Regression step-by-step using Microsoft Excel®

Notes prepared by Pamela Peterson Drake, James Madison University

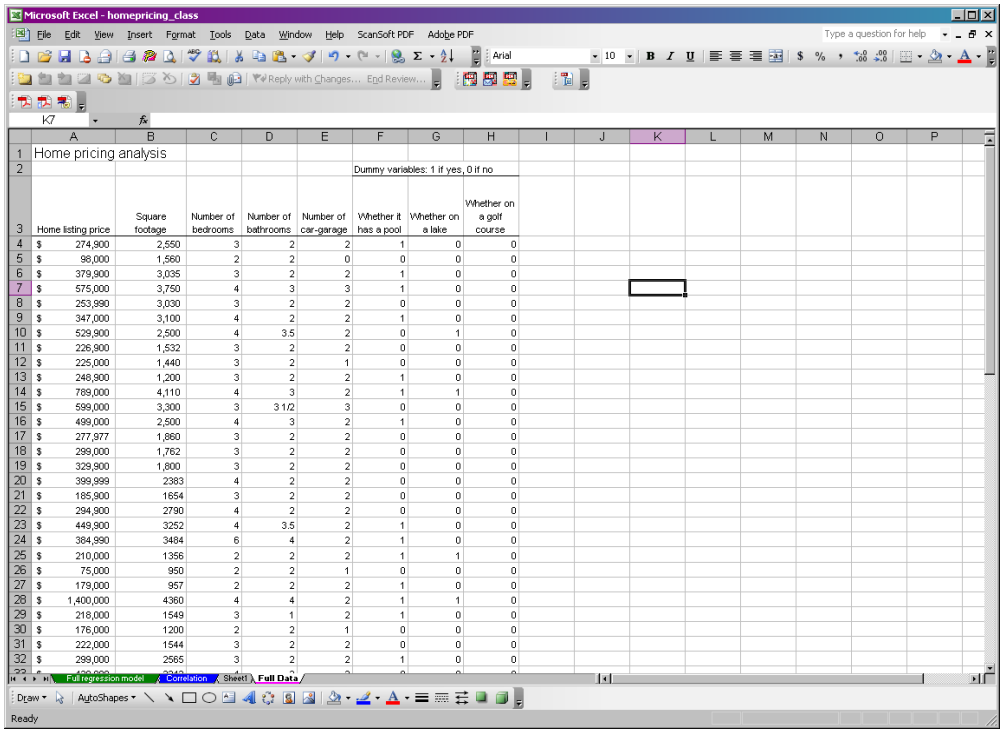
Step 1: Type the data into the spreadsheet

The example used throughout this "How to" is a regression model of home prices, explained by:

- square footage,
- number of bedrooms,
- number of bathrooms,
- number of garages,
- whether it has a pool,
- whether it is on a lake, and
- whether it is on a golf course.

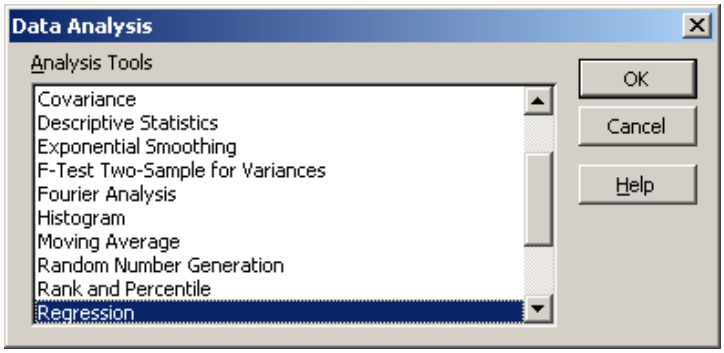
The objective is to explain the variation in home prices, using the variation in the independent variables. In other words, we are asking the question of "Why do home prices vary from home to home?" It may be because the homes have different features. Hence, we are using the variation in the features to explain the variation in the home prices.

You need to arrange the data in columns to use the built in regression function within Microsoft Excel. The first column contains the observations on the dependent variable and then the other, adjoining columns containing the observations on the independent variables. Include column headings to make it is easier to interpret your results.



