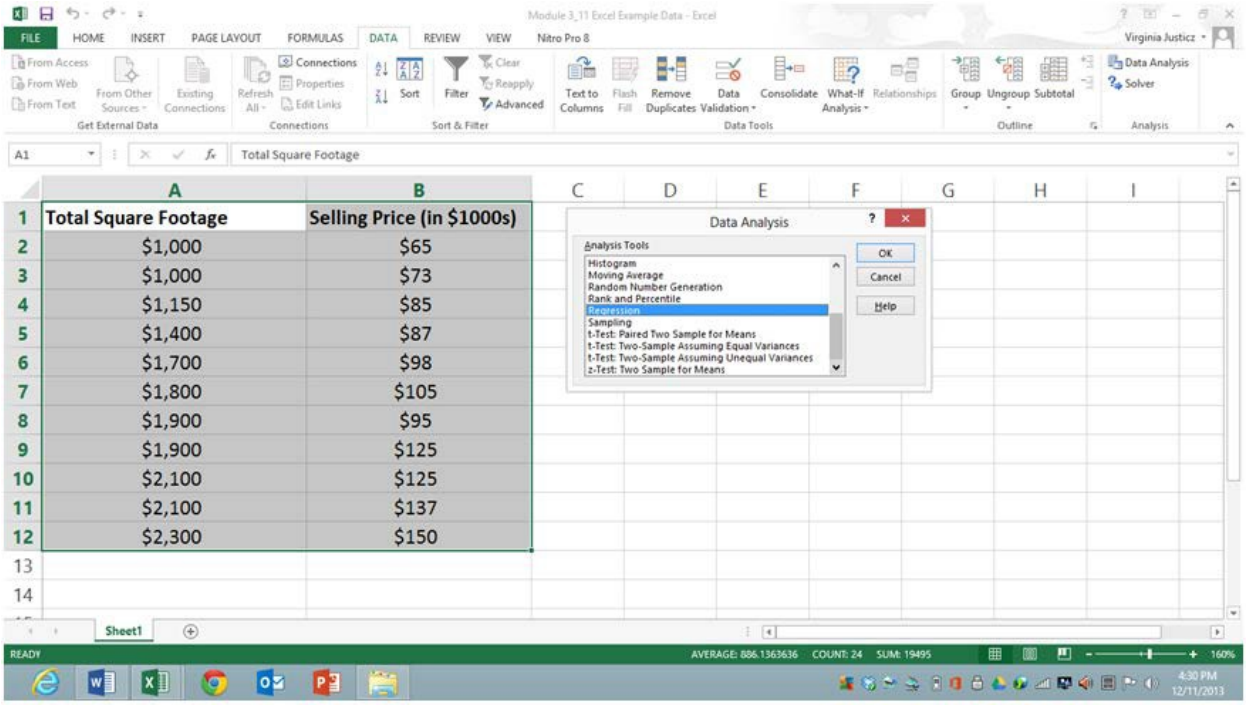
