COMP CRUNCHER

REGRESSION WITHIN COMPCRUNCHER:
Steps to Competency and Excellence

Mark R. Linné, MAI, SRA, CRE, CAE, ASA, FRICS
Managing Director
Education and Analytics

APPRAISALWORLD

Redefining the Landscape of Valuation

Regression within CompCruncher:
Helpful Hints to Make You an Expert

OK, you've taken the training, passed the test and are generally familiar with the CompCruncher application. Most of it makes a great deal of sense and flows the way you would ideally perform an appraisal. But then the regression section comes up. How should you evaluate the regression output? What is the "best" model? How can you make sure that you are getting the "best" value? How can you feel comfortable with the output?

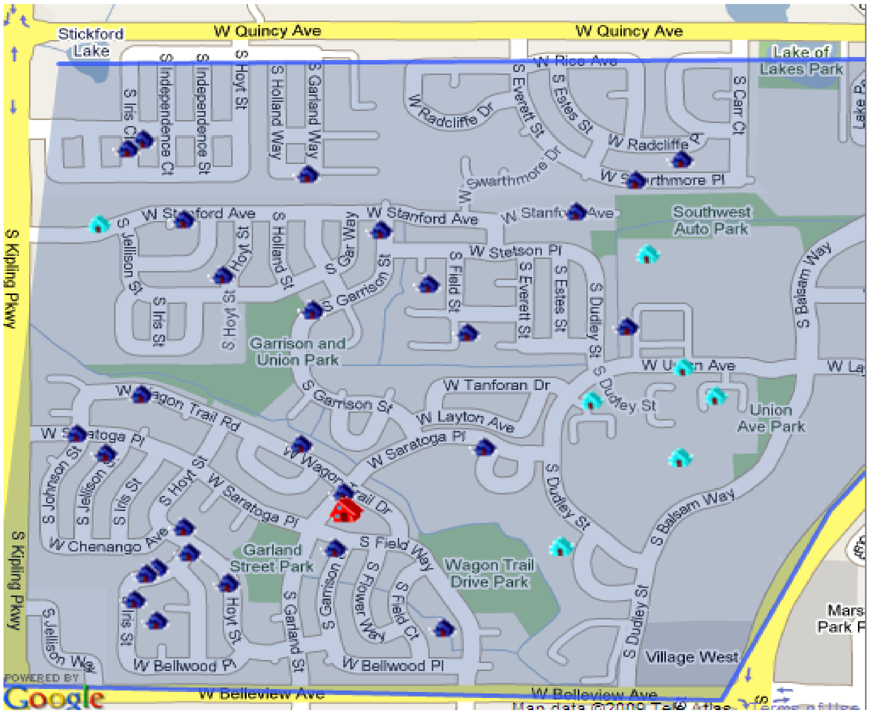
All of these questions and more, will be answered in the pages that follow. This guide to regression within CompCruncher is intended to help explain what constitutes a "good" outcome for the regression and will help guide you through this process.

In tandem with the education and testing that you have already successfully passed, as well as the experience you will gain by simply using the application, you will become a proficient and successful user of the revolutionary technology that will change our profession.

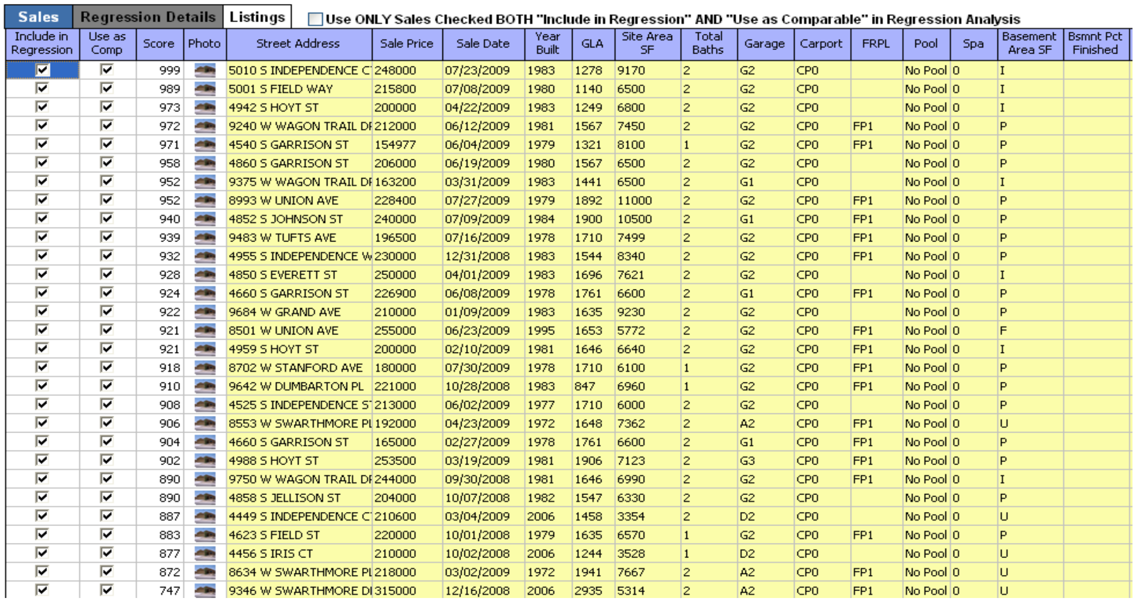
Lets get started!

