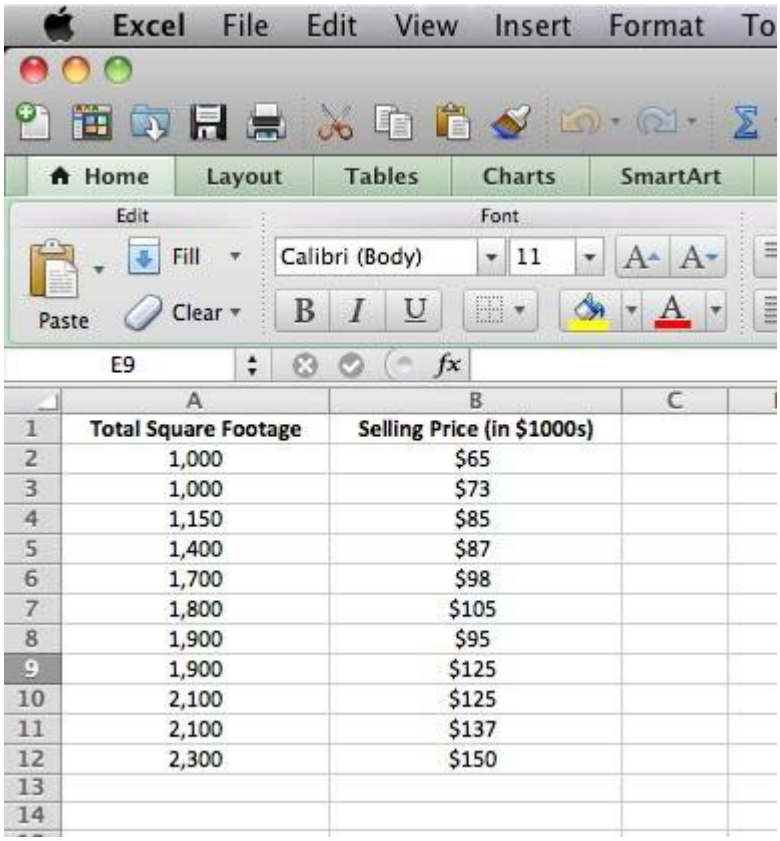
MindEdge LEARNING

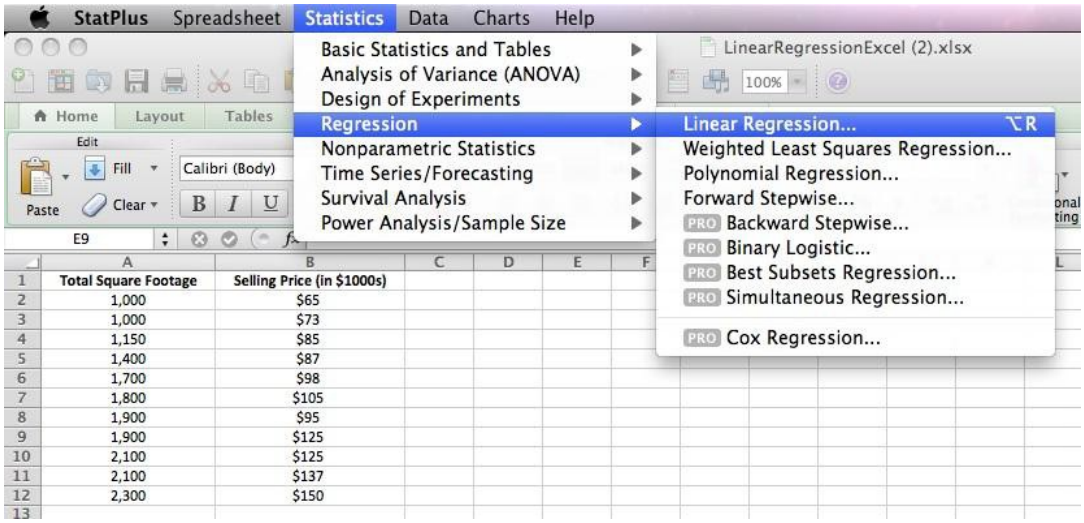
Excel on Macs: Linear Regression



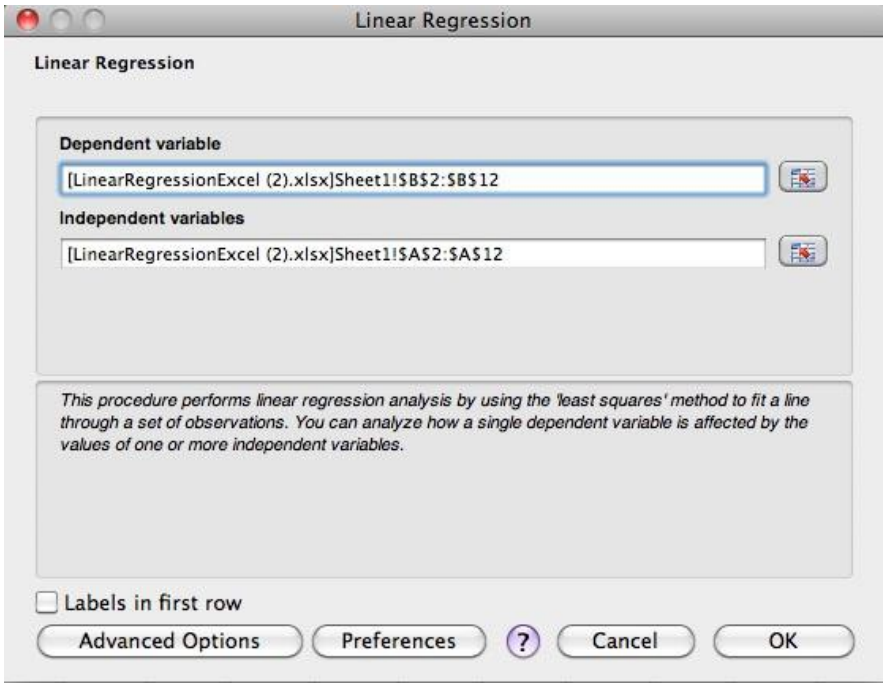
	A	B	C
1	Total Square Footage	Selling Price (in $1000s)	
2	1,000	$65	
3	1,000	$73	
4	1,150	$85	
5	1,400	$87	
6	1,700	$98	
7	1,800	$105	
8	1,900	$95	
9	1,900	$125	
10	2,100	$125	
11	2,100	$137	
12	2,300	$150	
13			
14			

In this video, we'll use the functions in Excel and StatPlus to create Linear Regression output from a dataset. In this case, the data set are test scores, listed in this column here. In this problem, we're looking at the relationship between houses' square footage and selling price.