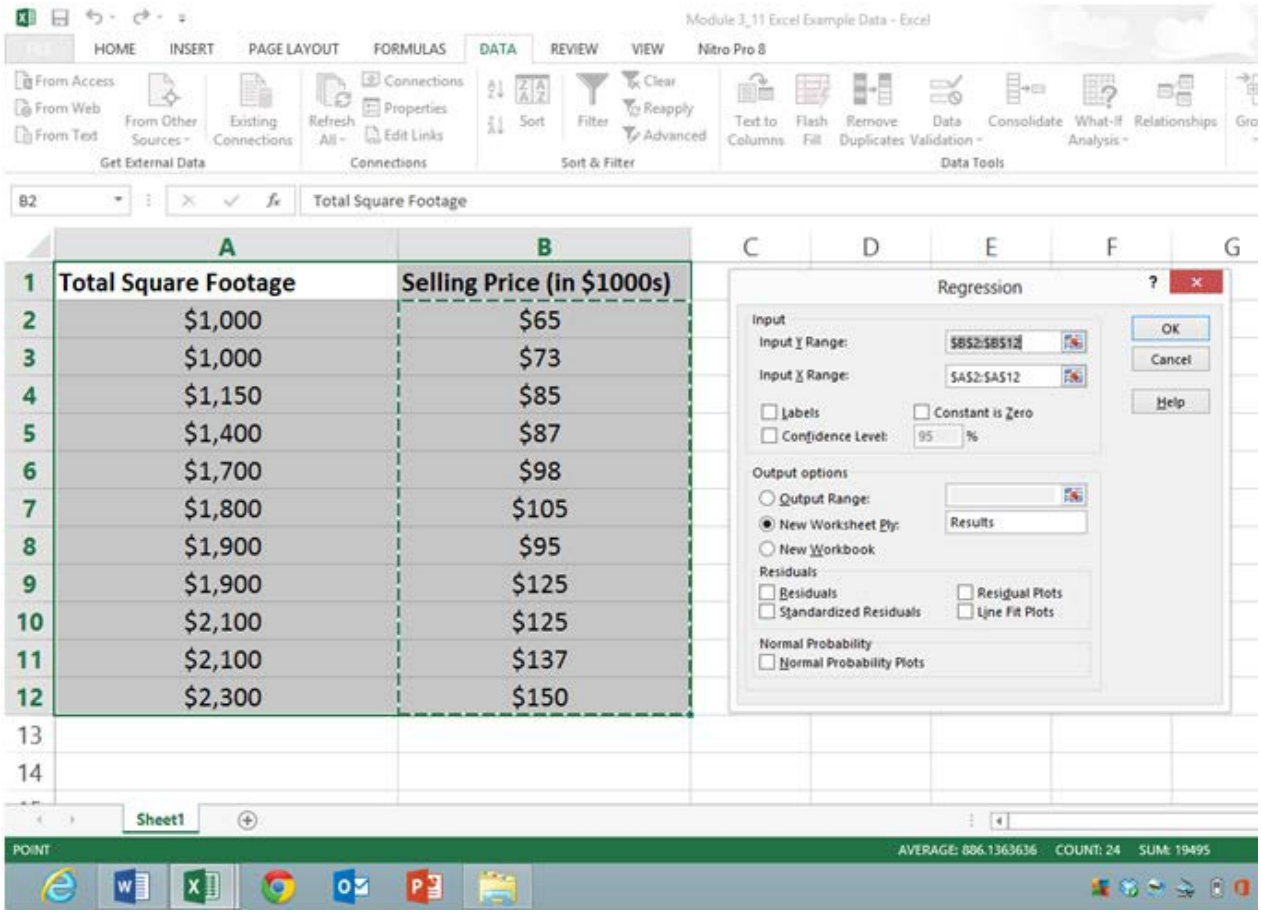
Linear Regression