1. The Little Picture is better than the Big Picture: Why regression works in CompCruncher:

Regression within CompCruncher (CC) works well because the appraiser gets to define the neighborhood. This is important in appraisal practice, and it helps to make the regression process relatively straight-forward. This concept is called small-market modeling, and it means that instead to creating large models that might take up an entire metro area, and then breaking down the different neighborhood locations and coding them within the model as different sub-markets, we have created a smaller model that has a consistent neighborhood definition. As with all things, however, there is a price to pay for benefits to small-market modeling. The first and foremost is that sometimes we do not get as many differences in the sales as we would normally like. Remember-regression likes differences. If there are no differences, then regression cannot determine which characteristics that are "different" lead to the difference in sales price. It would be the same if you were doing matched-pair analysis and all of your houses were the same. How could you tell the value of a garage if every house had a two-car garage? How could you tell the value of a bath if all of the houses had two baths. Regression has the same challenges when it looks at houses that have too much similarity. That is why more data is better in a set of sales. There is bound to be more chances for differences in the data-and that makes for a better regression model and better information on the sales components that contribute to value.



First Steps First: Understanding your data and what it means:



□ Use ONLY Sales Checked BOTH "Include in Regression" AND "Use as Comparable" in Regression Analysis

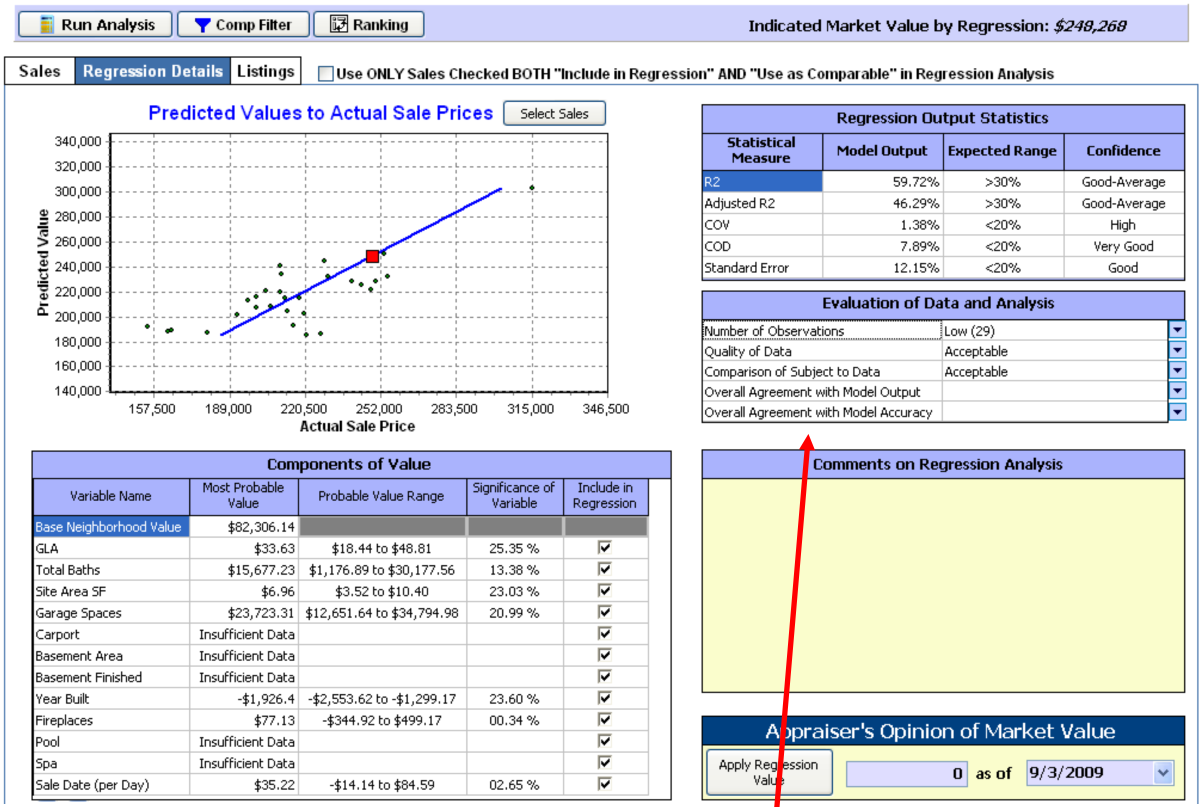
Include in Regression	Use as Comp	Score	Photo	Street Address	Sale Price	Sale Date	Year Built	GLA	Site Area SF	Total Baths	Garage	Carport	FRPL	Pool	Spa	Basement Area SF	Bsmnt Pct Finished
☑	☑	999		5010 S INDEPENDENCE C	248000	07/23/2009	1983	1278	9170	2	G2	CP0		No Pool	0	I	
☑	☑	989		5001 S FIELD WAY	215800	07/08/2009	1980	1140	6500	2	G2	CP0		No Pool	0	I	
☑	☑	973		4942 S HOYT ST	200000	04/22/2009	1983	1249	6800	2	G2	CP0		No Pool	0	I	
☑	☑	972		9240 W WAGON TRAIL DI	212000	06/12/2009	1981	1567	7450	2	G2	CP0	FP1	No Pool	0	P	
☑	☑	971		4540 S GARRISON ST	154977	06/04/2009	1979	1321	8100	1	G2	CP0	FP1	No Pool	0	P	
☑	☑	958		4860 S GARRISON ST	206000	06/19/2009	1980	1567	6500	2	G2	CP0		No Pool	0	P	
☑	☑	952		9375 W WAGON TRAIL DI	163200	03/31/2009	1983	1441	6500	2	G1	CP0		No Pool	0	I	
☑	☑	952		8993 W UNION AVE	228400	07/27/2009	1979	1892	11000	2	G2	CP0	FP1	No Pool	0	P	
☑	☑	940		4852 S JOHNSON ST	240000	07/09/2009	1984	1900	10500	2	G1	CP0	FP1	No Pool	0	P	
☑	☑	939		9483 W TUFTS AVE	196500	07/16/2009	1978	1710	7499	2	G2	CP0	FP1	No Pool	0	P	
☑	☑	932		4955 S INDEPENDENCE W	230000	12/31/2008	1983	1544	8340	2	G2	CP0	FP1	No Pool	0	P	
☑	☑	928		4850 S EVERETT ST	250000	04/01/2009	1983	1696	7621	2	G2	CP0		No Pool	0	I	