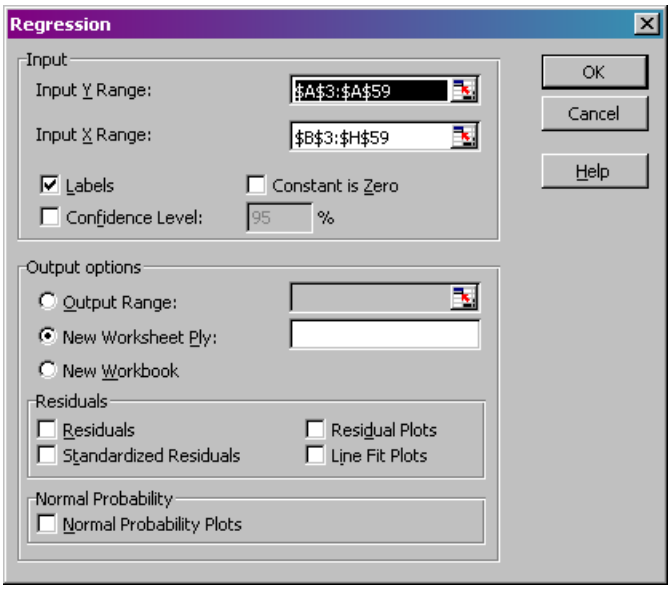
Step 2: Use Excel®'s Data Analysis program, Regression

In the **Tools** menu, you will find a **Data Analysis** option.[1] Within **Data Analysis**, you should then choose **Regression**:



Step 3: Specify the regression data and output

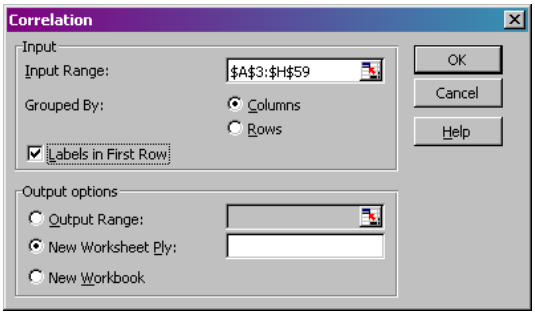
You will see a pop-up box for the regression specifications. Using this screen, you can then specify the dependent variable [Input Y Range] and the columns of the independent variables [Input X Range].



If you include the variable names in the column headings and these column headings are part of the range of observations that you specified, be sure to check the **Labels** box.

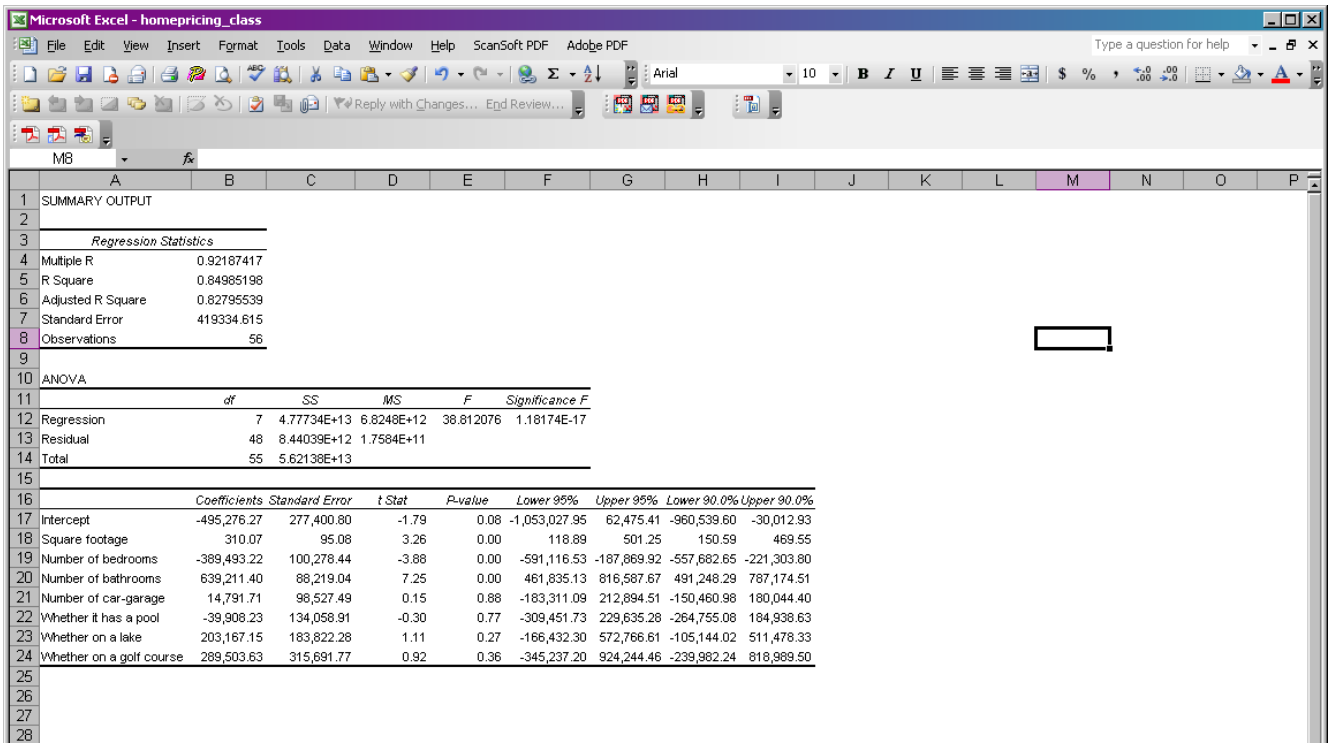
You can then specify where you would like to place the results. If you leave the default checked as **New Worksheet Ply**, a new worksheet is created that will contain the results.

