Linear Regression Workshop

Dr. Kamesam

Learning Objectives

- Introduction to linear regression models, and building regression models in Excel.
- **Understand R-square, Adjusted R-Square**: measures of the variation in the dependent variable explained by the variation in the independent variables
- Evaluate the model through tests of hypothesis (**F** and **t** tests)
- **Backward Elimination method** to identify and eliminate the independent variables that may not be needed in the regression model. This method will help build a parsimonious model (compact model).

Data

We will use the dataset **house prices data for regression.xlsx** for this workshop. You can download files from Blackboard. The first 10 house data is in Figure 2.1.

Introduction

Regression analysis is a statistical process to estimate the relationships among variables. The focus is on the relationship between a **dependent** variable (**Y**) and one or more other changing variables, a.k.a. **independent** variables (**X**). Independent variables are also known as ***explanatory*** variables. Regression analysis can be used for prediction and forecasting. Regression analysis helps in understanding which independent variables have the predictive power to predict the value of the dependent variable. Often, regression analysis is used to understand how the independent variables influence the dependent variable **Y**. We need to make some assumptions about the form of the relationship. In this tutorial, we only want to consider **LINEAR** relationships between dependent and independent variables. Further assumptions are shown at the end of this document.

In this workshop, we will build several linear regression models on our data. Based on the results, we will learn how to evaluate the relationships between house prices and other variables, e.g. property size, house size, rooms etc. as well as correlations among independent variables

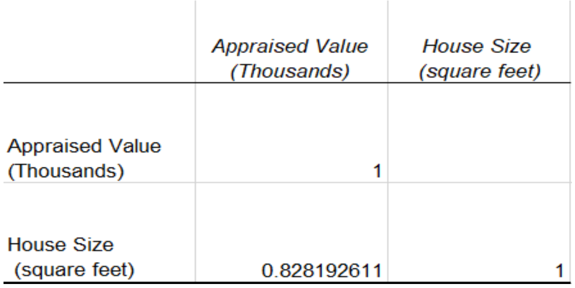
Excel **Data Analysis** ToolPak must be enabled to run Regression analysis. (See **Appendix 1)**.

1. Correlations Analysis

The very first step in conducting a regression analysis is to conduct a correlation analysis to measure and understand the strength of linear association between the dependent variable and the independent variable. In this example, we want to see the correlation between the dependent variable, **house price**, and independent variable, **house size** (sq. ft.).

Excel **Data Analysis** ToolPak must be enabled to run Correlation analysis. (See **Appendix 2)**.

- **Interpret the Output Generated by the Module (Single Independent Variable)**



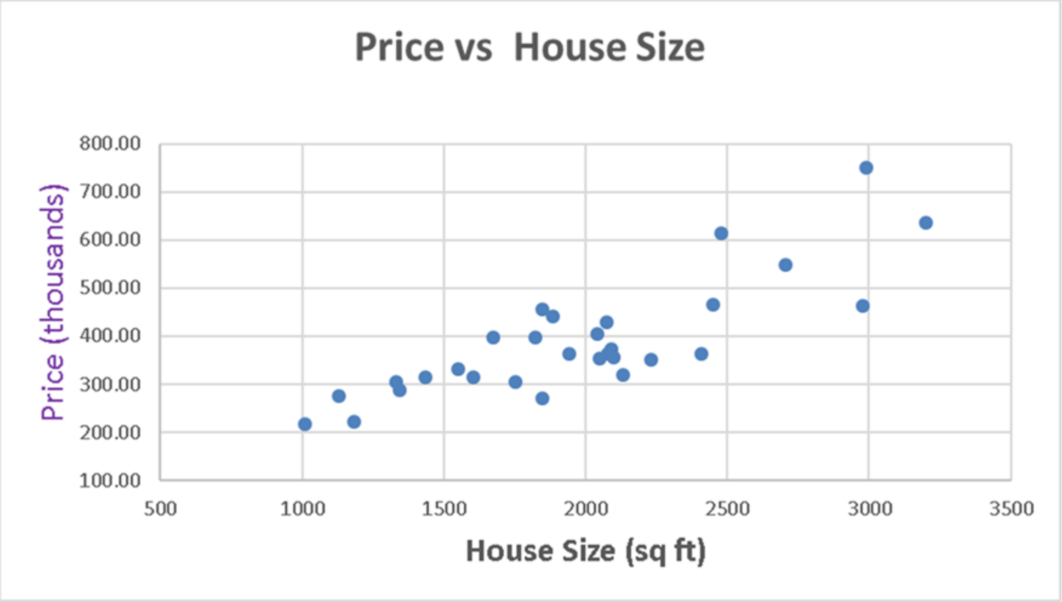
	Appraised Value (Thousands)	*House Size (square feet)*
Appraised Value (Thousands)	1	
House Size (square feet)	0.828192611	1

Figure 1.1 Correlations Analysis (Single Explanatory Variable)

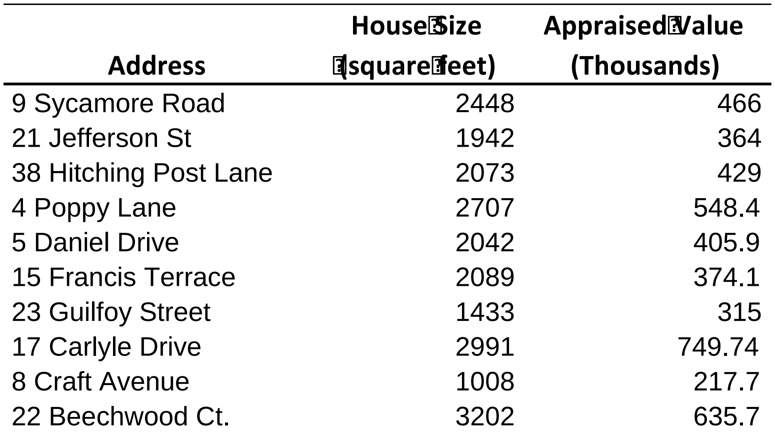
Correlations analysis is to look at the correlation between the dependent variable, house price, and independent variable, house size. It also displays the pair-wise linear correlation coefficient between the independent variables. In this example, we see that the correlation is 0.828, which is a strong correlation. Hence we can expect House-size to be able to predict Price fairly well. If this correlation is small (like 0.2, 0.3) the regression model cannot be that good.

Inspecting and understanding the data:
Before you conduct analysis or build a data mining model, an important step is to study the data and develop an understanding of the data. SPSS Modeler has tools to do so and you will learn to use those. If you are about to conduct a regression analysis, it is always a good idea to take a look at the scatter plot of the data. If there are several independent variables, you can look at pair-wise scatter plots.

Observations: Correlation analysis showed a strong positive correlation between House Size and its Price. You can see that in this scatter diagram. Prices of houses with 2500 sq. ft. are much more variable in price compared to houses with smaller sq. ft.



2. Simple Regression Analysis



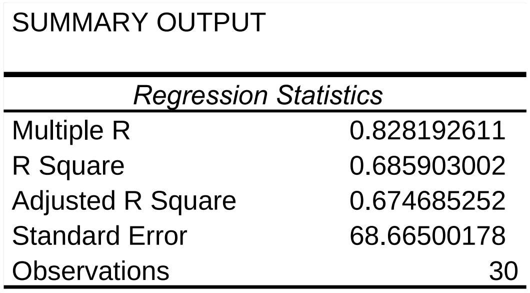
Address	House Size (square feet)	Appraised Value (Thousands)
9 Sycamore Road	2448	466
21 Jefferson St	1942	364
38 Hitching Post Lane	2073	429
4 Poppy Lane	2707	548.4
5 Daniel Drive	2042	405.9
15 Francis Terrace	2089	374.1
23 Guilfoy Street	1433	315
17 Carlyle Drive	2991	749.74
8 Craft Avenue	1008	217.7
22 Beechwood Ct.	3202	635.7

Figure 2.1 Simple Regression Data – first 10 rows

Appendix 3 shows how to create a simple regression model using Excel **Data Analysis** ToolPak

- **Interpret the Output Generated by The Regression Module**

An output worksheet has been generated, named as “**Simple Regression**”, we will interpret the meaning of the output section by section starting from **Summary Output.**



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.828192611
R Square	0.685903002
Adjusted R Square	0.674685252
Standard Error	68.66500178
Observations	30

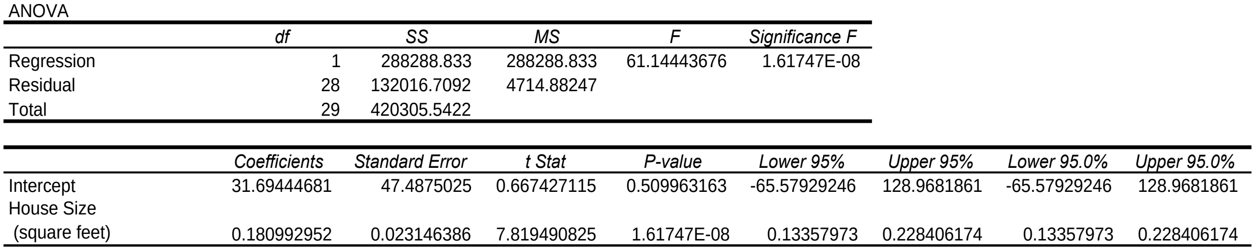
Figure 2.2 Simple Regression Summary Output

- **R-Square** is the first measure you should look at. It measures how close the data are to the fitted regression line. **R Square** ranges from 0 to 1. An R-square of 1 indicates that the regression line perfectly fits the data.
- Note: **R-Square measures the percentage of variation in Y (house prices) that is explained by the variation in House Size. Figure 1.2 shows that 68.5% of the variation in prices is explained by the variation in house size (sq. ft.).**
- Adjusted R Square is a measure that is more relevant when you create a regression model with more than one independent variable and want to compare models (explained in the next section.)
- **Multiple R in Figure 2.2**: if you calculate the correlation coefficient between the Predicted house values (Figure 2.5) and actual appraised values (in Figure 2.2), you will see that the correlation coefficient is exactly equal to the Multiple R in Fig. 2.2, which is the positive square root of **R-Square**.
- **Standard Error** is a measure of the error in prediction (In this example, the error in predicting the house price, given house size in sq. ft.). In general, a strong regression model

is expected to have a high R Square and low standard error. The standard error is the standard deviation of the residuals (see Fig 2.5) after accounting for the fact that **two parameter estimates** are also involved.