☑	☑	924		4660 S GARRISON ST	226900	06/08/2009	1978	1761	6600	2	G1	CP0	FP1	No Pool	0	P	
☑	☑	922		9684 W GRAND AVE	210000	01/09/2009	1983	1635	9230	2	G2	CP0		No Pool	0	P	
☑	☑	921		8501 W UNION AVE	255000	06/23/2009	1995	1653	5772	2	G2	CP0	FP1	No Pool	0	F	
☑	☑	921		4959 S HOYT ST	200000	02/10/2009	1981	1646	6640	2	G2	CP0	FP1	No Pool	0	I	
☑	☑	918		8702 W STANFORD AVE	180000	07/30/2009	1978	1710	6100	1	G2	CP0	FP1	No Pool	0	P	
☑	☑	910		9642 W DUMBARTON PL	221000	10/28/2008	1983	847	6960	1	G2	CP0	FP1	No Pool	0	P	
☑	☑	908		4525 S INDEPENDENCE S	213000	06/02/2009	1977	1710	6000	2	G2	CP0		No Pool	0	P	
☑	☑	906		8553 W SWARTHMORE PL	192000	04/23/2009	1972	1648	7362	2	A2	CP0	FP1	No Pool	0	U	
☑	☑	904		4660 S GARRISON ST	165000	02/27/2009	1978	1761	6600	2	G1	CP0	FP1	No Pool	0	P	
☑	☑	902		4988 S HOYT ST	253500	03/19/2009	1981	1906	7123	2	G3	CP0	FP1	No Pool	0	P	
☑	☑	890		9750 W WAGON TRAIL DI	244000	09/30/2008	1981	1646	6990	2	G2	CP0	FP1	No Pool	0	I	
☑	☑	890		4858 S JELLISON ST	204000	10/07/2008	1982	1547	6330	2	G2	CP0		No Pool	0	P	
☑	☑	887		4449 S INDEPENDENCE C	210600	03/04/2009	2006	1458	3354	2	D2	CP0		No Pool	0	U	
☑	☑	883		4623 S FIELD ST	220000	10/01/2008	1979	1635	6570	1	G2	CP0	FP1	No Pool	0	P	
☑	☑	877		4456 S IRIS CT	210000	10/02/2008	2006	1244	3528	1	D2	CP0		No Pool	0	U	
☑	☑	872		8634 W SWARTHMORE PL	218000	03/02/2009	1972	1941	7667	2	A2	CP0	FP1	No Pool	0	U	
☑	☑	747		9346 W SWARTHMORE DI	315000	12/16/2008	2006	2935	5314	2	A2	CP0	FP1	No Pool	0	U	

Doing a good job requires adequate preparation. The same rule holds true for applying regression in CC. When you first get to the regression screen you will see the sales that will be used in the regression. This will help you understand what type of data you will be running the regression on. Take a moment to look the data over and decide what challenges you might have. There always seem to be some challenges in how data is reported. Look at the site area, the garage area and the basement information. Remember that you are dealing with public record and MLS data. Sometimes data is filled in differently. Sometimes jurisdictions report data differently. In some jurisdictions, basement is reported as sq ft of basement area, and sq ft of basement that is finished. In others, like the example above, there is no square footage information other than unfinished and partially finished-which really isn't a lot of help. When you notice that in the data, you can anticipate that basement may not come in well in the regression, and you may have to take steps to deal with that.

Garage works similarly. Sometimes there is a code which indicates square footage of the garage. Sometimes there is an indication of a one-car versus a two-car. Watch this closely to see what happens later on with the regression.

Remember-looking at your data and being prepared for what comes later will save you time and give you a greater understanding of your data. Ultimately, its all about your data!