First create a table in Excel with two columns of data.

Click on the "Data Analysis" option in the "Data" tab. Then choose "Regression" from the options.



	A	B
1	Total Square Footage	Selling Price (in $1000s)
2	$1,000	$65
3	$1,000	$73
4	$1,150	$85
5	$1,400	$87
6	$1,700	$98
7	$1,800	$105
8	$1,900	$95
9	$1,900	$125
10	$2,100	$125
11	$2,100	$137
12	$2,300	$150

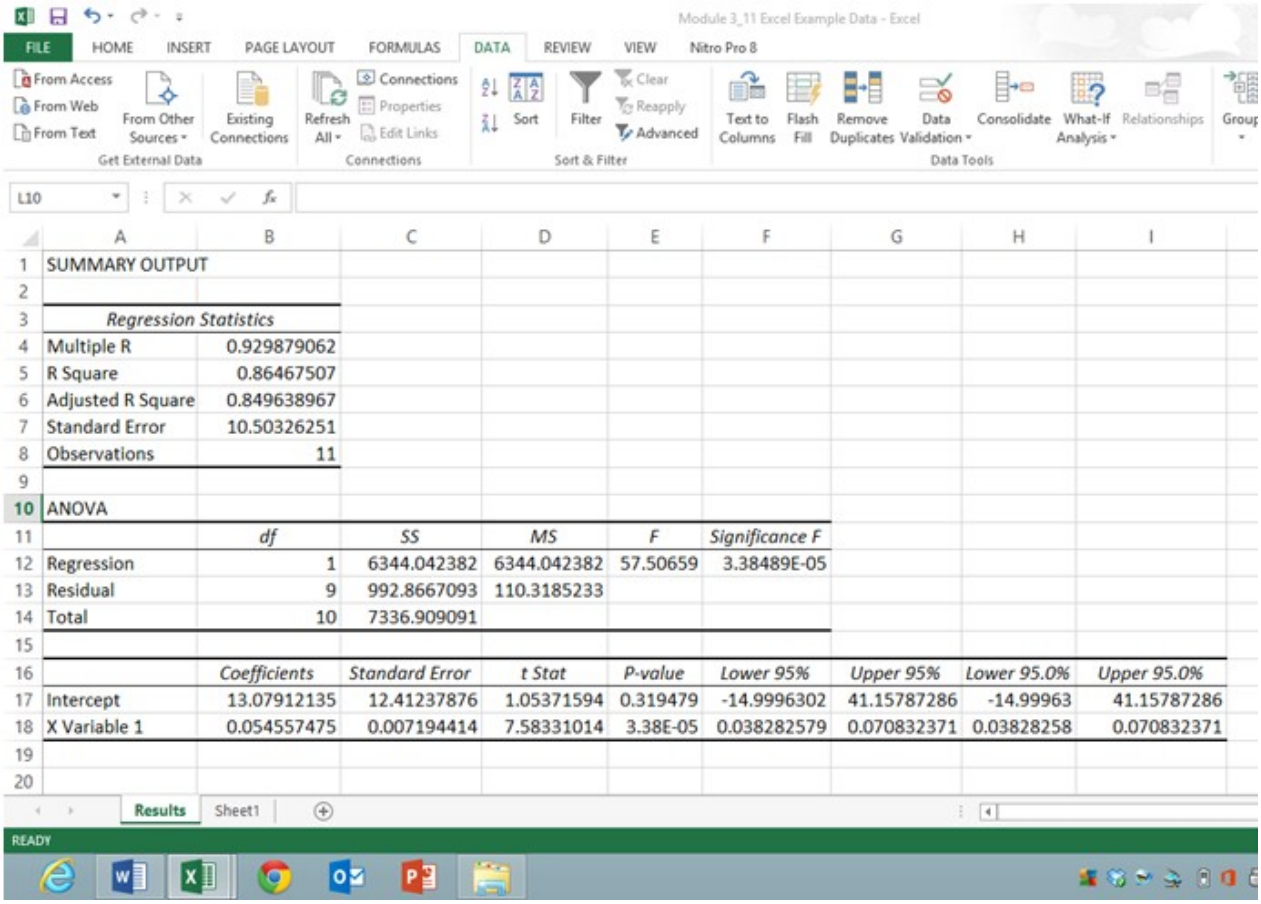
In the "input Y range" box, highlight the cells containing the sample observations for the dependent variable (in this case, selling prices, or B2-B12).

In the "input X range" box, highlight the cells containing the sample observations for the independent variable (in this case, square footage, or A2-A12).

Then specify the output range where we want the regression results to be reported (in the case below, we selected the New Worksheet Ply option to indicate that we want the regression results placed on a new sheet named "Results").

Click OK.

As you can see, the regression function produces a number of results based on the sample observation data. For now, we will focus on only a few key values in the summary output below.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.929879062							
5	R Square	0.86467507							
6	Adjusted R Square	0.849638967							
7	Standard Error	10.50326251							
8	Observations	11							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	6344.042382	6344.042382	57.50659	3.38489E-05			
13	Residual	9	992.8667093	110.3185233					
14	Total	10	7336.909091						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	13.07912135	12.41237876	1.05371594	0.319479	-14.9996302	41.15787286	-14.99963	41.15787286
18	X Variable 1	0.054557475	0.007194414	7.58331014	3.38E-05	0.038282579	0.070832371	0.03828258	0.070832371
19									
20									

As you can see, the regression function produces a number of results based on the sample observation data. For now, we will focus on only a few key values in the summary output below.A regression line is defined by the equation: y = mx + b.