- **Observations** is the number of records used to create the regression model**.**

Analysis of Variance:



ANOVA

	df	SS	MS	F	Significance F
Regression	1	288288.833	288288.833	61.14443676	1.61747E-08
Residual	28	132016.7092	4714.88247		
Total	29	420305.5422			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	31.69444681	47.4875025	0.667427115	0.509963163	-65.57929246	128.9681861	-65.57929246	128.9681861
House Size (square feet)	0.180992952	0.023146386	7.819490825	1.61747E-08	0.13357973	0.228406174	0.13357973	0.228406174

Figure 2.3 Simple Regression ANOVA Table

- The ANOVA (Analysis of variance) table shows a number of statistical estimates such as SS (sum of squares), significance F, t Stats, P-value, etc.
- The estimated linear relation: **Price = 31.69 + 0.18 * House_size**
- The above Linear relation is estimated based on a random sample of 30 houses from a population of houses

Q: Does this relationship apply to all houses in the population? In other words, can this model be used to predict the price of other houses in the population (given sq. ft.)? We need to run a **statistical test of Hypothesis**. Often, this question is also stated as

"Is the estimated linear relationship statistically significant?"

Test of Hypothesis

H_0: There is No Linear relation between Price and Size in the Population of all houses.

H_1: There is a linear relation in the population between Price and Sq. ft.

$\mathbf{H_0}$ can be restated as "The slope of the linear relation in the population = 0"

$\mathbf{H_1}$ can be restated as "the slope of the linear relation in the population is not equal to 0"

$$H_0: \quad \beta_1 = 0$$
$$H_1: \quad \beta_1 \neq 0$$

If we can reject H_0 with a high confidence, then we can conclude that $\mathbf{H_1}$ holds. The **F** test is used for this. The **F** value is already calculated and given to us. If Prob (F >= 61.44 | $\mathbf{H_0}$) < 0.05, we can Reject $\mathbf{H_0}$ with 95% or more confidence.

In the above regression model, since Significance F (F-test) is less than 0.05, **with more than 95% confidence, we can reject Null hypothesis that there is No Linear relation between house**

size and price. This is good. So, house size has a role in predicting house price, and the estimated model is useful. The predictions, of course, will have errors and we have an idea of the size of those errors (standard error in Fig. 2.2)

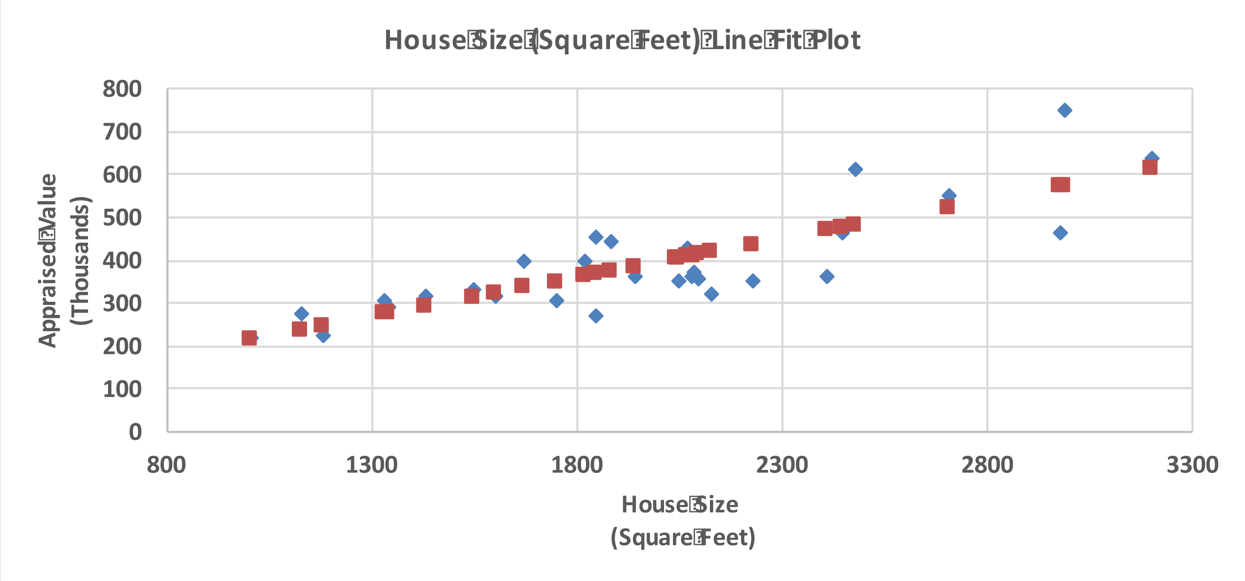
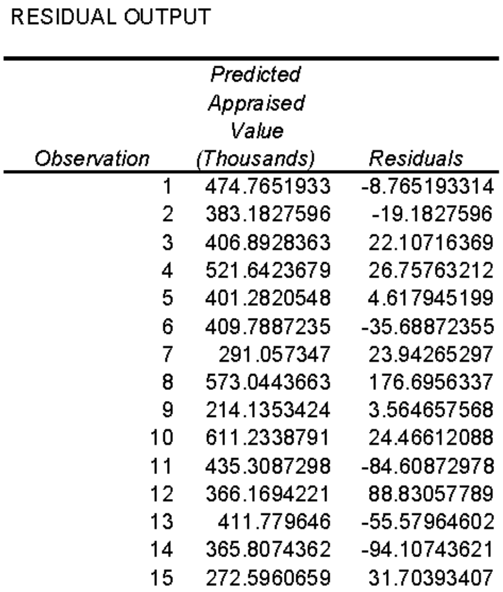


Figure 2.4 Simple Regression: scatter Plot with fitted Regression line

Figure 2.4 demonstrates how the predicted house price (Red Square) fits the actual house price (Blue Diamond). Again, based on the R-square = 0.685, 68.5% of the variation in house prices is explained by the variation in house sq. ft. This should not be a surprise since house price depends on many factors besides the square feet of the house.

Sample Residuals output from Excel



RESIDUAL OUTPUT

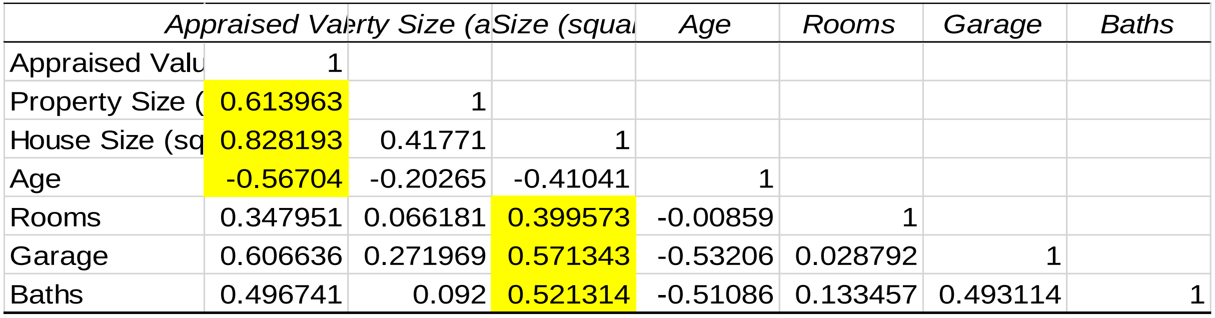
Observation	*Predicted Appraised Value (Thousands)*	*Residuals*
1	474.7651933	-8.765193314
2	383.1827596	-19.1827596
3	406.8928363	22.10716369
4	521.6423679	26.75763212
5	401.2820548	4.617945199
6	409.7887235	-35.68872355
7	291.057347	23.94265297
8	573.0443663	176.6956337
9	214.1353424	3.564657568
10	611.2338791	24.46612088
11	435.3087298	-84.60872978
12	366.1694221	88.83057789
13	411.779646	-55.57964602
14	365.8074362	-94.10743621
15	272.5960659	31.70393407

The residual output table shows the difference between predicted house price and the actual house price. The smaller the residuals, better the regression model.

3. Multiple Regression

If a regression model contains more than one independent variable, then the regression is known as **Multiple Regression**. (As opposed to **Simple Regression**). Before conducting multiple regression analysis, again, the first step is to use correlation analysis to inspect the correlation (linear association) between the dependent variable and the independent variables. In this example, we want to see the correlation between the dependent variable, **house price**, and the independent variables, **property size** (acres), **house size** (sq. ft.), **age**, **rooms**, **baths**, and **garages**. The closer the correlation coefficient is to 1 or -1, the stronger the linear relationship is. The correlations shown below are **pair-wise** correlations.

Interpreting the Output Generated by the Correlation Analysis



	Appraised Va	erty Size (a	Size (squa	Age	Rooms	Garage	Baths
Appraised Valu	1						
Property Size (	0.613963	1					
House Size (sq	0.828193	0.41771	1				
Age	-0.56704	-0.20265	-0.41041	1			
Rooms	0.347951	0.066181	0.399573	-0.00859	1		
Garage	0.606636	0.271969	0.571343	-0.53206	0.028792	1	
Baths	0.496741	0.092	0.521314	-0.51086	0.133457	0.493114	1

Figure 3.0 Correlations Analysis (Multiple Explanatory Variables)

Observations from the Correlation Analysis:

- **House Price** is strongly correlated with **property size, House Size, Garages, and Age** (Age is negatively correlated to House Price).
- House Size is also positively correlated with **#Rooms**, **#Garages** and **#Baths**. So larger houses tend to have more rooms, garages, as well as more bath rooms.
- If two independent variables are strongly correlated (to each other), then you may not need both of them in the model. Such strong correlations can also lead to problems in model building.

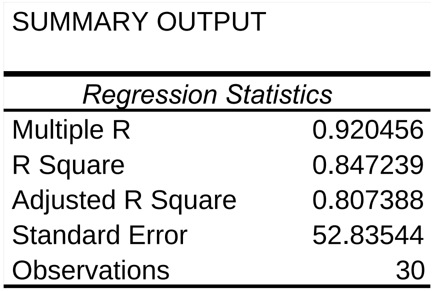
Creating and evaluating the Multiple Regression Model

Appendix 3 shows how to create a multiple regression model using the **Data Analysis** ToolPack

First build a multiple regression model with all the independent variables available in the dataset. The output is stored in the worksheet named "**Multiple Regression 6 indep var**", we will interpret the meaning of the output, section by section starting from **Summary Output.**

R square has a weakness. As you keep adding more independent variables, **R Square** never decreases, even if the additional variables do not have any explanatory power.

Adjusted R-Square adjusts **R-square** based on the number of independent variables in the model. In Multiple Regression, we should also look at **Adjusted R square**. If you add a variable that is irrelevant, Adjusted R Square goes down.