1. Starting Simply: Looking at a Homogeneous Neighborhood and Initially good Output:



Indicated Market Value by Regression: *$248,268*

□ Use ONLY Sales Checked BOTH "Include in Regression" AND "Use as Comparable" in Regression Analysis

Regression Output Statistics			
Statistical Measure	Model Output	Expected Range	Confidence
R2	59.72%	>30%	Good-Average
Adjusted R2	46.29%	>30%	Good-Average
COV	1.38%	<20%	High
COD	7.89%	<20%	Very Good
Standard Error	12.15%	<20%	Good

Evaluation of Data and Analysis	
Number of Observations	Low (29)
Quality of Data	Acceptable
Comparison of Subject to Data	Acceptable
Overall Agreement with Model Output	
Overall Agreement with Model Accuracy	

Components of Value				
Variable Name	Most Probable Value	Probable Value Range	Significance of Variable	Include in Regression
Base Neighborhood Value	$82,306.14			
GLA	$33.63	$18.44 to $48.81	25.35 %	☑
Total Baths	$15,677.23	$1,176.89 to $30,177.56	13.38 %	☑
Site Area SF	$6.96	$3.52 to $10.40	23.03 %	☑
Garage Spaces	$23,723.31	$12,651.64 to $34,794.98	20.99 %	☑
Carport	Insufficient Data			☑
Basement Area	Insufficient Data			☑
Basement Finished	Insufficient Data			☑
Year Built	-$1,926.4	-$2,553.62 to -$1,299.17	23.60 %	☑
Fireplaces	$77.13	-$344.92 to $499.17	00.34 %	☑
Pool	Insufficient Data			☑
Spa	Insufficient Data			☑
Sale Date (per Day)	$35.22	-$14.14 to $84.59	02.65 %	☑

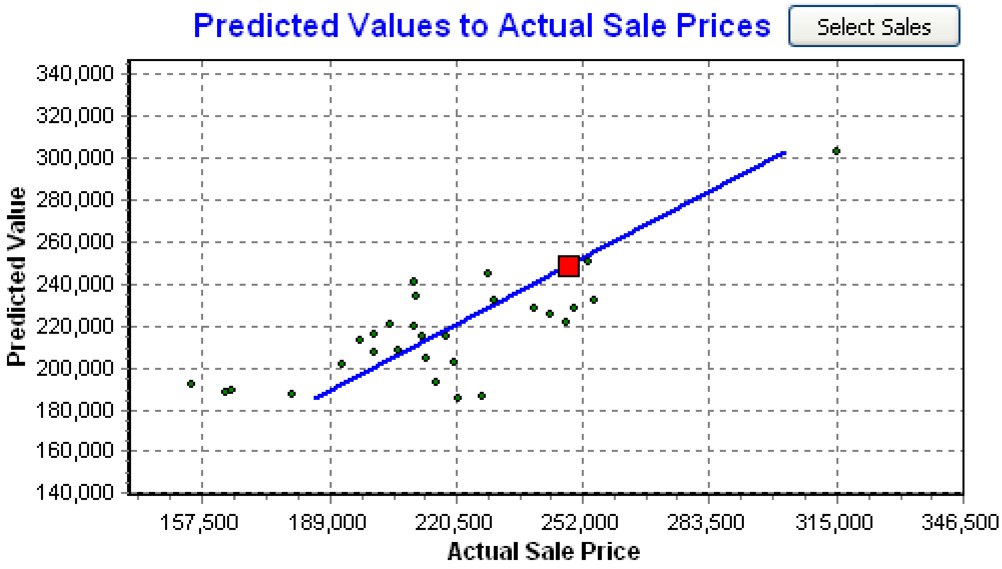
Comments on Regression Analysis

Appraiser's Opinion of Market Value

0 as of 9/3/2009

Initial Thoughts:

Looking at the data-we can see that there isn't a lot-so we need to be careful about taking too much out of the data set. At the same time, there seems to be some spread in the data. We can go to the "Evaluation of Data and Analysis" section of the screen and comment on the number of sales, the quality of the data and other key factors. Now lets look at each screen more specifically.

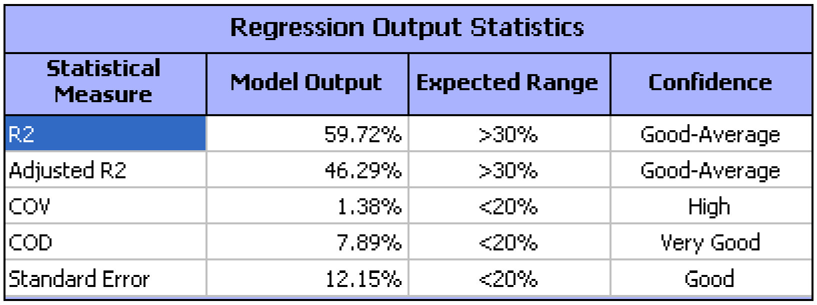


What is this graph telling us?