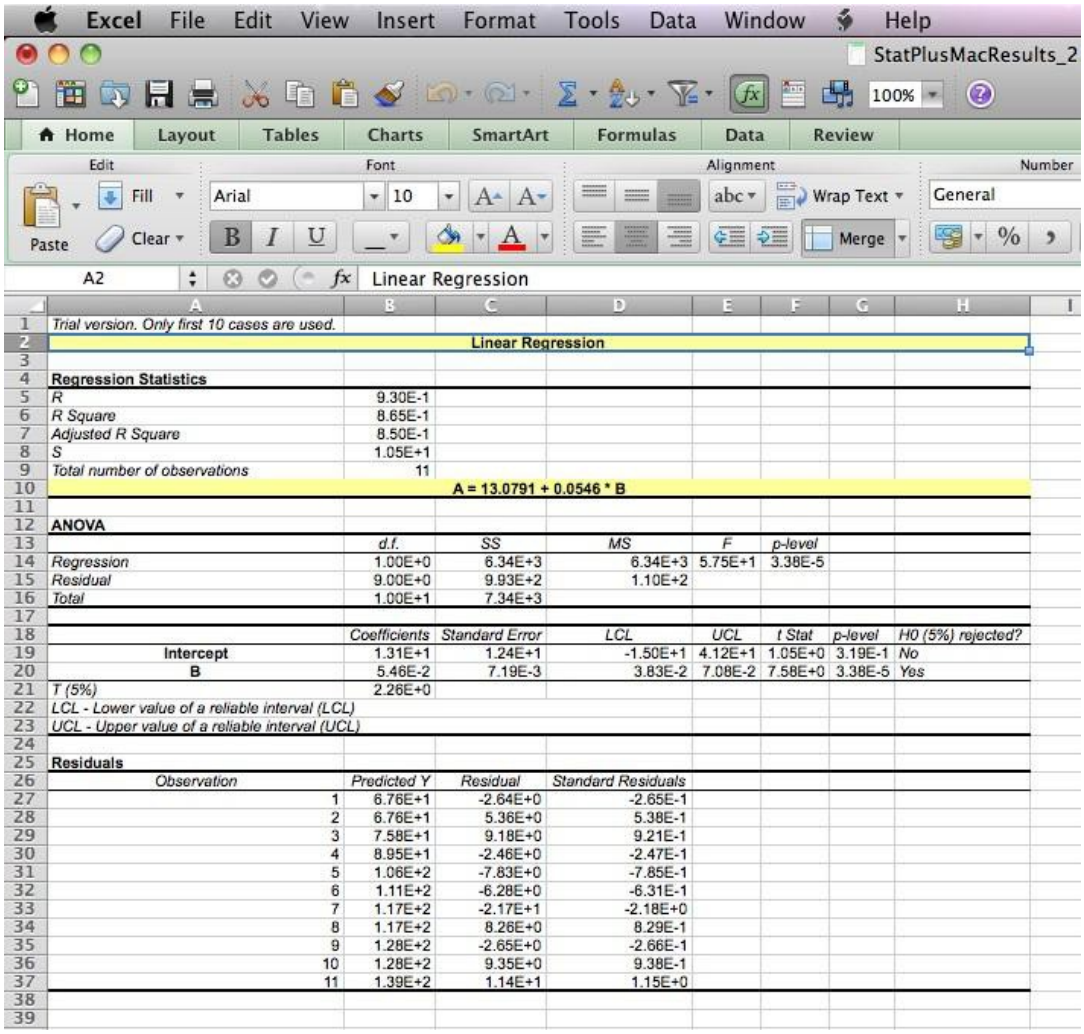
It is important to have opened both Microsoft Excel AND StatPlus.



After having opened the data in Microsoft Excel, the next step is to go to the StatPlus Application, select "Statistics", "Regression", and "Linear Regression…" which should open the following window:

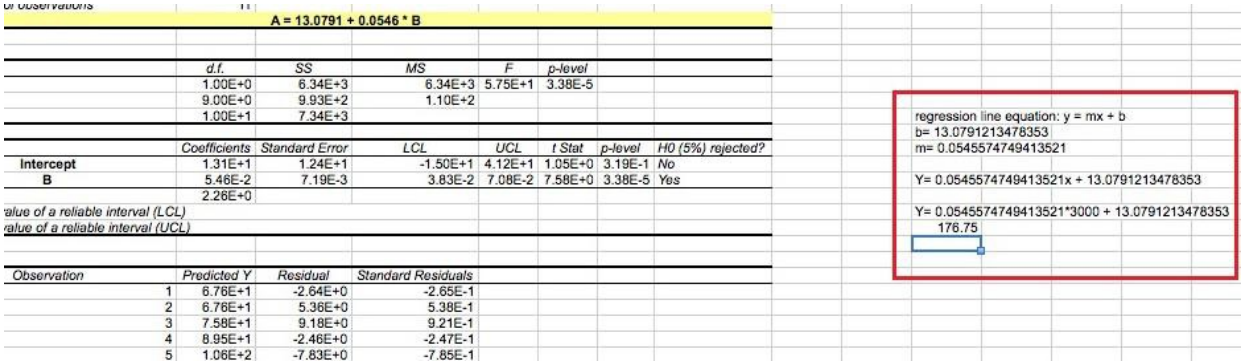


In this window, it is important to select the data. To do this, you should click on the button to the right of the variables empty box and return to the Excel sheet where the dependent variable is the selling price dataset and the independent variable is the square footage dataset. Choose not to include “Labels in first row”. Once both variables have been selected and the function knows the labels are not in the first row, click “OK”.



	A	B	C	D	E	F	G	H	I
1	*Trial version. Only first 10 cases are used.*								
2				**Linear Regression**					
3									
4	**Regression Statistics**								
5	*R*	9.30E-1							
6	*R Square*	8.65E-1							
7	*Adjusted R Square*	8.50E-1							
8	*S*	1.05E+1							
9	*Total number of observations*	11							
10				**A = 13.0791 + 0.0546 * B**					
11									
12	**ANOVA**								
13		*d.f.*	*SS*	*MS*	*F*	*p-level*			
14	*Regression*	1.00E+0	6.34E+3	6.34E+3	5.75E+1	3.38E-5			
15	*Residual*	9.00E+0	9.93E+2	1.10E+2					
16	*Total*	1.00E+1	7.34E+3						
17									
18		*Coefficients*	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*	
19	**Intercept**	1.31E+1	1.24E+1	-1.50E+1	4.12E+1	1.05E+0	3.19E-1	*No*	
20	**B**	5.46E-2	7.19E-3	3.83E-2	7.08E-2	7.58E+0	3.38E-5	*Yes*	
21	*T (5%)*	2.26E+0							
22	*LCL - Lower value of a reliable interval (LCL)*								
23	*UCL - Upper value of a reliable interval (UCL)*								
24									
25	**Residuals**								
26	*Observation*	*Predicted Y*	*Residual*	*Standard Residuals*					
27	1	6.76E+1	-2.64E+0	-2.65E-1					
28	2	6.76E+1	5.36E+0	5.38E-1					
29	3	7.58E+1	9.18E+0	9.21E-1					
30	4	8.95E+1	-2.46E+0	-2.47E-1					
31	5	1.06E+2	-7.83E+0	-7.85E-1					
32	6	1.11E+2	-6.28E+0	-6.31E-1					
33	7	1.17E+2	-2.17E+1	-2.18E+0					
34	8	1.17E+2	8.26E+0	8.29E-1					
35	9	1.28E+2	-2.65E+0	-2.66E-1					
36	10	1.28E+2	9.35E+0	9.38E-1					
37	11	1.39E+2	1.14E+1	1.15E+0					
38									
39									

This output for the Linear Regression should appear.



A = 13.0791 + 0.0546 * B

	d.f.	SS	MS	F	p-level
	1.00E+0	6.34E+3	6.34E+3	5.75E+1	3.38E-5
	9.00E+0	9.93E+2	1.10E+2		
	1.00E+1	7.34E+3			

	Coefficients	Standard Error	LCL	UCL	t Stat	p-level	H0 (5%) rejected?
Intercept	1.31E+1	1.24E+1	-1.50E+1	4.12E+1	1.05E+0	3.19E-1	No
B	5.46E-2	7.19E-3	3.83E-2	7.08E-2	7.58E+0	3.38E-5	Yes
	2.26E+0						

alue of a reliable interval (LCL)
alue of a reliable interval (UCL)