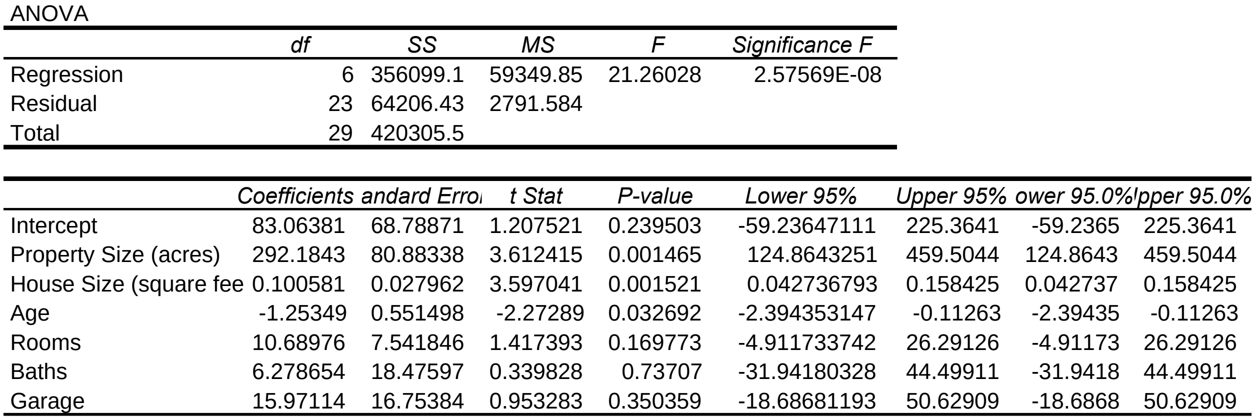


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.920456
R Square	0.847239
Adjusted R Square	0.807388
Standard Error	52.83544
Observations	30

Figure 3.1 Multiple Regression Summary Output

- First, note that R square is 0.847, which means **84.7%** of the variation in house prices is explained by the variations in all the independent variables put together. Note that this is considerably higher than the **67.8%** obtained in the simple regression. Hence, all the additional independent variables have helped build a stronger model.
- The standard error has reduced from **68.665** (in simple regression) to **52.835**.
- In the ANOVA table below, significance F < 0.05. See comments after Figure 3.2. Hence, the **multiple regression model is a stronger model** both in terms of **R Square** and in terms of **standard error**.



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	356099.1	59349.85	21.26028	2.57569E-08
Residual	23	64206.43	2791.584		
Total	29	420305.5			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	83.06381	68.78871	1.207521	0.239503	-59.23647111	225.3641	-59.2365	225.3641
Property Size (acres)	292.1843	80.88338	3.612415	0.001465	124.8643251	459.5044	124.8643	459.5044
House Size (square fee	0.100581	0.027962	3.597041	0.001521	0.042736793	0.158425	0.042737	0.158425
Age	-1.25349	0.551498	-2.27289	0.032692	-2.394353147	-0.11263	-2.39435	-0.11263
Rooms	10.68976	7.541846	1.417393	0.169773	-4.911733742	26.29126	-4.91173	26.29126
Baths	6.278654	18.47597	0.339828	0.73707	-31.94180328	44.49911	-31.9418	44.49911
Garage	15.97114	16.75384	0.953283	0.350359	-18.68681193	50.62909	-18.6868	50.62909

Figure 3.2 Multiple Regression ANOVA Table

ANOVA table gives an overall **F** value and displays **P-value** associated with the F.

H_0: There is No Linear relation between Price and all the independent variables.

H_1: There is a linear relation in the population between the independent variables and Price

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1: $\beta_j \neq 0$, for at least one value of j

Now, since significance F is smaller than 0.05, **with more than 95% confidence we can reject the Null Hypothesis that there is no linear relation between house prices and all the independent variables.** Hence the above multiple regression model is **statistically significant**

Since the multiple Regression model is a stronger model with more explanatory power and has smaller predictive errors, we will choose the multiple regression model and discard the simple regression model.

Conclusion: A strong multiple regression model was built with 6 independent (explanatory) variables. **We can use the model to predict house prices** given the values of the independent variables. However, an important question remains.

Are all the 6 independent variables necessary in the model?

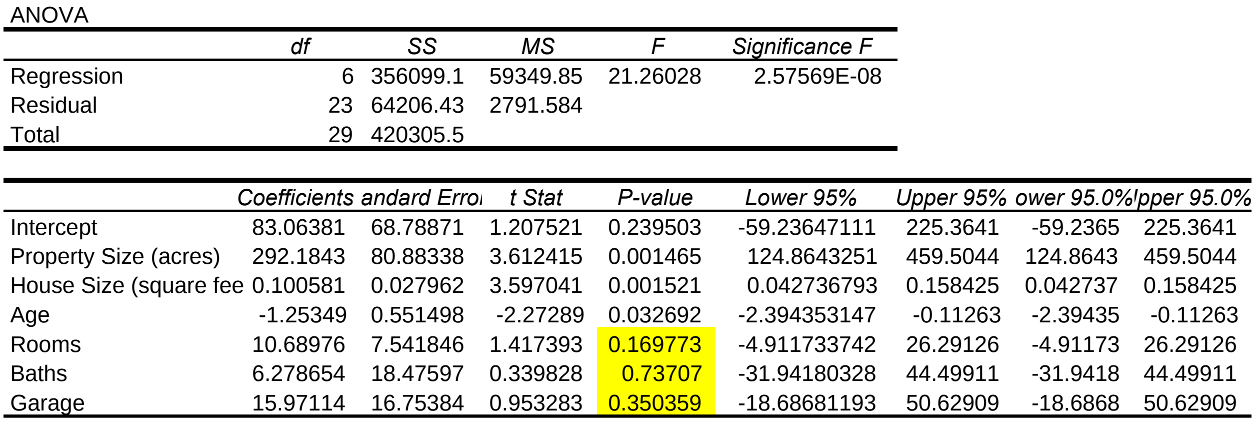
Stated another way, Can we build an equally (approximately) strong model with **less number of independent variables?** Our aim is to build model with as few independent variables as possible.
The backward elimination method (next section) helps in identifying which independent variables are needed in the model and hence which independent variables can be dropped from the model.

Just like the **F**-test to determine the statistical significance of the overall model, we will conduct a series of **t**-tests to check if each of the independent variables are needed or not. The value of the **t** statistic and the corresponding **P**-value are already present in **figure 3.2** in columns 4, 5.

4. Backward Elimination in Regression

- **Interpret the Output Generated by The Regression Module**

First of all, we will look at the "**Multiple Regression 6 indep var**" sheet.



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	356099.1	59349.85	21.26028	2.57569E-08
Residual	23	64206.43	2791.584		
Total	29	420305.5			

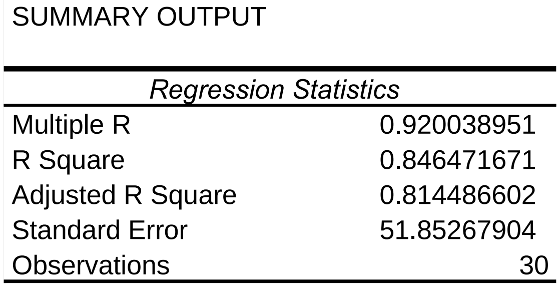
	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	83.06381	68.78871	1.207521	0.239503	-59.23647111	225.3641	-59.2365	225.3641
Property Size (acres)	292.1843	80.88338	3.612415	0.001465	124.8643251	459.5044	124.8643	459.5044
House Size (square fee	0.100581	0.027962	3.597041	0.001521	0.042736793	0.158425	0.042737	0.158425
Age	-1.25349	0.551498	-2.27289	0.032692	-2.394353147	-0.11263	-2.39435	-0.11263
Rooms	10.68976	7.541846	1.417393	0.169773	-4.911733742	26.29126	-4.91173	26.29126
Baths	6.278654	18.47597	0.339828	0.73707	-31.94180328	44.49911	-31.9418	44.49911
Garage	15.97114	16.75384	0.953283	0.350359	-18.68681193	50.62909	-18.6868	50.62909

Figure 4.1 Backward Elimination Regression ANOVA Table – 6 independent variables

A high P-values (> 0.05) associated with the **t** statistic is an indication that this independent variable **may not be needed** in the regression model. We will rebuild the model by eliminating one independent variable. Although 3 independent variables are candidates for elimination, choose the one with the highest p-value. **Note: Drop only one variable at a time and check Significance F, R square and Adjusted R-Square.**

From correlation analysis, you saw that a house with a large Size (sq. ft.) tends have more rooms, more garages, bathrooms. So possibly these variables are not adding that much more information. **We should drop one variable at a time.** Therefore, let us drop #Baths, which has the highest P-value.

Run the regression model again with only 5 independent variables to get the following output.



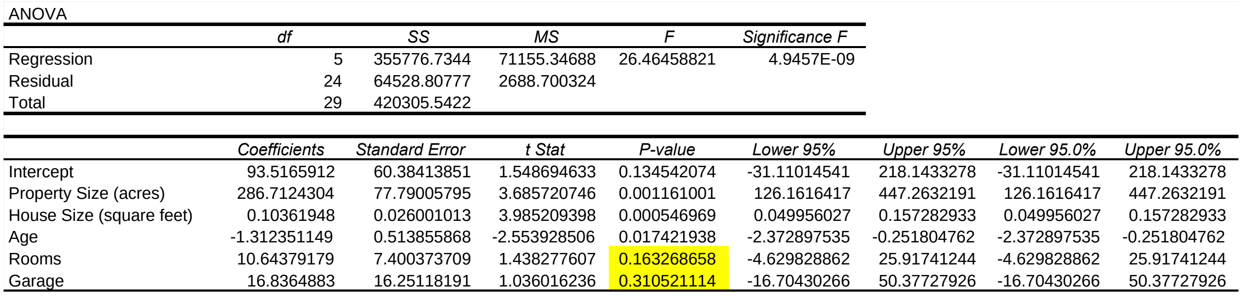
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.920038951
R Square	0.846471671
Adjusted R Square	0.814486602
Standard Error	51.85267904
Observations	30

Figure 4.2 Backward Elimination Regression Summary – 5 independent variables

After dropping **#Baths** and rebuilding the model, we see that

- In choosing between two models, it is helpful to compare the adjusted R square. The model with 5 independent variables has higher Adjusted R square (81.4% vs. 80.7%). This confirms our suspicion that #baths is not needed in the model, given the other independent variables in the model. If the Adjusted R-Square decreases then do not remove the variable from the model
- Standard Error has also improved
- **Conclusion**: Given all the other independent Variables in the model, **#baths** is not adding to the model's predictive power. Hence, we can drop **#Baths** from the model.