Note the value labeled “Coefficients” in the "Intercept" row of the Summary Output below. This represents the optimal value for "b" in the regression equation for this particular set of data. The Coefficient value in the "X Variable 1" row represents the optimal value for "m" in the equation. If we write the equation for the regression line using the values from our summary output for the regression problem, we get the following:

Y= 0.054557475x + 13.07912135

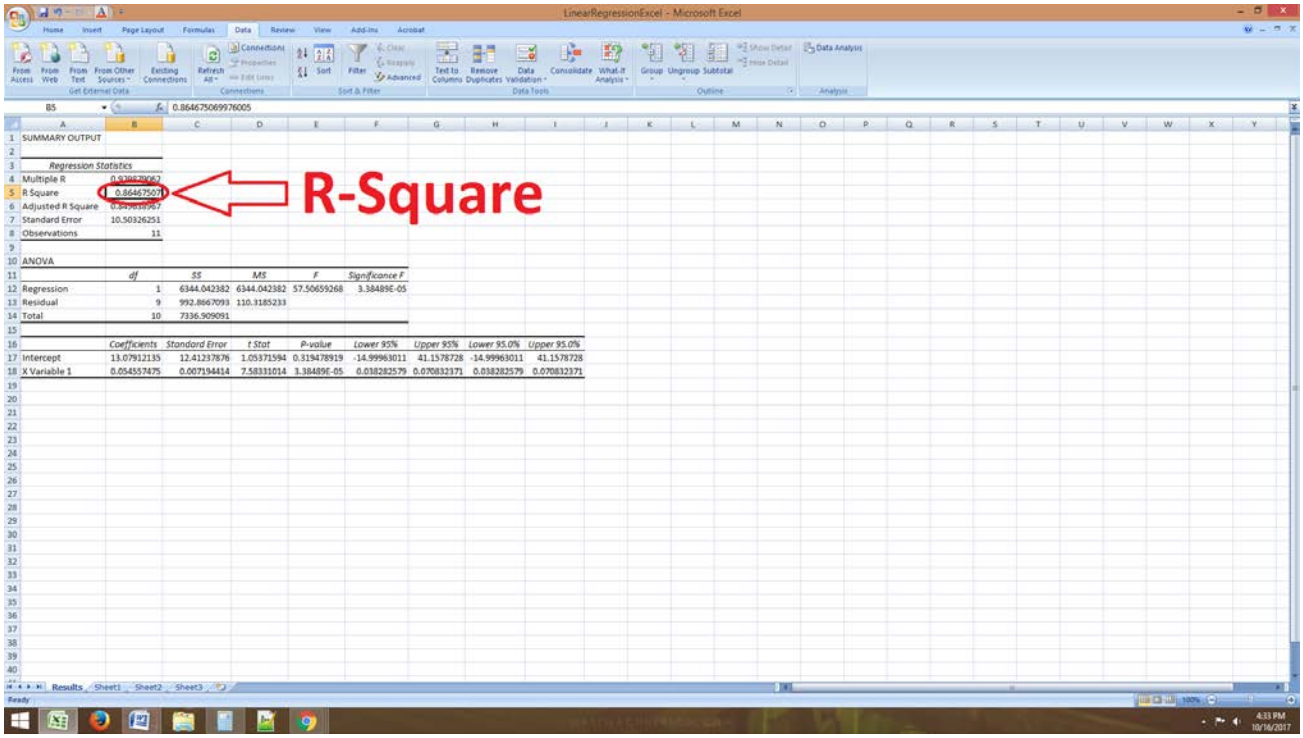
In this example, x is the independent variable (square footage) and y is the dependent variable (selling price). So we can use the equation to predict the selling price by putting in a specific square footage value for x and calculating the likely selling price (y).