Observation	Predicted Y	Residual	Standard Residuals
1	6.76E+1	-2.64E+0	-2.65E-1
2	6.76E+1	5.36E+0	5.38E-1
3	7.58E+1	9.18E+0	9.21E-1
4	8.95E+1	-2.46E+0	-2.47E-1
5	1.06E+2	-7.83E+0	-7.85E-1

regression line equation: y = mx + b
b= 13.0791213478353
m= 0.0545574749413521

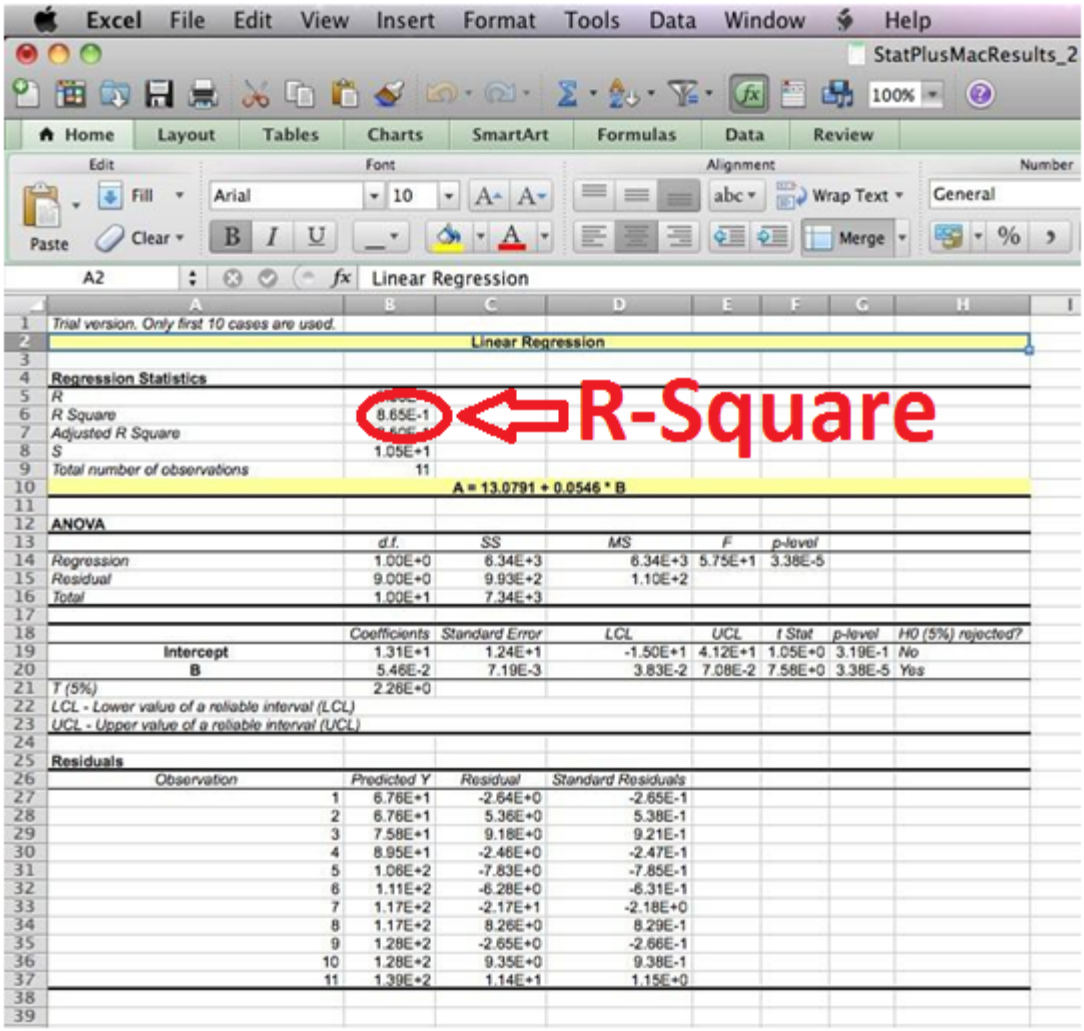
Y= 0.0545574749413521x + 13.0791213478353

Y= 0.0545574749413521*3000 + 13.0791213478353
176.75

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 "Intercept" in the Summary Output above. This represents the optimal value for "b" in the regression equation for this particular set of data. The value labeled "B" 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
Y = 176.75
Therefore, we can assume that a house with 3,000 square feet would cost near $176,750.