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	355776.7344	71155.34688	26.46458821	4.9457E-09
Residual	24	64528.80777	2688.700324		
Total	29	420305.5422			

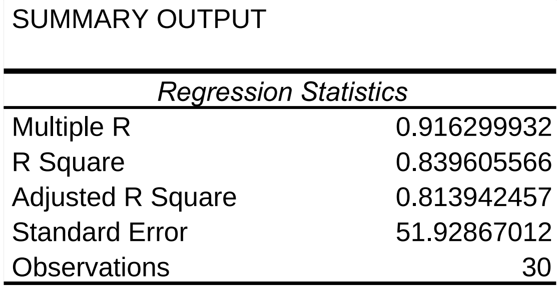
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	93.5165912	60.38413851	1.548694633	0.134542074	-31.11014541	218.1433278	-31.11014541	218.1433278
Property Size (acres)	286.7124304	77.79005795	3.685720746	0.001161001	126.1616417	447.2632191	126.1616417	447.2632191
House Size (square feet)	0.10361948	0.026001013	3.985209398	0.000546969	0.049956027	0.157282933	0.049956027	0.157282933
Age	-1.312351149	0.513855868	-2.553928506	0.017421938	-2.372897535	-0.251804762	-2.372897535	-0.251804762
Rooms	10.64379179	7.400373709	1.438277607	0.163268658	-4.629828862	25.91741244	-4.629828862	25.91741244
Garage	16.8364883	16.25118191	1.036016236	0.310521114	-16.70430266	50.37727926	-16.70430266	50.37727926

Figure 4.3 Backward Elimination Regression ANOVA table – 5 independent variables

Significance F is still really small. **With more than 95% confidence we can reject the Null Hypothesis that there is no linear relation between house prices and the independent variables.** This is good. So we can use the model to predict house prices given the values of the five independent variables above.

However, since #**garages'** P-value is the highest and higher than 0.05, we can check to see if **#garages** is needed in the model or not. As before, compare the two models, one with **#garages** in the model and another without **#garages**

Run the regression model again with 4 independent variables to get the following output.

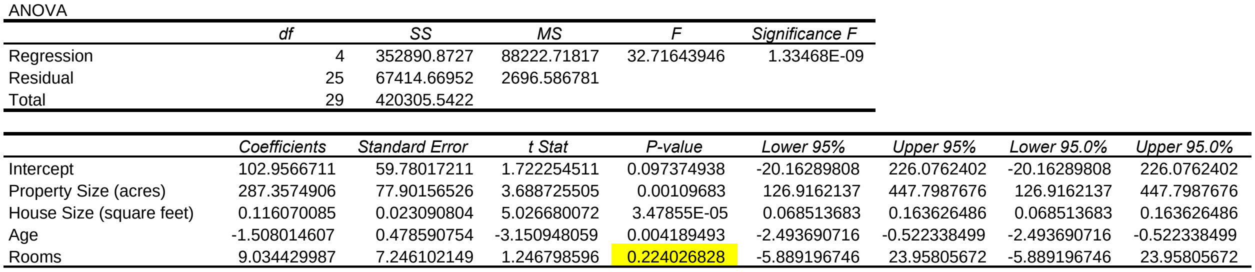


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.916299932
R Square	0.839605566
Adjusted R Square	0.813942457
Standard Error	51.92867012
Observations	30

Figure 4.4 Backward Elimination Regression Summary – 4 independent variables

Again, after dropping #Garages and #Baths, **Adjusted R Square is 81.396%,** reduced slightly from 81.4%, but not much. It indicates that 81.3% of the variation in house prices can be explained by the variation in these 4 independent variables.



ANOVA

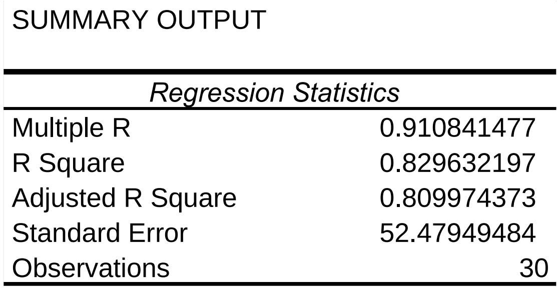
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	352890.8727	88222.71817	32.71643946	1.33468E-09
Residual	25	67414.66952	2696.586781		
Total	29	420305.5422			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	102.9566711	59.78017211	1.722254511	0.097374938	-20.16289808	226.0762402	-20.16289808	226.0762402
Property Size (acres)	287.3574906	77.90156526	3.688725505	0.00109683	126.9162137	447.7987676	126.9162137	447.7987676
House Size (square feet)	0.116070085	0.023090804	5.026680072	3.47855E-05	0.068513683	0.163626486	0.068513683	0.163626486
Age	-1.508014607	0.478590754	-3.150948059	0.004189493	-2.493690716	-0.522338499	-2.493690716	-0.522338499
Rooms	9.034429987	7.246102149	1.246798596	0.224026828	-5.889196746	23.95805672	-5.889196746	23.95805672

Figure 4.5 Backward Elimination Regression ANOVA table – 4 independent variables

Significance F is still really small (< 0.05). **With more than 95% confidence we can reject the Null Hypothesis that there is no linear relation between house prices and the independent variables.** This is good. We can use the model to predict house prices given the values of the four independent variables in the model. The P-value of **#Rooms** in the house remains higher than 0.05. Next, test to see if this variable can be dropped from the model without affecting the predictive power of the model.

Run the regression model again with only 3 independent variables to get the following output.

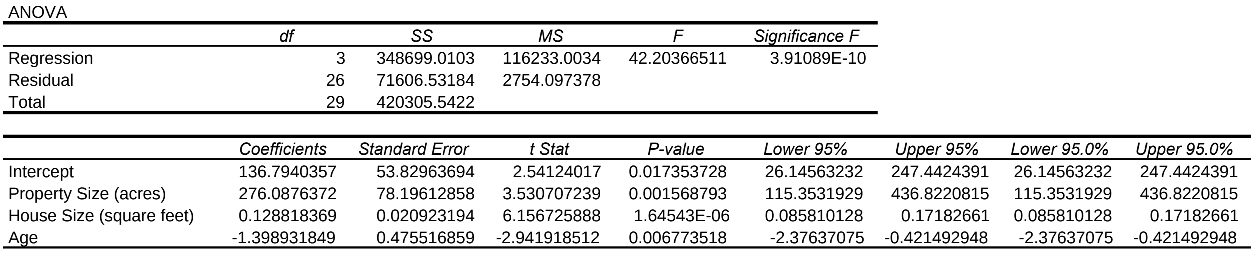


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.910841477
R Square	0.829632197
Adjusted R Square	0.809974373
Standard Error	52.47949484
Observations	30

Figure 4.6 Backward Elimination Regression Summary – 3 independent variables

After dropping #Bath rooms, #Garages and #Rooms, the Adjusted R Square now is 81%.



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	348699.0103	116233.0034	42.20366511	3.91089E-10
Residual	26	71606.53184	2754.097378		
Total	29	420305.5422			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	136.7940357	53.82963694	2.54124017	0.017353728	26.14563232	247.4424391	26.14563232	247.4424391
Property Size (acres)	276.0876372	78.19612858	3.530707239	0.001568793	115.3531929	436.8220815	115.3531929	436.8220815
House Size (square feet)	0.128818369	0.020923194	6.156725888	1.64543E-06	0.085810128	0.17182661	0.085810128	0.17182661
Age	-1.398931849	0.475516859	-2.941918512	0.006773518	-2.37637075	-0.421492948	-2.37637075	-0.421492948

Figure 4.7 Backward Elimination Regression ANOVA table – 3 independent variables

Significance F is smaller than 0.05, **with more than 95% confidence we can reject the Null Hypothesis that there is no linear relation between house prices and the independent variables.** Since the P-value associated with **all the 3 remaining independent variables are less than 0.05**, we arrived at the **final model**.

Note that you may want to keep **#rooms** in the model. Because if you take it out of the model, **Adjusted R-Square** decreased slightly, and standard error went up. On the other hand, it may also be OK to drop **#rooms** with the justification that the drop in **Adjusted R-Square** is very small.

Not all statisticians recommend backward elimination. Even among those who recommend, there are variations. Some recommend eliminating independent variables (one by one) with p value > 0.1. We will discuss this in class.

Theory and some Definitions

1. **Dependent Variable** The variable y being investigated. In SPSS Modeler, it is called **Target**
2. **Independent Variable** Any of the variables $x_1, x_2, \dots, x_k$ that might affect the values of the dependent variable. Other names for the independent variable are **predictor**, **regressor**, and **explanatory** variable. A continuous independent variable is called a **covariate**; a discrete independent variable is also called a **factor** or a **grouping variable**. A **dummy variable** is an independent variable that can only take on the values 0 and 1.
3. **Linear Regression Equation** An equation of the following form that could predict the value of the dependent variable y from the values of the dependent variables $x_1, x_2, \dots, x_k$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

4. The **coefficients** of the regression model are $\beta_0, \beta_1, \dots, \beta_k$

5. The **Linear Regression Model** shows the relationship between the **actual dependent variable values**, $y_1, y_2, \dots, y_n$, the **coefficients** $\beta_0, \beta_1, \dots, \beta_k$, the **independent variables** x_{ij}, $i = 1, \dots, n$, $j = 0, \dots, k$, and the random errors or **residuals** $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ik} + \varepsilon_i$$

- In practice, the coefficients of the regression model and the residuals are unknown. The coefficients are estimated from a sample, with a procedure known as obtaining a least squares estimation (LSE).
- It is worth re-emphasizing that we are interested in the relationship between the independent variable and the dependent variables in the **population**, which is unknown. Hence we try to estimate the relation from a sample. (Assumed to be a random representative sample from the population). The least squares estimation process gives us estimates of the regression coefficients $\boldsymbol{\beta_0, \beta_1, \dots, \beta_k}$
- We shall denote these estimates obtained from the sample as $\mathbf{b_0, b_1, b_2, \dots\dots., b_k}$
- The regression equation estimated from the sample looks like

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots\dots\dots\dots\dots\dots + b_k x_k$$

- **ASSUMPTIONS**: The estimation process is based on the following assumptions (*acronym*: **LINE**)
  - **LINEARITY**: The relationship between the independent variable and the dependent variables in the population is Linear. Hence we choose the linear form shown above.
  - **INDEPENDENCE**: The errors are statistically independent of each other.
  - **NORMALITY**: The Errors are **normally** distributed.
  - **EQUAL VARIANCE**: The probability distribution of the errors has constant variance