This process is similar to the correlation specification. For example, if we want to view the correlation among the dependent variables, we would use a similar process using the Data Analysis function "Correlation":



[1] If you do not find this option, you will want to click on **Add-ins** and then specify **Data Analysis** as an option.

The results of the regression analysis appear as follows:



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.92187417							
5	R Square	0.84985198							
6	Adjusted R Square	0.82795539							
7	Standard Error	419334.615							
8	Observations	56							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	7	4.77734E+13	6.8248E+12	38.812076	1.18174E-17			
13	Residual	48	8.44039E+12	1.7584E+11					
14	Total	55	5.62138E+13						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
17	Intercept	-495,276.27	277,400.80	-1.79	0.08	-1,053,027.95	62,475.41	-960,539.60	-30,012.93
18	Square footage	310.07	95.08	3.26	0.00	118.89	501.25	150.59	469.55
19	Number of bedrooms	-389,493.22	100,278.44	-3.88	0.00	-591,116.53	-187,869.92	-557,682.65	-221,303.80
20	Number of bathrooms	639,211.40	88,219.04	7.25	0.00	461,835.13	816,587.67	491,248.29	787,174.51
21	Number of car-garage	14,791.71	98,527.49	0.15	0.88	-183,311.09	212,894.51	-150,460.98	180,044.40
22	Whether it has a pool	-39,908.23	134,058.91	-0.30	0.77	-309,451.73	229,635.28	-264,755.08	184,938.63
23	Whether on a lake	203,167.15	183,822.28	1.11	0.27	-166,432.30	572,766.61	-105,144.02	511,478.33
24	Whether on a golf course	289,503.63	315,691.77	0.92	0.36	-345,237.20	924,244.46	-239,982.24	818,989.50
25									
26									
27									
28									

Step 4: Interpret the results

Summary information

Regression Statistics	
Multiple R	0.92187417
R Square	0.84985198
Adjusted R Square	0.82795539
Standard Error	419334.615
Observations	56

1. The **multiple correlation coefficient** is 0.92187417. This indicates that the correlation among the independent and dependent variables is positive. This statistic, which ranges from -1 to +1, does not indicate statistical significance of this correlation.
2. The **coefficient of determination**, R^2, is 84.99%. This means that close to 85% of the variation in the dependent variable (home prices) is explained by the independent variables.
3. The **adjusted R-square**, a measure of explanatory power, is 0.82795539. This statistic is not generally interpreted because it is neither a percentage (like the R^2), nor a test of significance (such as the F-statistic).
4. The **standard error of the regression** is $419,334, which is an estimate of the variation of the observed home prices, in dollar terms, about the regression line.[2]

Analysis of variance

[2] The standard error is in the same unit of measurement as the dependent variable.

The analysis of variance information provides the breakdown of the total variation of the dependent variable in this case home prices) in to the explained and unexplained portions.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	4.77734E+13	6.8248E+12	38.812076	1.18174E-17
Residual	48	8.44039E+12	1.7584E+11		
Total	55	5.62138E+13			

1. The **SS Regression** is the variation explained by the regression line; **SS Residual** is the variation of the dependent variable that is not explained.
2. The **F-statistic** is calculated using the ratio of the mean square regression (**MS Regression**) to the mean square residual (**MS Residual**). This is statistic can then be compared with the critical F value for 7 and 48 degrees of freedom (available from an F-table) to test the null hypothesis:

   H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

   v.