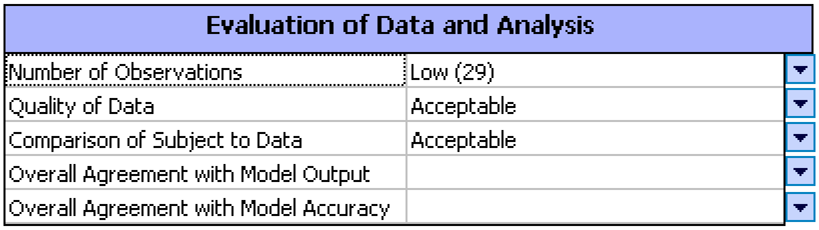
Most of the data appears to center on the subject-this is good. There appear to be a few outliers at the bottom of the regression line and one at the extreme upper-end of the range. These two groups of data can have a dramatic impact on moving the line in their direction. Remember that the outliers and their difference from the line are squared in the equation-this means that any extreme distances at the upper or lower-end of the line can have a significant impact.

One thing we can see right away, is that the subject seems to lie at the extreme upper-end of the range of data. We saw earlier when we looked at our data, that this neighborhood is a fairly homogeneous neighborhood-the subject should be right in the middle of the data. Somehow, something in the model is exaggerating the value of the subject-its likely too high. We need to look at the data more closely to see what the problem might be.

In the meantime, lets look at the actual model itself.



Regression Output Statistics			
Statistical Measure	Model Output	Expected Range	Confidence
R2	59.72%	>30%	Good-Average
Adjusted R2	46.29%	>30%	Good-Average
COV	1.38%	<20%	High
COD	7.89%	<20%	Very Good
Standard Error	12.15%	<20%	Good



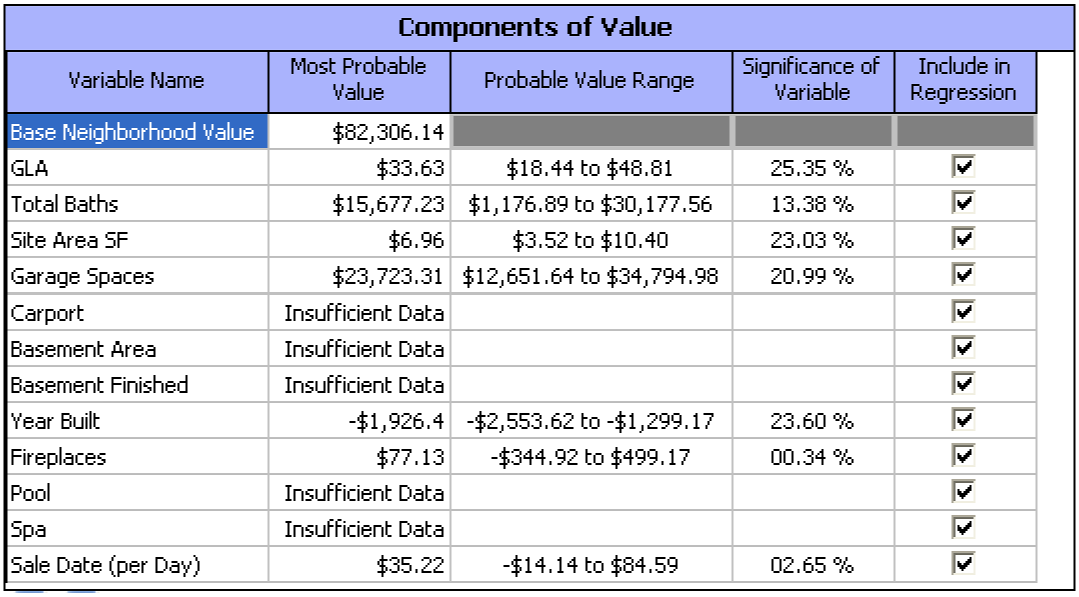
Evaluation of Data and Analysis	
Number of Observations	Low (29)
Quality of Data	Acceptable
Comparison of Subject to Data	Acceptable
Overall Agreement with Model Output	
Overall Agreement with Model Accuracy	

The **R**^{2} and the Adjusted **R**^{2} are both fairly good. Remember, in small neighborhoods, our expectation is that the **R**^{2} will be somewhat lower than if we have a larger multi-neighborhood model where there is a lot of variation. There is less variation in this neighborhood to begin with, so regression has a tougher time finding it.

The **COV** and **COD** are excellent, telling us that there is not very much variation in our data. This is excellent. It means that the mean and median data in our neighborhood is relevant. We likely have a normal distribution- a “bell curve”.

The **standard error** is also good. It tells us that the values in this model are +/- 12.15%. You can therefore know how “wrong” the values in this data set can ultimately be.

Let’s move on and look at the coefficients.



Components of Value

Variable Name	Most Probable Value	Probable Value Range	Significance of Variable	Include in Regression
Base Neighborhood Value	$82,306.14			
GLA	$33.63	$18.44 to $48.81	25.35 %	☑
Total Baths	$15,677.23	$1,176.89 to $30,177.56	13.38 %	☑
Site Area SF	$6.96	$3.52 to $10.40	23.03 %	☑
Garage Spaces	$23,723.31	$12,651.64 to $34,794.98	20.99 %	☑
Carport	Insufficient Data			☑
Basement Area	Insufficient Data			☑
Basement Finished	Insufficient Data			☑
Year Built	-$1,926.4	-$2,553.62 to -$1,299.17	23.60 %	☑
Fireplaces	$77.13	-$344.92 to $499.17	00.34 %	☑
Pool	Insufficient Data			☑
Spa	Insufficient Data			☑
Sale Date (per Day)	$35.22	-$14.14 to $84.59	02.65 %	☑

This is the key part of this screen-where you will likely do most of your work and analysis.