For example, if we want to know the likely selling price in this neighborhood for a house with 3,000 square feet, we would calculate the following:

0.054557475 * 3,000 + 13.07912135 = y

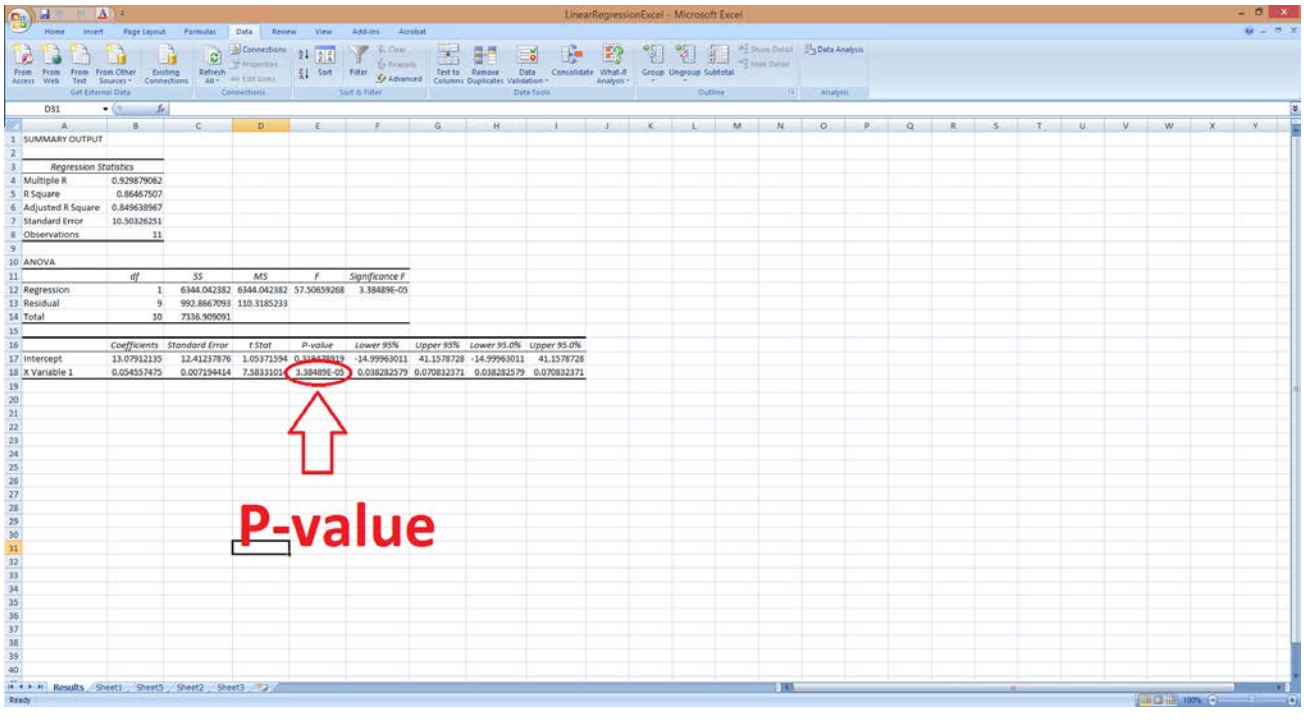
The next thing to check is the accuracy of this analysis. Is our regression analysis predictive and useful?

To answer this question, the next value we’ll look at in the summary output is R Square. If you look under Regression Statistics, you can see the R Square value of 0.86. As we know, an R-square value close to one indicates that the data closely aligns to the regression line.