Appendix 1 - How to Add the Data Analysis ToolPak

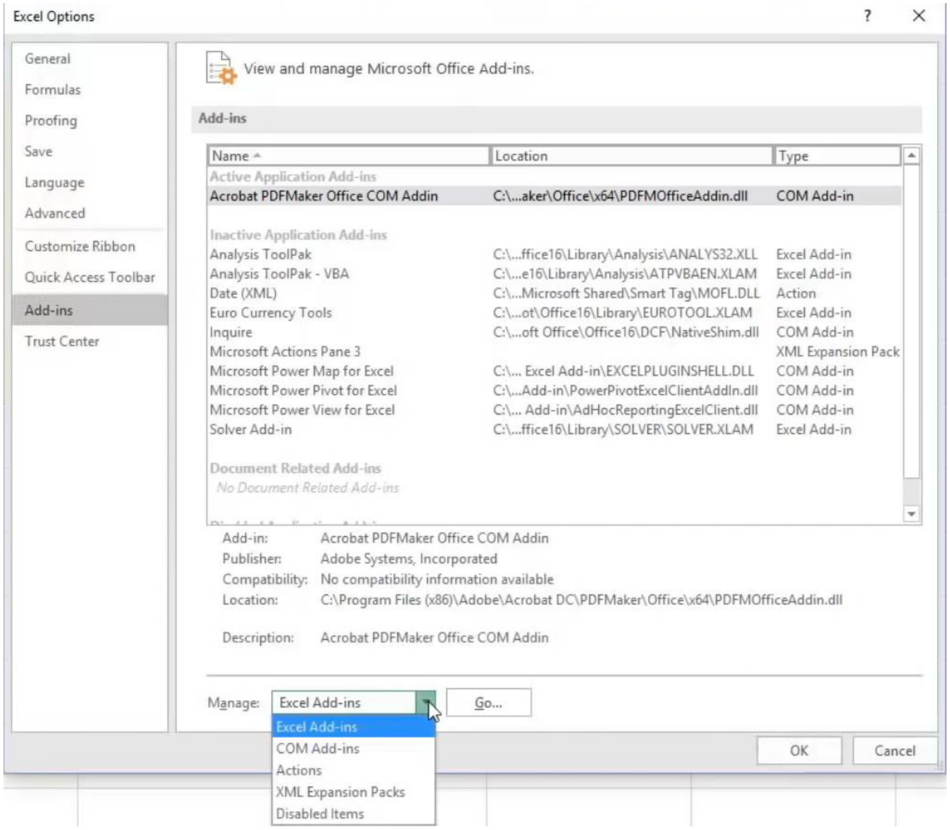
If you already have the Data Analysis ToolPak Installed in your Excel, you can skip this step.

1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.

   If you're using Excel 2007, click the **Microsoft Office Button** , and then click **Excel Options**

2. In the **Manage** box, select **Excel Add-ins** and then click **Go**.

   If you're using Excel for Mac, in the file menu go to **Tools** > **Excel Add-ins**.



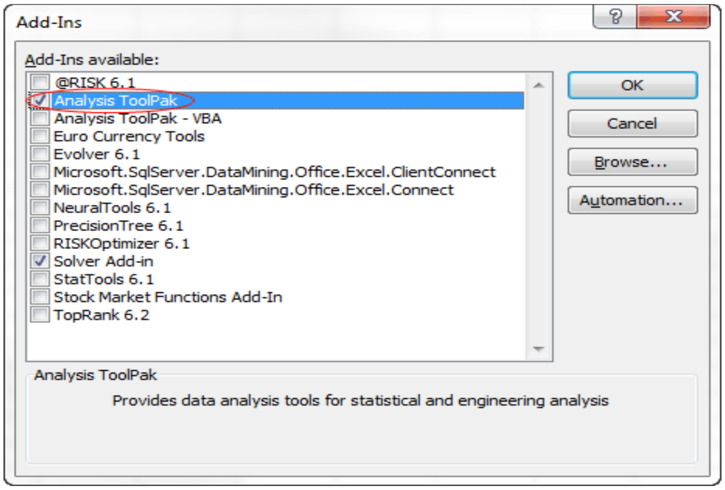
View and manage Microsoft Office Add-ins.

Name ▲	Location	Type
Active Application Add-ins		
Acrobat PDFMaker Office COM Addin	C:\...aker\Office\x64\PDFMOfficeAddin.dll	COM Add-in
Inactive Application Add-ins		
Analysis ToolPak	C:\...ffice16\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...e16\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Date (XML)	C:\...Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...ot\Office16\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\...oft Office\Office16\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Power Map for Excel	C:\... Excel Add-in\EXCELPLUGINSHELL.DLL	COM Add-in
Microsoft Power Pivot for Excel	C:\...Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Microsoft Power View for Excel	C:\... Add-in\AdHocReportingExcelClient.dll	COM Add-in
Solver Add-in	C:\...ffice16\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		

Add-in: Acrobat PDFMaker Office COM Addin
Publisher: Adobe Systems, Incorporated
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Adobe\Acrobat DC\PDFMaker\Office\x64\PDFMOfficeAddin.dll

Description: Acrobat PDFMaker Office COM Addin

3. In the **Add-Ins** box, check the **Analysis ToolPak** check box, and then click **OK**.



Add-Ins

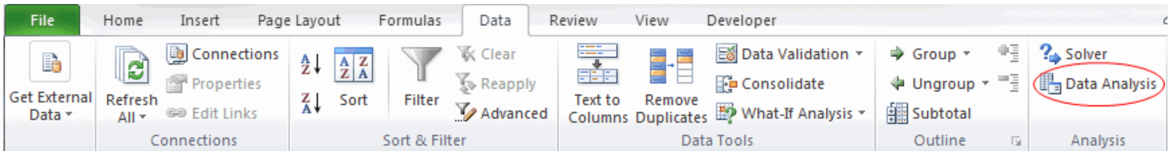
Add-Ins available:

- ☐ @RISK 6.1
- ☑ Analysis ToolPak
- ☐ Analysis ToolPak - VBA
- ☐ Euro Currency Tools
- ☐ Evolver 6.1
- ☐ Microsoft.SqlServer.DataMining.Office.Excel.ClientConnect
- ☐ Microsoft.SqlServer.DataMining.Office.Excel.Connect
- ☐ NeuralTools 6.1
- ☐ PrecisionTree 6.1
- ☐ RISKOptimizer 6.1
- ☑ Solver Add-in
- ☐ StatTools 6.1
- ☐ Stock Market Functions Add-In
- ☐ TopRank 6.2

Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

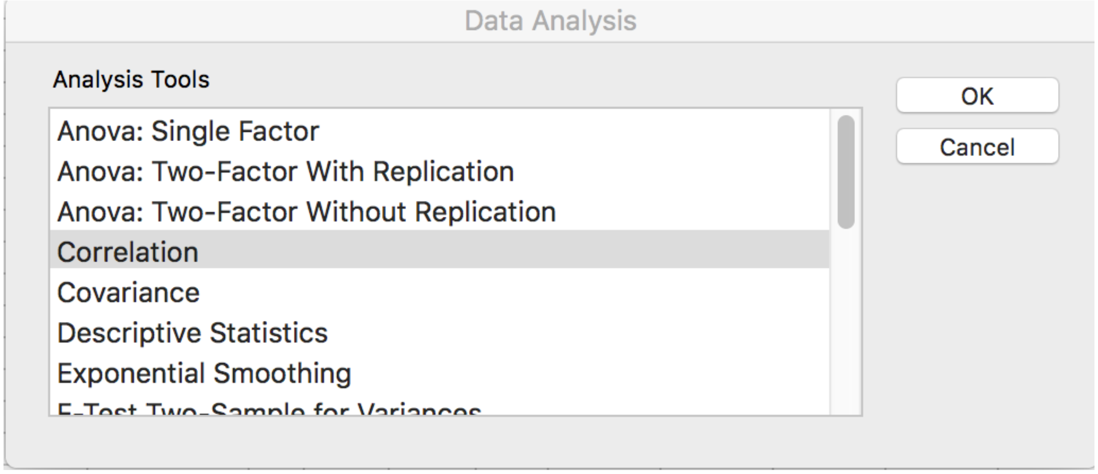
4. Once **Analysis ToolPak** is loaded, you will see a **Data Analysis** option on the **Data ribbon**.



Appendix 2 – Correlation Analysis with Data Analysis in Excel.

1. For Simple Explanatory Variable

Navigate to **Data Ribbon**, click **Data Analysis,** choose **Correlation,** and click OK.

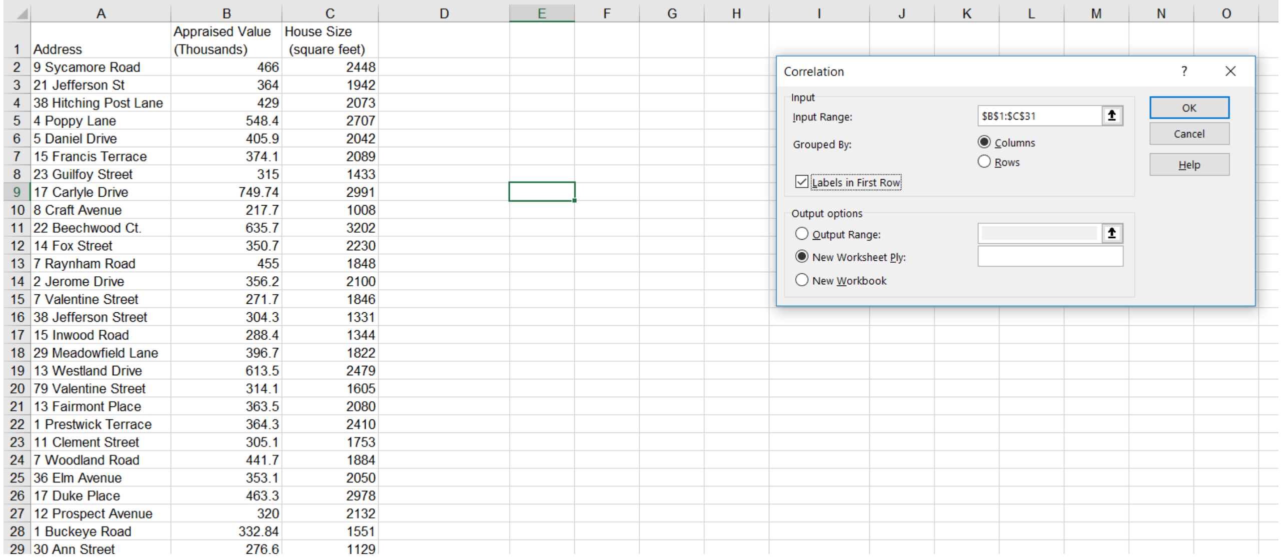


Data Analysis

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances

We will choose the details as follows:



	A	B	C
1	Address	Appraised Value (Thousands)	House Size (square feet)
2	9 Sycamore Road	466	2448
3	21 Jefferson St	364	1942
4	38 Hitching Post Lane	429	2073
5	4 Poppy Lane	548.4	2707
6	5 Daniel Drive	405.9	2042
7	15 Francis Terrace	374.1	2089
8	23 Guilfoy Street	315	1433
9	17 Carlyle Drive	749.74	2991
10	8 Craft Avenue	217.7	1008
11	22 Beechwood Ct.	635.7	3202
12	14 Fox Street	350.7	2230
13	7 Raynham Road	455	1848
14	2 Jerome Drive	356.2	2100
15	7 Valentine Street	271.7	1846
16	38 Jefferson Street	304.3	1331
17	15 Inwood Road	288.4	1344
18	29 Meadowfield Lane	396.7	1822
19	13 Westland Drive	613.5	2479
20	79 Valentine Street	314.1	1605
21	13 Fairmont Place	363.5	2080
22	1 Prestwick Terrace	364.3	2410
23	11 Clement Street	305.1	1753
24	7 Woodland Road	441.7	1884
25	36 Elm Avenue	353.1	2050
26	17 Duke Place	463.3	2978
27	12 Prospect Avenue	320	2132
28	1 Buckeye Road	332.84	1551
29	30 Ann Street	276.6	1129