The first variable is the **Base Neighborhood Value**. This is the “Y-Intercept”-where the equation crosses the y axis. It is the unexplained component of the equation, and represents the inherent or core value within the subject’s overall value.

The value of **GLA** is $33.63/square foot, which seems reasonable for this neighborhood (remember that this is the marginal value per square foot of GLA, since we already are starting with a base neighborhood value of $82,306.14 for the property as a whole).

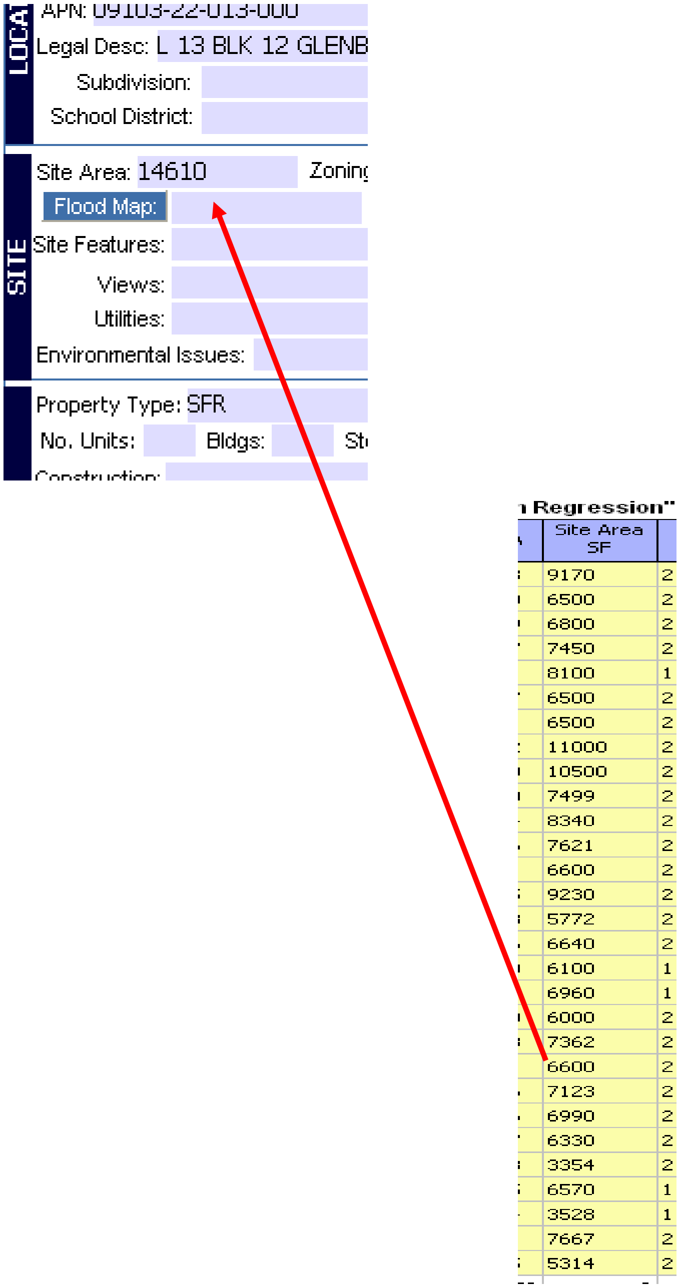
Total Baths at $15,677.23 per bath seems good for now.

The **Site Area SF** at $6.96 seems very high for a neighborhood of generally fairly standard lot sizes. We will have to look at this a bit closer in a minute. **Garage Spaces** are coming in $23,723.31-which also appears reasonable. **Basement Area** and **Basement Finished** information is coming in with insufficient data for analysis-which given the problems with that data field which we noticed earlier (public record only reports “P” for partially finished and “U” for unfinished and doesn’t give us any square footage information). Regression cannot operate without data of some kind.

The **Year Built** is a negative $1,926.40 per year. Its always good when this is a negative number-that makes intuitive sense. If any other variable is negative-in almost all cases, its best to just turn it off.

The final variable is **Sale Date (per day)** which is coming in at $35.22 per day-reasonable in this declining market.

OK: let's look at the data one more time. Is our subject in line with the data of the comparable sales?