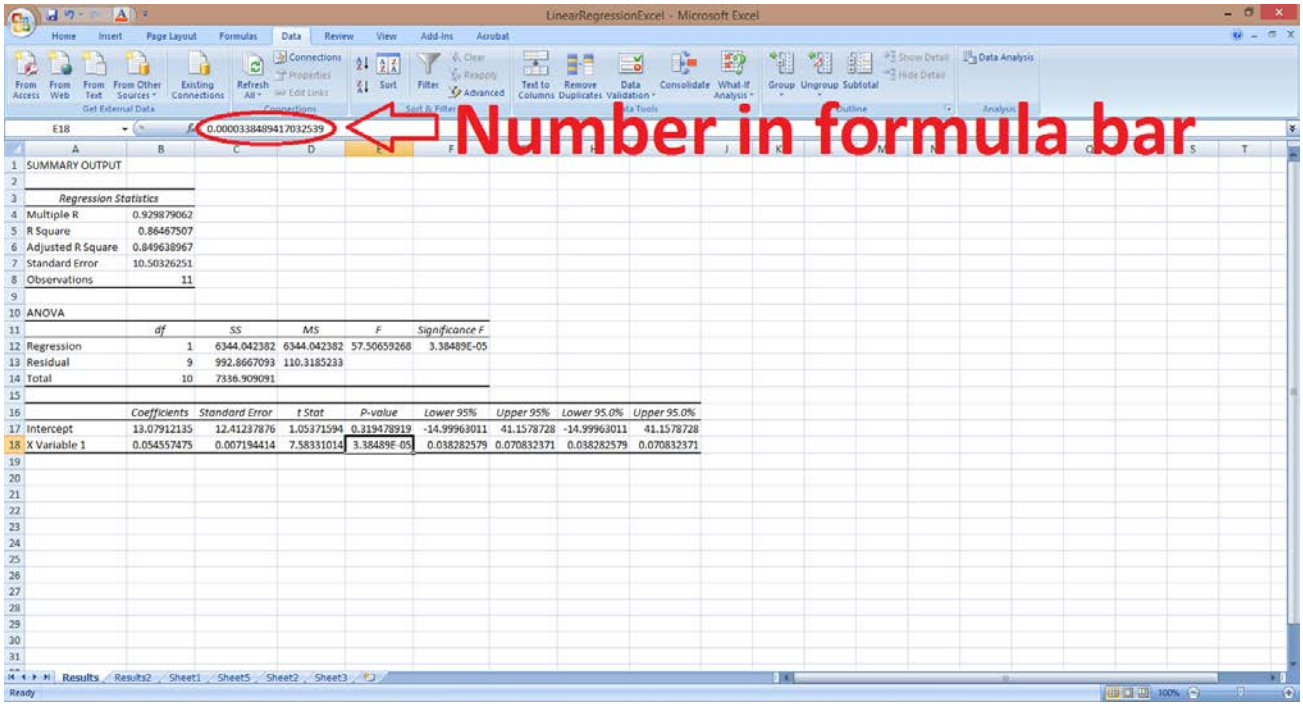


This value is close enough to 1 to apply linear regression here, but it is important to remember that linear regression is only an estimate, and isn't perfectly predictive. An R-Square value of 1.0 would indicate that the variables are perfectly correlated. Here, with an R-Square value of 0.86, we see that Total Square Footage is closely tied to Selling Price, but you can imagine that there are other factors at play. One home may be slightly smaller, but sell for more money because it has high end appliances and new siding.

Another value we'll examine is the p-value.



The X Variable 1 line displays summary statistics for the relationship between the x- and y-variables, which are Total Square Footage and Selling Price, respectively. This p-value is for the hypothesis test for a relationship between x and y. A p-value less than 0.05 indicates that there is a statistically significant relationship between x and y. Here, our p-value is so low that it is represented with scientific notation.



	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.929879062				
5	R Square	0.86467507				
6	Adjusted R Square	0.849638967				
7	Standard Error	10.50326251				
8	Observations	11				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	6344.042382	6344.042382	57.50659268	3.38489E-05
13	Residual	9	992.8667093	110.3185233		
14	Total	10	7336.909091			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	13.07912135	12.41237876	1.05371594	0.319478919	-14.99963011
18	X Variable 1	0.054557475	0.007194414	7.58331014	3.38489E-05	0.038282579

	G	H	I
16	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	41.1578728	-14.99963011	41.1578728
18	0.070832371	0.038282579	0.070832371

If we select the cell, we can see the number written in standard form in the formula bar. This number is obviously much smaller than 0.05. Therefore, there is a significant relationship between our x and y variables, Total Square Footage and Selling Price.