagram (XY chart). Include the least squares line.

Question 3: Using Microsoft Excel:

a. Count the number of variables
b. Calculate the mean and standard deviation of the independent variable
c. Calculate the mean and standard deviation of the dependent variable
d. Calculate the correlation coefficient, r, and interpret what it means
e. Calculate the coefficient of determination, r^2, and interpret what it means

Question 4: Using Excel, conduct a NHST to determine whether there is a correlation in the population. Use a 0.05 significance level**.**

Question 5: Using Excel, calculate the Slope of the line (b), the Y-Intercept (a), the Standard Error of the Estimate (SEE), and test the hypothesis that the slope of the line equals zero.

Question 6: Using Excel's TREND Function the calculate predicted values for Y (Ŷ) and their Confidence Intervals.

Question 7: Using Excel's TREND Function the calculate predicted values for Y and their Prediction Intervals.

Question 8: Using Excel's Regression Tool: a) Report the Regression Statistics, b) Interpret the ANOVA table, c) Interpret the t-test for the slope of the regression line, d) Interpret the residual output, and e) Interpret the Normal Probability Plot.

a) Regression Statistics:

b) ANOVA Table:

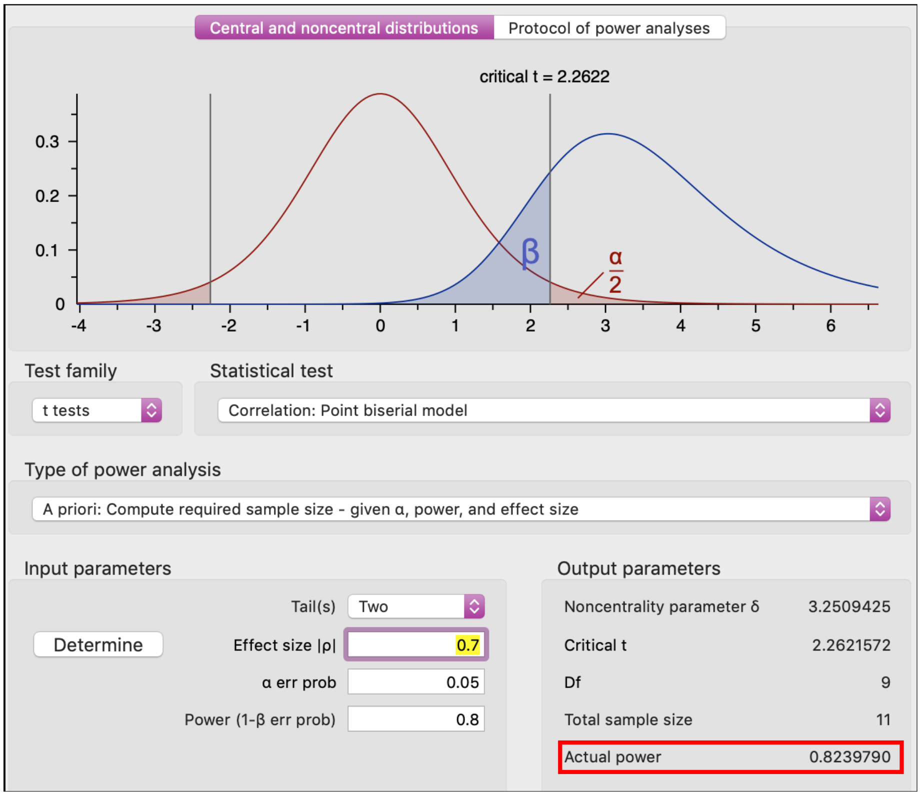
c) t-Tests:

d) Residual Output:

e) Normal Probability Plot:

Exercise 2: Used Cars: Mileage and Retail Selling Price

Terry is a recent college graduate who loves to fix up used cars. She thinks she can turn this hobby into a business. She has recently purchased a late model Toyota Camray, fixed it up, and sold it for a tidy profit. When selling this car, she noticed a large inverse correlation between cars' mileage and their asking price. She has now purchased a used, four-door Honda Accord LX. She has taken a survey of late model four-door Honda Accord LX from cars.com. Based on her experience with the Toyota, she anticipates a strong negative correlation between the car’s asking price mileage. Before conducting her survey, Terry conducted an *a priori* power test using G*Power to determine the size of her sample. She estimates the correlation coefficient as -0.70. She entered 0.7 in Effect size |ρ|. She set the α err prob at 0.50 and the Power (1- β err prob) at 0.8. The analysis in Figure 43, shows that a sample size of 11 cars would yield 82.24 percent power.