Correlation

Input

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

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

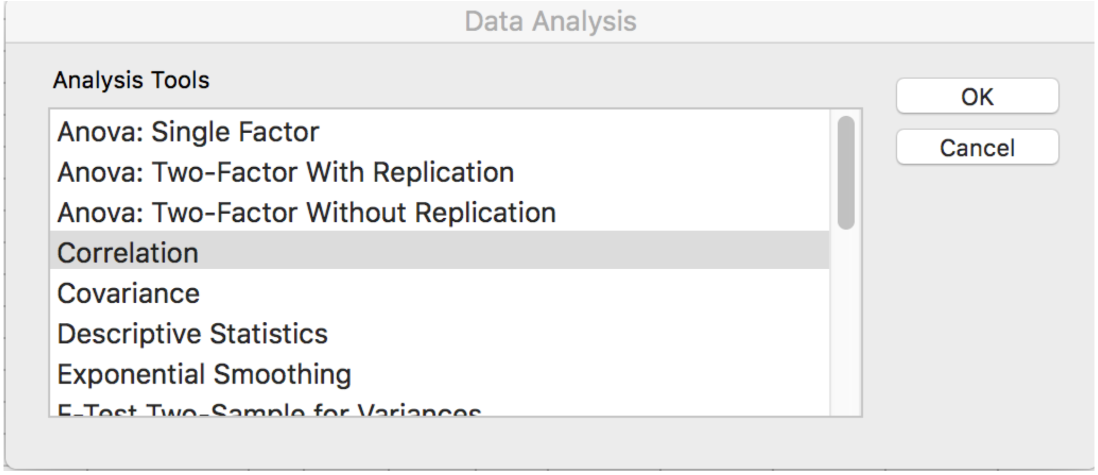
○ New Workbook

Click ↥ in **Input Range** and Choose **B1:C31**

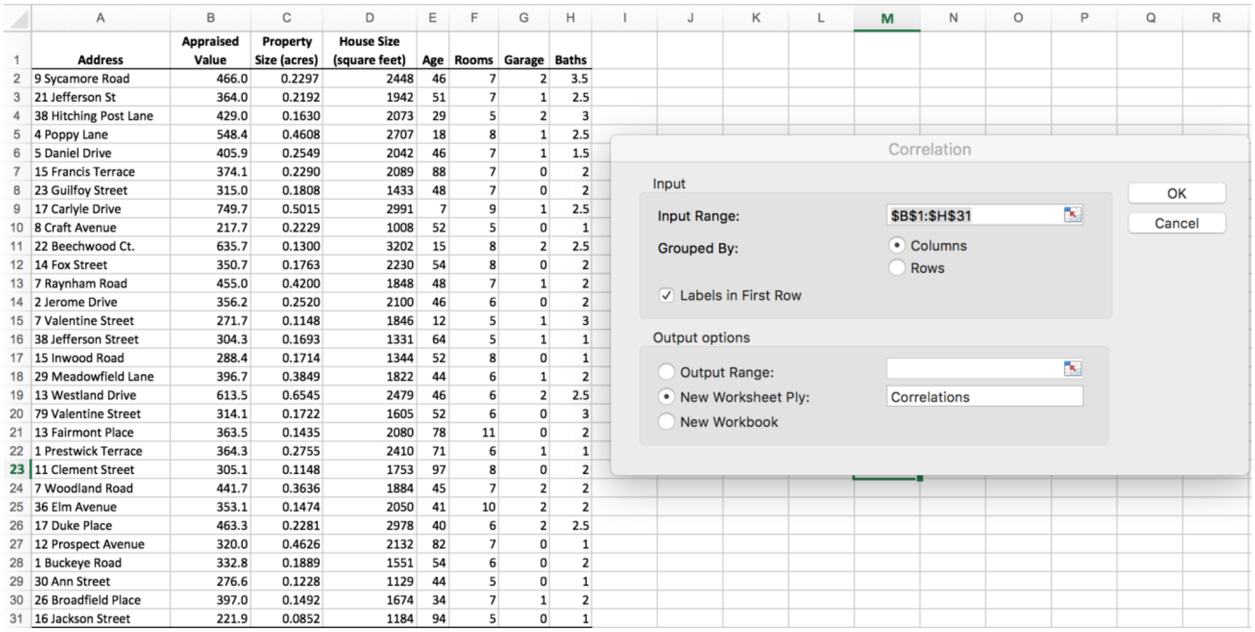
Select **Columns** in **Grouped By** and specify the output worksheet name in **New Worksheet Ply**

2. For Multiple Explanatory Variables

Navigate to **Data Ribbon**, click **Data Analysis,** choose **Correlation,** and click OK.



We will choose the details as follows:



	A	B	C	D	E	F	G	H
1	Address	Appraised Value	Property Size (acres)	House Size (square feet)	Age	Rooms	Garage	Baths
2	9 Sycamore Road	466.0	0.2297	2448	46	7	2	3.5
3	21 Jefferson St	364.0	0.2192	1942	51	7	1	2.5
4	38 Hitching Post Lane	429.0	0.1630	2073	29	5	2	3
5	4 Poppy Lane	548.4	0.4608	2707	18	8	1	2.5
6	5 Daniel Drive	405.9	0.2549	2042	46	7	1	1.5
7	15 Francis Terrace	374.1	0.2290	2089	88	7	0	2
8	23 Guilfoy Street	315.0	0.1808	1433	48	7	0	2
9	17 Carlyle Drive	749.7	0.5015	2991	7	9	1	2.5
10	8 Craft Avenue	217.7	0.2229	1008	52	5	0	1
11	22 Beechwood Ct.	635.7	0.1300	3202	15	8	2	2.5
12	14 Fox Street	350.7	0.1763	2230	54	8	0	2
13	7 Raynham Road	455.0	0.4200	1848	48	7	1	2
14	2 Jerome Drive	356.2	0.2520	2100	46	6	0	2
15	7 Valentine Street	271.7	0.1148	1846	12	5	1	3
16	38 Jefferson Street	304.3	0.1693	1331	64	5	1	1
17	15 Inwood Road	288.4	0.1714	1344	52	8	0	1
18	29 Meadowfield Lane	396.7	0.3849	1822	44	6	1	2
19	13 Westland Drive	613.5	0.6545	2479	46	6	2	2.5
20	79 Valentine Street	314.1	0.1722	1605	52	6	0	3
21	13 Fairmont Place	363.5	0.1435	2080	78	11	0	2
22	1 Prestwick Terrace	364.3	0.2755	2410	71	6	1	1
23	11 Clement Street	305.1	0.1148	1753	97	8	0	2
24	7 Woodland Road	441.7	0.3636	1884	45	7	2	2
25	36 Elm Avenue	353.1	0.1474	2050	41	10	2	2
26	17 Duke Place	463.3	0.2281	2978	40	6	2	2.5
27	12 Prospect Avenue	320.0	0.4626	2132	82	7	0	1
28	1 Buckeye Road	332.8	0.1889	1551	54	6	0	2
29	30 Ann Street	276.6	0.1228	1129	44	5	0	1
30	26 Broadfield Place	397.0	0.1492	1674	34	7	1	2
31	16 Jackson Street	221.9	0.0852	1184	94	5	0	1

Correlation

Input

Input Range: $B$1:$H$31

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply: Correlations

○ New Workbook

Click in **Input Range** and Choose **B1:H31**

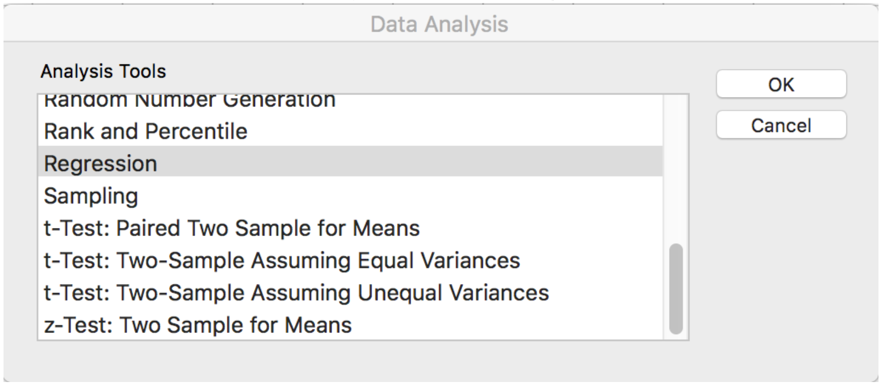
Select **Columns** in **Grouped By** and specify the output worksheet name in **New Worksheet Ply**

Appendix 3 – Regression model with Data Analysis ToolPak in Excel

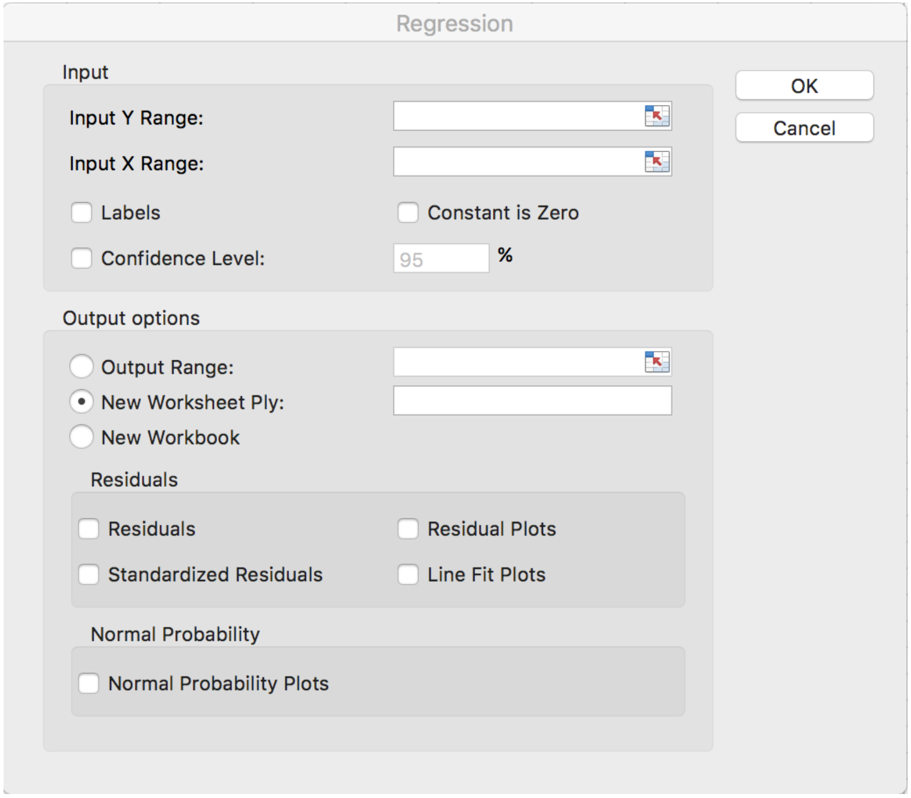
1. For Simple Regression Analysis

In this part, we will run a simple regression analysis between **two variables**: house prices (**Appraised Value in Thousands**) and **house sizes in square feet.**

Navigate to **Data Ribbon**, click **Data Analysis,** choose **Regression,** and click OK.