   H_A: at least one β_i not equal to 0
3. The p-value associated with the calculated F-statistic is probability beyond the calculated value. Comparing this value with 5%, for example, indicates rejection of the null hypothesis.

The estimated regression line

The results of the estimated regression line include the estimated coefficients, the standard error of the coefficients, the calculated t-statistic, the corresponding p-value, and the bounds of both the 95% and the 90% confidence intervals.

The independent variables that statistically significant in explaining the variation in the home prices are the square footage, the number of bedrooms, and the number of bathrooms, as indicated by (1) calculated t-statistics that exceed the critical values, and (2) the calculated p-values that are less than the significance level of 5%.

1. The relationship between square footage and home prices is positive: the larger the square footage, the higher the home price. The coefficient of 310.07 indicates, on average, an additional square foot increases the home price by $310.07.
2. The number of bedrooms is negatively related to the home price, but this may be due to an interaction with the square footage variable because larger homes tend to have more bedrooms.
3. The number of bathrooms is positively related to home prices. Adding a bathroom, apart from the effect on square footage, increases the home price.
4. The other independent variables do not add not significantly in explaining the variation in home prices.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	-495,276.27	277,400.80	-1.79	0.08	-1,053,027.95	62,475.41	-960,539.60	-30,012.93
Square footage	310.07	95.08	3.26	0.00	118.89	501.25	150.59	469.55
Number of bedrooms	-389,493.22	100,278.44	-3.88	0.00	-591,116.53	-187,869.92	-557,682.65	-221,303.80
Number of bathrooms	639,211.40	88,219.04	7.25	0.00	461,835.13	816,587.67	491,248.29	787,174.51
Number of car-garage	14,791.71	98,527.49	0.15	0.88	-183,311.09	212,894.51	-150,460.98	180,044.40
Whether it has a pool	-39,908.23	134,058.91	-0.30	0.77	-309,451.73	229,635.28	-264,755.08	184,938.63
Whether on a lake	203,167.15	183,822.28	1.11	0.27	-166,432.30	572,766.61	-105,144.02	511,478.33
Whether on a golf course	289,503.63	315,691.77	0.92	0.36	-345,237.20	924,244.46	-239,982.24	818,989.50

Correlations

When using multiple regression to estimate a relationship, there is always the possibility of correlation among the independent variables. This correlation may be pair-wise or multiple correlation. Looking at the correlation, generated by the Correlation function within Data Analysis, we see that there is positive correlation among several variables:

Correlation coefficients	Home listing price	Square footage	Number of bedrooms	Number of bathrooms	Number of car-garage	Whether it has a pool	Whether on a lake	Whether on a golf course
Home listing price	1.000							
Square footage	0.765	1.000						
Number of bedrooms	0.282	0.642	1.000					
Number of bathrooms	0.873	0.795	0.499	1.000				
Number of car-garage	0.445	0.554	0.371	0.465	1.000			
Whether it has a pool	0.345	0.430	0.311	0.381	0.463	1.000		
Whether on a lake	0.598	0.527	0.130	0.534	0.349	0.280	1.000	
Whether on a golf course	0.067	0.004	-0.083	-0.003	0.038	0.179	-0.095	1.000

However, we cannot conclude that any of these correlations are important until we test for significance. calculating the test statistic for each of the pair-wise correlations above, we see that there are many statistically significant correlations (indicated in green), suggesting that multicollinearity may be a problem.

Test statistics	Square footage	Number of bedrooms	Number of bathrooms	Number of car-garage	Whether it has a pool	Whether on a lake	Whether on a golf course
Square footage							
Number of bedrooms	6.156						
Number of bathrooms	9.625	4.235					
Number of car-garage	4.890	2.936	3.861				
Whether it has a pool	3.495	2.404	3.029	3.840			
Whether on a lake	4.555	0.962	4.644	2.735	2.144		
Whether on a golf course	0.027	-0.613	-0.023	0.282	1.338	-0.702	

In the case of multicollinearity, we could either:

- Increase the sample size (which will often reduce the correlation among the independent variables, or
- Re-specify the regression model, removing or restating the independent variables such that there is less correlation among them.