The first 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.



Trial version. Only first 10 cases are used.

Linear Regression

Regression Statistics

R	
R Square	8.65E-1
Adjusted R Square	
S	1.05E+1
Total number of observations	11

A = 13.0791 + 0.0546 * B

ANOVA

	d.f.	SS	MS	F	p-level
Regression	1.00E+0	6.34E+3	6.34E+3	5.75E+1	3.38E-5
Residual	9.00E+0	9.93E+2	1.10E+2		
Total	1.00E+1	7.34E+3			

	Coefficients	Standard Error	LCL	UCL	t Stat	p-level	H0 (5%) rejected?
Intercept	1.31E+1	1.24E+1	-1.50E+1	4.12E+1	1.05E+0	3.19E-1	No
B	5.46E-2	7.19E-3	3.83E-2	7.08E-2	7.58E+0	3.38E-5	Yes
T (5%)	2.26E+0						

LCL - Lower value of a reliable interval (LCL)

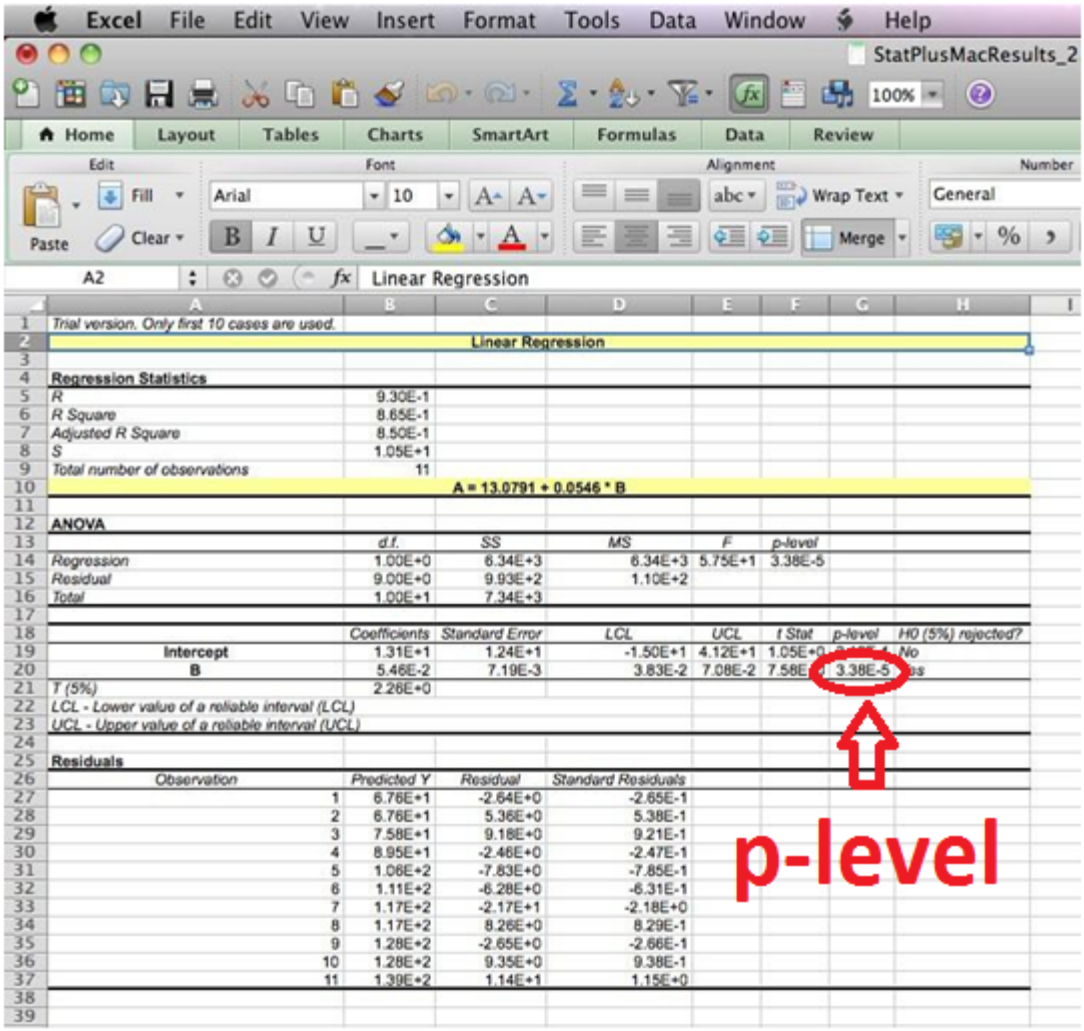
UCL - Upper value of a reliable interval (UCL)