Central and noncentral distributions
Protocol of power analyses
critical t = 2.2622
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
5
6
β
α
2
Determine

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s)	Two
Effect size |ρ|	0.7
α err prob	0.05
Power (1-β err prob)	0.8

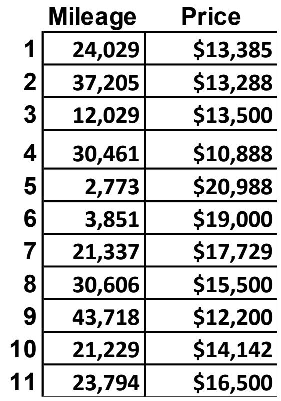
Output parameters

Noncentrality parameter δ	3.2509425
Critical t	2.2621572
Df	9
Total sample size	11
Actual power	0.8239790

Figure 43: A Priori Power Analysis

The data can be found in the Excel file, 18_Exercises. Look for the worksheets with titles that start with “Exercise2_.”

Figure 44 shows the results of Terry’s survey:



	Mileage	Price
1	24,029	$13,385
2	37,205	$13,288
3	12,029	$13,500
4	30,461	$10,888
5	2,773	$20,988
6	3,851	$19,000
7	21,337	$17,729
8	30,606	$15,500
9	43,718	$12,200
10	21,229	$14,142
11	23,794	$16,500

Figure 44: Terry's Sample - Mileage and Asking Price

Question 1: Which variable is the independent variable and which is the dependent.

Question 2: Using Microsoft Excel, construct a scatter diagram (XY chart). Include the least squares line.

Question 3: Using Microsoft Excel:

a. Count the number of variables
b. Calculate the correlation coefficient, r, and interpret what it means
c. Calculate the coefficient of determination, r^2, and interpret what it means
d. Calculate the mean and standard deviation of the independent variable
e. Calculate the mean and standard deviation of the dependent variable

Question 4: Using Excel, conduct a NHST to determine whether there is a correlation in the population. Use a 0.05 significance level.

Question 5: Using Excel calculate the Slope of the line (b), the Y-Intercept, the Standard Error of the Estimate, and test the hypothesis that the slope of the line equals zero.

Question 6: Using Excel's TREND Function, calculate predict values for Y and their Confidence Intervals.

Question 7: Using Excel's Regression Tool: a) report the Regression Statistics, b) interpret the ANOVA table, c) interpret the t-test for the slope of the regression line, d) interpret the residual output, and e) interpret the Normal Probability Plot.

a) Regression Statistics:

b) ANOVA Table:

c) t-Tests:

d) Residual Output:

e) Normal Probability Plot:

* * *

[1] Edward Tufte, *The Cognitive Style of Powerpoint*, (Cheshire, CT, Graphics Press, 2006), p. 5.

[2] George Udny Yule and Maurice G. Kendall, *An Introduction to the Theory of Statistics,* 14th Edition 5th Impressio*n*, (London, UK: Charles Griffin & Company, 1968), p. 213. The first edition of this classic was published in 1911.

[3] George E. P. Box and Norman R. Draper, *Empirical Model-Building and Response Surfaces*," (New York: John Wiley & Sons, 1987), p. 425.

[4] George E. P. Box and Norman R. Draper, *Empirical Model-Building and Response Surfaces*," (New York: John Wiley & Sons, 1987), p. 74.

[5] Brian E. Clauser, "The Life and Labors of Francis Galton: A Review of Four Recent Books About the Father of Behavioral Statistics," *Journal of Educational and Behavioral Statistics*, Vol. 32, No. 4. December 1, 2007, pp. 440-444. Michael Bulmer, "Galton's Law of Ancestral Heredity," *Heredity*, Vol. 81, No. 5 1998. pp. 579-585.

[6] Karl Pearson, *Francis Galton: A Centenary Appreciation*, (Cambridge, UK: Cambridge University Press, 1922). Karl Pearson, *The Life Letters and Labors of Francis Galton*, (Cambridge, UK: Cambridge University Press, 1930).

[7] David Skinner, "The Age of Female Computers," *The New Atlantis: A Journal of technology and Society*, Number 12, Spring 2006, p. 97.

[8] Karl Pearson, *The Life, Letters, and Labours of Francis Galton, Vol. IIIA*, (Cambridge, UK, Cambridge University Press, 1930), p. 1.

[9] Judea Pearl, Madelyn Glymour, and Nicholas p. Jewell, *Statistical Inference in Statistics: A Primer*, (West Sussex, UK: John Wiley & Sons, 2016), p. 1.

[10] Francis Bacon, *Novum Organum: Aphorisms Concerning the Interpretation of Nature and the Kingdom of Man*, III, (New York: P. F. Collier, 1902), p. 11.

[11] Ronald A. Fisher, Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, pp. 155-156.

"Dangers of Cigarette-Smoking," *The British Medical Journal*, June 20, 1958, p. 1518.