APN: 09103-22-013-000

Legal Desc: L 13 BLK 12 GLENB

Subdivision:

School District:

Site Area: 14610 Zoning

Flood Map:

Site Features:

Views:

Utilities:

Environmental Issues:

Property Type: SFR

No. Units: Bldgs: St

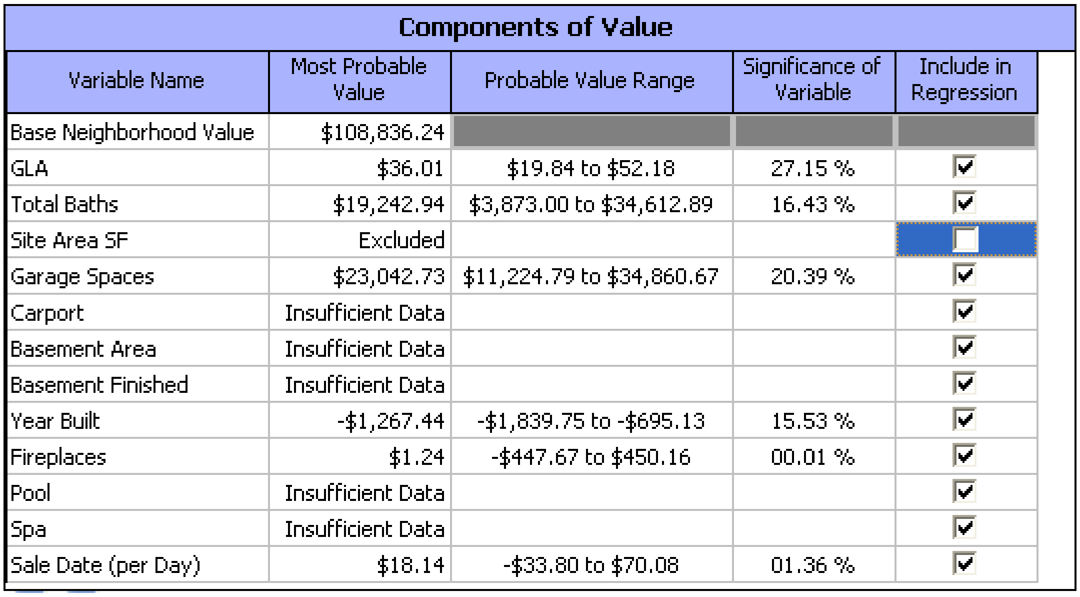
Construction:

Regression"

Site Area SF	
9170	2
6500	2
6800	2
7450	2
8100	1
6500	2
6500	2
11000	2
10500	2
7499	2
8340	2
7621	2
6600	2
9230	2
5772	2
6640	2
6100	1
6960	1
6000	2
7362	2
6600	2
7123	2
6990	2
6330	2
3354	2
6570	1
3528	1
7667	2
5314	2

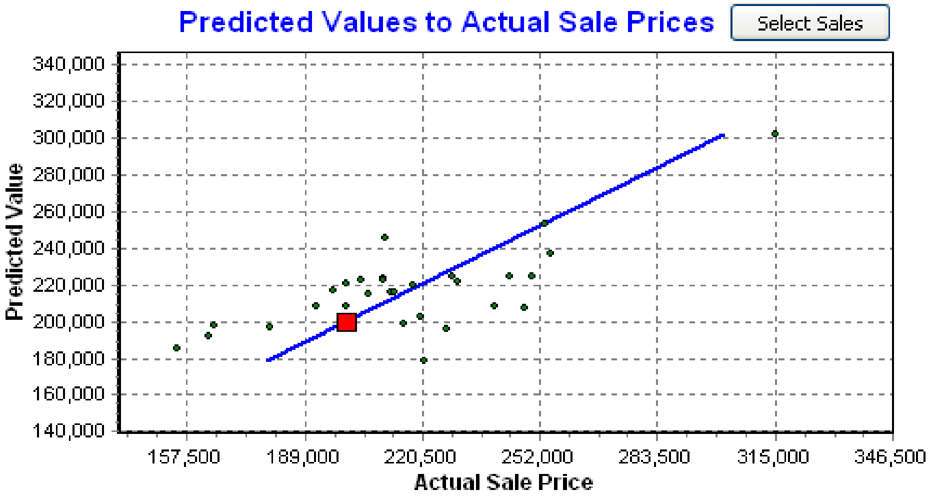
What do we notice about the data? The subject property has 14,610 sq ft, while the majority of the comparable sales have less than 8,000 sq ft; only a few are above that level. Yet our model is making an adjustment of $6.96 per square foot for the subject. The market likely does not value site differences in this neighborhood in this manner. It's probably better to de-select this variable, since it likely will over-value the subject as a result.

We remember in our original glance at the scatterplot, that the subject was being valued at the extreme upper-end of the range of comparable sales. Now we can see why.



Components of Value				
Variable Name	Most Probable Value	Probable Value Range	Significance of Variable	Include in Regression
Base Neighborhood Value	$108,836.24			
GLA	$36.01	$19.84 to $52.18	27.15 %	☑
Total Baths	$19,242.94	$3,873.00 to $34,612.89	16.43 %	☑
Site Area SF	Excluded			☐
Garage Spaces	$23,042.73	$11,224.79 to $34,860.67	20.39 %	☑
Carport	Insufficient Data			☑
Basement Area	Insufficient Data			☑
Basement Finished	Insufficient Data			☑
Year Built	-$1,267.44	-$1,839.75 to -$695.13	15.53 %	☑
Fireplaces	$1.24	-$447.67 to $450.16	00.01 %	☑
Pool	Insufficient Data			☑
Spa	Insufficient Data			☑
Sale Date (per Day)	$18.14	-$33.80 to $70.08	01.36 %	☑

Now that we have taken out the variable **Site Area SF**, what happens to our model? Most of the variables are essentially the same; the **Sale Date** has gone down a bit, as has the Bathroom variable-regression is re-setting the model slightly to accommodate your changes. That is OK. Overall our value has dropped to $199,000 (from roughly $248,00). That is OK. Now look at the scatter-plot below:



Now the subject property is within the range of sales that we have selected-it is no longer at the extreme top of the range.