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

A window will pop up as shown below, it is the **Regression Module** that we mainly use to conduct all analysis in this workshop.



Input Fields:

Input Y Range: Dependent variables data range in Excel
Input X Range: Independent variables data range in Excel (Can be multiple columns)

Choices of **Labels**: whether the data include label in the first row
Constant is Zero: Force the regression line through the origin (intercept = 0)
Confidence Level: Input the confidence Level

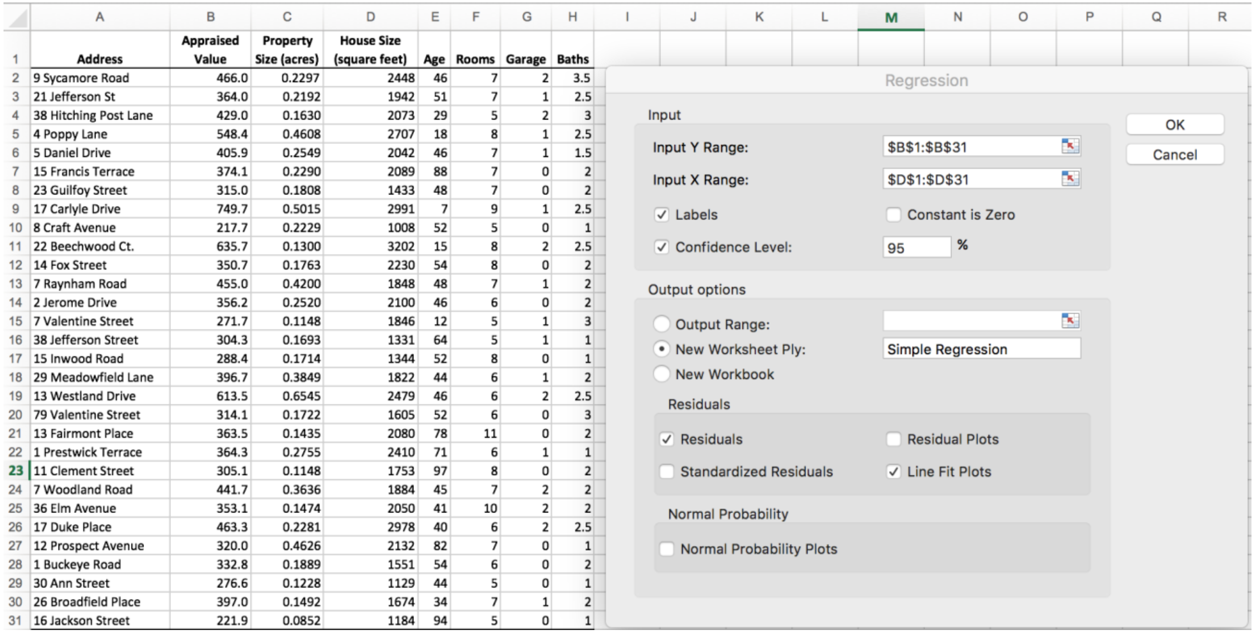
Output Fields:

Output Range: The output will be in the same excel sheet within the range specified
New Worksheet Ply: The output will be in a new sheet and named by the following input
New Workbook: The output will be in a new excel file and named by the following input

Residuals:

Choices of whether the output includes "**Residuals**", "**Residual Plots**", "**Standardized Residuals**", "**Line Fit Plots**".

For the simple regression analysis, we will select as follows:



	A	B	C	D	E	F	G	H
1	Address	Appraised Value	Property Size (acres)	House Size (square feet)	Age	Rooms	Garage	Baths
2	9 Sycamore Road	466.0	0.2297	2448	46	7	2	3.5
3	21 Jefferson St	364.0	0.2192	1942	51	7	1	2.5
4	38 Hitching Post Lane	429.0	0.1630	2073	29	5	2	3
5	4 Poppy Lane	548.4	0.4608	2707	18	8	1	2.5
6	5 Daniel Drive	405.9	0.2549	2042	46	7	1	1.5
7	15 Francis Terrace	374.1	0.2290	2089	88	7	0	2
8	23 Guilfoy Street	315.0	0.1808	1433	48	7	0	2
9	17 Carlyle Drive	749.7	0.5015	2991	7	9	1	2.5
10	8 Craft Avenue	217.7	0.2229	1008	52	5	0	1
11	22 Beechwood Ct.	635.7	0.1300	3202	15	8	2	2.5
12	14 Fox Street	350.7	0.1763	2230	54	8	0	2
13	7 Raynham Road	455.0	0.4200	1848	48	7	1	2
14	2 Jerome Drive	356.2	0.2520	2100	46	6	0	2
15	7 Valentine Street	271.7	0.1148	1846	12	5	1	3
16	38 Jefferson Street	304.3	0.1693	1331	64	5	1	1
17	15 Inwood Road	288.4	0.1714	1344	52	8	0	1
18	29 Meadowfield Lane	396.7	0.3849	1822	44	6	1	2
19	13 Westland Drive	613.5	0.6545	2479	46	6	2	2.5
20	79 Valentine Street	314.1	0.1722	1605	52	6	0	3
21	13 Fairmont Place	363.5	0.1435	2080	78	11	0	2
22	1 Prestwick Terrace	364.3	0.2755	2410	71	6	1	1
23	11 Clement Street	305.1	0.1148	1753	97	8	0	2
24	7 Woodland Road	441.7	0.3636	1884	45	7	2	2
25	36 Elm Avenue	353.1	0.1474	2050	41	10	2	2
26	17 Duke Place	463.3	0.2281	2978	40	6	2	2.5
27	12 Prospect Avenue	320.0	0.4626	2132	82	7	0	1
28	1 Buckeye Road	332.8	0.1889	1551	54	6	0	2
29	30 Ann Street	276.6	0.1228	1129	44	5	0	1
30	26 Broadfield Place	397.0	0.1492	1674	34	7	1	2
31	16 Jackson Street	221.9	0.0852	1184	94	5	0	1

Regression

Input
Input Y Range: $B$1:$B$31
Input X Range: $D$1:$D$31
☑ Labels
□ Constant is Zero
☑ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply: Simple Regression
○ New Workbook

Residuals
☑ Residuals
□ Residual Plots
□ Standardized Residuals
☑ Line Fit Plots

Normal Probability
□ Normal Probability Plots

Click in **Input Y Range** and Choose **B1:B31**; Click in **Input X Range** and choose **D1:D31**

Select **Labels** and **Confidence Level** (default is 95%).

For the output options, select **New Worksheet Ply** and input “**Simple Regression**” to name the new worksheet that will be generated later on.

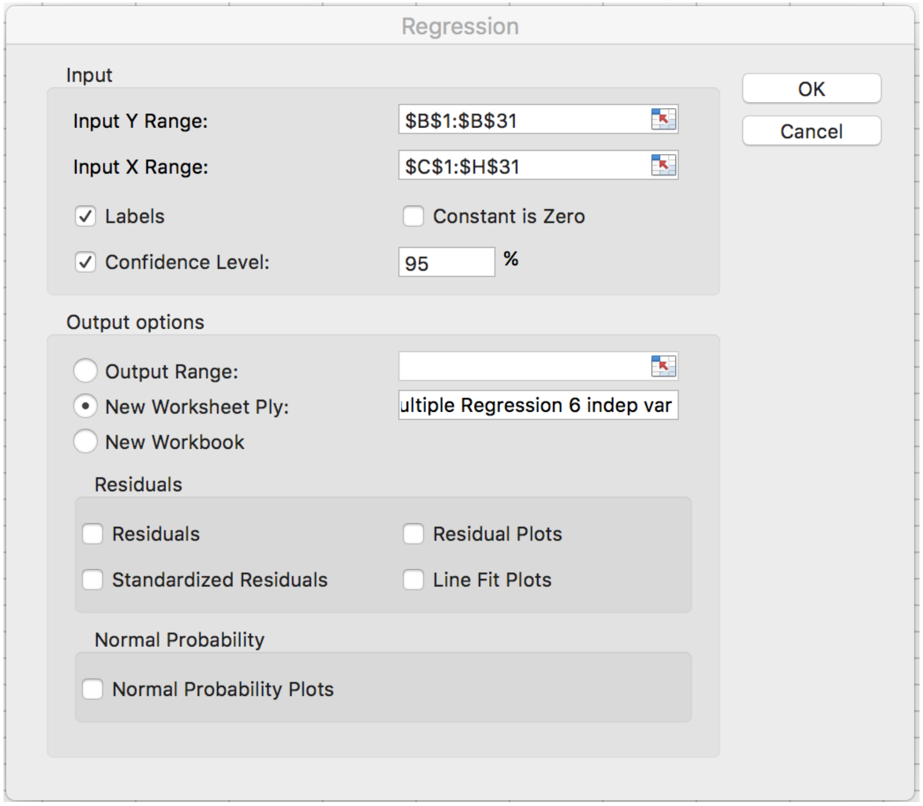
In the **Residuals** section, select both “**Residuals**” and “**Line Fit Plots**”.

Click **OK** and a new worksheet with results will be created.

Browse the new worksheet to see whether you can explain the meaning of the outputs on your own.

2. For Multiple Regression Analysis

Similar to the simple regression model, multiple regression will set the fields as follows:



Input

Input Y Range: $B$1:$B$31

Input X Range: $C$1:$H$31

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: ultiple Regression 6 indep var

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Click in **Input Y Range** and Choose **B1:B31**; Click in **Input X Range** and choose **C1:H31** (All 6 independent variables will be selected)

Select **Labels** and **Confidence Level** (default is 95%).

For the output options, select **New Worksheet Ply** and input “**Multiple Regression 6 indep var**” to name the new worksheet that will be generated later on.

Click **OK** and a new worksheet with results will be created.