Residuals

Observation	Predicted Y	Residual	Standard Residuals
1	6.76E+1	-2.64E+0	-2.65E-1
2	6.76E+1	5.36E+0	5.38E-1
3	7.58E+1	9.18E+0	9.21E-1
4	8.95E+1	-2.46E+0	-2.47E-1
5	1.06E+2	-7.83E+0	-7.85E-1
6	1.11E+2	-6.28E+0	-6.31E-1
7	1.17E+2	-2.17E+1	-2.18E+0
8	1.17E+2	8.26E+0	8.29E-1
9	1.28E+2	-2.65E+0	-2.66E-1
10	1.28E+2	9.35E+0	9.38E-1
11	1.39E+2	1.14E+1	1.15E+0

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.



1	*Trial version. Only first 10 cases are used.*							
2				Linear Regression				
3								
4	**Regression Statistics**							
5	*R*	9.30E-1						
6	*R Square*	8.65E-1						
7	*Adjusted R Square*	8.50E-1						
8	*S*	1.05E+1						
9	*Total number of observations*	11						
10				A = 13.0791 + 0.0546 * B				
11								
12	**ANOVA**							
13		*d.f.*	*SS*	*MS*	*F*	*p-level*		
14	*Regression*	1.00E+0	6.34E+3	6.34E+3	5.75E+1	3.38E-5		
15	*Residual*	9.00E+0	9.93E+2	1.10E+2				
16	*Total*	1.00E+1	7.34E+3					
17								
18		*Coefficients*	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*
19	**Intercept**	1.31E+1	1.24E+1	-1.50E+1	4.12E+1	1.05E+0		*No*
20	**B**	5.46E-2	7.19E-3	3.83E-2	7.08E-2	7.58E	3.38E-5	*es*
21	*T (5%)*	2.26E+0						
22	*LCL - Lower value of a reliable interval (LCL)*							
23	*UCL - Upper value of a reliable interval (UCL)*							
24								
25	**Residuals**							
26	*Observation*	*Predicted Y*	*Residual*	*Standard Residuals*				
27	1	6.76E+1	-2.64E+0	-2.65E-1				
28	2	6.76E+1	5.36E+0	5.38E-1				
29	3	7.58E+1	9.18E+0	9.21E-1				
30	4	8.95E+1	-2.46E+0	-2.47E-1				
31	5	1.06E+2	-7.83E+0	-7.85E-1				
32	6	1.11E+2	-6.28E+0	-6.31E-1				
33	7	1.17E+2	-2.17E+1	-2.18E+0				
34	8	1.17E+2	8.26E+0	8.29E-1				
35	9	1.28E+2	-2.65E+0	-2.66E-1				
36	10	1.28E+2	9.35E+0	9.38E-1				
37	11	1.39E+2	1.14E+1	1.15E+0				
38								
39								

The second 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, labeled p-level in this program, 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 much smaller than 0.05. Therefore, there is a significant relationship between our x and y variables, Total Square Footage and Selling Price.