[12] Ronald A. Fisher, "Alleged Dangers of Cigarette-Smoking," *The British Medical Journal*, Volume II, June 29, 1958, p. 269.

[13] David Salsburg: *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, (New York: Henry Holt and Company, 2001), p. 158.

[14] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, p. 151.

[15] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, p. 151.

[16] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, pp. 154-155.

[17] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, p. 162.

[18] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, pp. 162-3.

[19] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, p. 163.

[20] Ronald A. Fisher, "Cancer and Smoking," *Nature*, Vol. 182, No. 596, August 30, 1958.

[21] Austin Bradford Hill, "The Environment and Disease: Association or Causation?" Proceedings of the Royal Society of Medicine, Vol. 58, No. 5, 1965. https://doi.org/10.1177/003591576505800503.

[22] David Spiegelhalter, *The Art of Statistics: How to Learn from Data*, (New York: Basic Books, 2019), p. 128.

[23] Judea Pearl and Dana MacKenzie, *The Book of Why: The New Science of Cause and Effect*, (New York: Basic Books, 2018), p. 7.

[24] Ronald A. Fisher, "Cigarettes, Cancer, and Statistics," *The Centennial Review of Arts & Sciences*, Vol. 2, 1958, p. 154.

[25] Karl Pearson, "On a Form of Spurious Correlation Which May Arise when Indices are Used in the Measurement of Organs," *Proceedings of the Royal Society of London*. Vol. 60. Issue 359-367, January 1, 1897, pp. 489-498.

[26] Andrew Gelman, "Statistics for Cigarette Sellers," *Chance*, Vol. 25.3, 2013, p. 43.

[27] "What is a Spurious Correlation?" *Statistics How To: Statistics For the Rest of Us!* https://www.statisticshowto.datasciencecentral.com/spurious-correlation/

[28] Anahad O'Connor, "Coca-Cola Funds Scientists Who Shift Blame for Obesity Away From Bad Diets," The *New York Times*, August 9, 2015. https://well.blogs.nytimes.com/2015/08/09/coca-cola-funds-scientists-who-shift-blame-for-obesity-away-from-bad-diets/?ref=business&_r=0.

[29] "Measles Vaccination: Myths and Facts," *Infectious Diseases Society of America (IDSA),* https://www.idsociety.org/public-health/measles/myths-and-facts/.

[30] "Addressing Misconceptions on Measles Vaccination," *The European Centre for Disease Prevention and Control*. https://www.ecdc.europa.eu/en/measles/prevention-and-control/addressing-misconceptions-measles.

[31] "What is Autism? There is Not One Type of Autism, but Many," *Autism Speaks*, https://www.autismspeaks.org/what-autism.

[32] Andrew J. Wakefield et al, "Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children," *The Lancet*, Vol. 351, February 18, 1998, p. 641.

[33] Azhar Hussain, Syed Ali, Madiha Ahmed, and Sheharyar Hussain, The Anti-Vaccination Movement: A Regression in Modern Medicine," *Cureus*, July, 3, 2018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6122668/#REF2.

[34] "Ten Threats to Global Health in 2019," World Health Organization, https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.

[35] "Measles Vaccination: Myths and Facts," *Infectious Diseases Society of America (IDSA),* https://www.idsociety.org/public-health/measles/myths-and-facts/.

[36] Brian Deer, "How the Case Against the MMR Vaccine Was Fixed," *The British Medical Journal*, January 6, 2011. https://www.bmj.com/content/342/bmj.c5347.

[37] Alice Park. Doctor Behind Vaccine-Autism Link Loses License," *Time*, May 24, 2010. http://healthland.time.com/2010/05/24/doctor-behind-vaccine-autism-link-loses-license/.

[38] Susan Dominus, "The Crash and Burn of An Autism Guru," *The New York Times Magazine*, April 20, 2011. https://www.nytimes.com/2011/04/24/magazine/mag-24Autism-t.html.

[39] Andrew Buncombe, "Andrew Wakefield: How a Disgraced UK Doctor Has Remade Himself in Anti-Vaxxer Trump's America," *Independent*, May 4, 2018, https://www.independent.co.uk/news/world/americas/andrew-wakefield-anti-vaxxer-trump-us-mmr-autism-link-lancet-fake-a8331826.html.

[40] Rachel Loomes, Laura Hill, William Polmear Locke Mandy, "What is the Male-to-Female Ratio in Autism Spectrum Disorder A Systematic Review and Meta-Analysis," *Journal of the American Academy of Child Adolescent Psychiatry*, Vol. 56, No. 6, June 2017, pp. 466-474. https://www.ncbi.nlm.nih.gov/pubmed/28545751.

[41] "What Causes Autism," *Autism Speaks*, https://www.autismspeaks.org/what-causes-autism.