Appendix 4: How to Drop a Column If the column is in the Middle of a Dataset

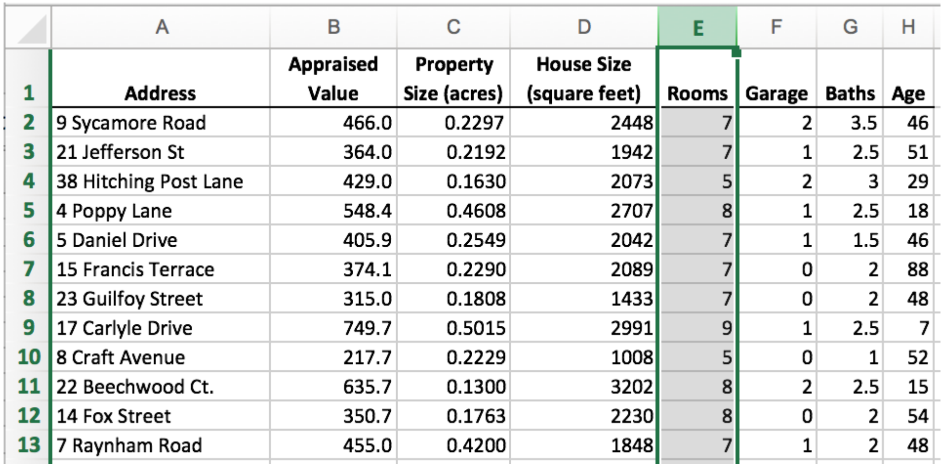
In Excel Regression model, input area has to be **contiguous reference**, which means if you want to drop a column from the regression, you have to either delete it or move it to the side, so that the **Input X Range** is contiguous.

In order to not lose any data, we choose to move column (variable) that we want to drop to the last column of the dataset, using **Cut** and **Insert Cut Cells.**

An example is given below:

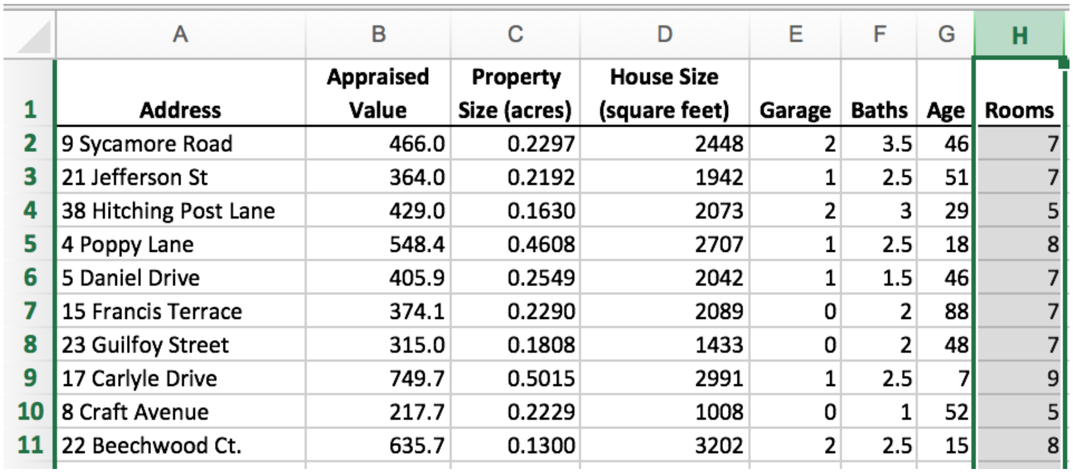
We want to drop **#Rooms** variable which locates in the middle of the independent variables.

Select **#Room** Column (Col E) and right click to choose **cut.**



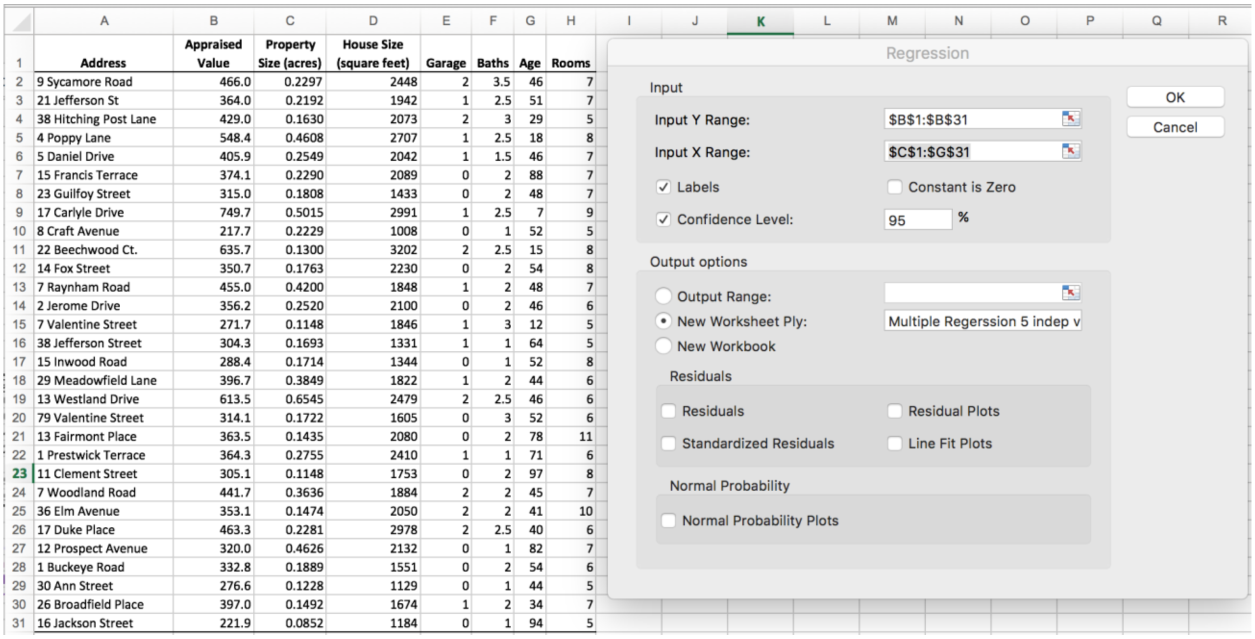
	A	B	C	D	E	F	G	H
1	Address	Appraised Value	Property Size (acres)	House Size (square feet)	Rooms	Garage	Baths	Age
2	9 Sycamore Road	466.0	0.2297	2448	7	2	3.5	46
3	21 Jefferson St	364.0	0.2192	1942	7	1	2.5	51
4	38 Hitching Post Lane	429.0	0.1630	2073	5	2	3	29
5	4 Poppy Lane	548.4	0.4608	2707	8	1	2.5	18
6	5 Daniel Drive	405.9	0.2549	2042	7	1	1.5	46
7	15 Francis Terrace	374.1	0.2290	2089	7	0	2	88
8	23 Guilfoy Street	315.0	0.1808	1433	7	0	2	48
9	17 Carlyle Drive	749.7	0.5015	2991	9	1	2.5	7
10	8 Craft Avenue	217.7	0.2229	1008	5	0	1	52
11	22 Beechwood Ct.	635.7	0.1300	3202	8	2	2.5	15
12	14 Fox Street	350.7	0.1763	2230	8	0	2	54
13	7 Raynham Road	455.0	0.4200	1848	7	1	2	48

Go to the column after last variable column, right click to choose **Insert Cut Cells**.



	A	B	C	D	E	F	G	H
1	Address	Appraised Value	Property Size (acres)	House Size (square feet)	Garage	Baths	Age	Rooms
2	9 Sycamore Road	466.0	0.2297	2448	2	3.5	46	7
3	21 Jefferson St	364.0	0.2192	1942	1	2.5	51	7
4	38 Hitching Post Lane	429.0	0.1630	2073	2	3	29	5
5	4 Poppy Lane	548.4	0.4608	2707	1	2.5	18	8
6	5 Daniel Drive	405.9	0.2549	2042	1	1.5	46	7
7	15 Francis Terrace	374.1	0.2290	2089	0	2	88	7
8	23 Guilfoy Street	315.0	0.1808	1433	0	2	48	7
9	17 Carlyle Drive	749.7	0.5015	2991	1	2.5	7	9
10	8 Craft Avenue	217.7	0.2229	1008	0	1	52	5
11	22 Beechwood Ct.	635.7	0.1300	3202	2	2.5	15	8

Open regression module and select the first five independent variables as **Input X Range:**



	A	B	C	D	E	F	G	H
1	Address	Appraised Value	Property Size (acres)	House Size (square feet)	Garage	Baths	Age	Rooms
2	9 Sycamore Road	466.0	0.2297	2448	2	3.5	46	7
3	21 Jefferson St	364.0	0.2192	1942	1	2.5	51	7
4	38 Hitching Post Lane	429.0	0.1630	2073	2	3	29	5
5	4 Poppy Lane	548.4	0.4608	2707	1	2.5	18	8
6	5 Daniel Drive	405.9	0.2549	2042	1	1.5	46	7
7	15 Francis Terrace	374.1	0.2290	2089	0	2	88	7
8	23 Guilfoy Street	315.0	0.1808	1433	0	2	48	7
9	17 Carlyle Drive	749.7	0.5015	2991	1	2.5	7	9
10	8 Craft Avenue	217.7	0.2229	1008	0	1	52	5
11	22 Beechwood Ct.	635.7	0.1300	3202	2	2.5	15	8
12	14 Fox Street	350.7	0.1763	2230	0	2	54	8
13	7 Raynham Road	455.0	0.4200	1848	1	2	48	7
14	2 Jerome Drive	356.2	0.2520	2100	0	2	46	6
15	7 Valentine Street	271.7	0.1148	1846	1	3	12	5
16	38 Jefferson Street	304.3	0.1693	1331	1	1	64	5
17	15 Inwood Road	288.4	0.1714	1344	0	1	52	8
18	29 Meadowfield Lane	396.7	0.3849	1822	1	2	44	6
19	13 Westland Drive	613.5	0.6545	2479	2	2.5	46	6
20	79 Valentine Street	314.1	0.1722	1605	0	3	52	6
21	13 Fairmont Place	363.5	0.1435	2080	0	2	78	11
22	1 Prestwick Terrace	364.3	0.2755	2410	1	1	71	6
23	11 Clement Street	305.1	0.1148	1753	0	2	97	8
24	7 Woodland Road	441.7	0.3636	1884	2	2	45	7
25	36 Elm Avenue	353.1	0.1474	2050	2	2	41	10
26	17 Duke Place	463.3	0.2281	2978	2	2.5	40	6
27	12 Prospect Avenue	320.0	0.4626	2132	0	1	82	7
28	1 Buckeye Road	332.8	0.1889	1551	0	2	54	6
29	30 Ann Street	276.6	0.1228	1129	0	1	44	5
30	26 Broadfield Place	397.0	0.1492	1674	1	2	34	7
31	16 Jackson Street	221.9	0.0852	1184	0	1	94	5

Regression

Input

Input Y Range: $B$1:$B$31

Input X Range: $C$1:$G$31

☑ Labels □ Constant is Zero

☑ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply: Multiple Regerssion 5 indep v

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

In this case, it is **C1:G31.**

Keep moving dropped columns to the right until we reach the final model.