Regression Conclusions:

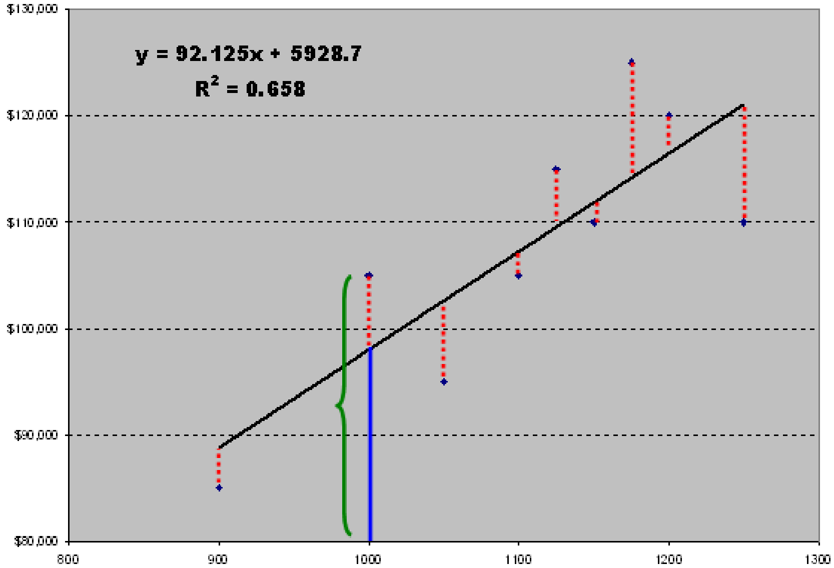
We ended up with a value of $199,000. That seems reasonable. We could have trimmed a few sales and our value would have gone to $210,000, but all of our statistical measures would have started to degrade and make less sense. Remember, you want to examine the data and potentially trim the data, but once you start to eliminate too many sales, regression will begin to produce some very odd answers.

A value of roughly $200,000 makes sense from the market perspective; all model statistics are good; we find from looking at our sales that a value of $210,000 to $215,000 is likely and supportable. Regression supports that value and the conclusion would likely be anywhere in the $200,000 to $210,000 based on both the direct sales comparison and regression outputs.

Life is good.

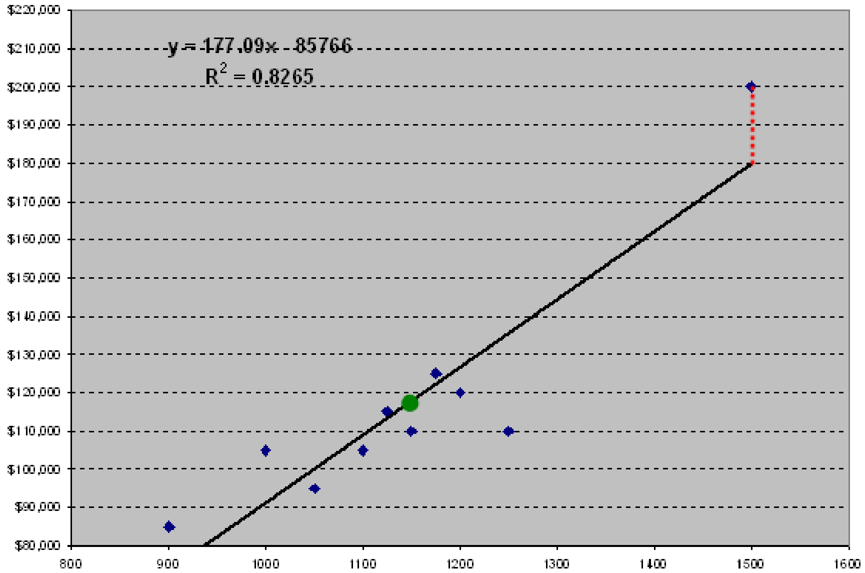
TOP TEN STEPS TO A FABULOUS REGRESSION OUTCOME

1. Look at your data first; anticipate challenges
2. Run the regression analysis with all of the data first-don't de-select until after the first regression run
3. Look at the model output first:
    a. **R^2** and Adjusted **R^2** : how good is my model?
    b. **COV/COD**: How much is my data spread around the mean or median?
    c. **Standard Error**: How wrong could I be?



Remember the Standard Error is based on the mean.

4. Look at the scatterplot-it's a visual representation of your data
    a. Do you see any patterns in the data?
    b. What about outliers?
        i. Could they be having an impact?



5. When you trim data-you have two choices:
    a. Polygon to exclude data
    b. Polygon to include and concentrate data

6. Lets say you cut too much-what do you do?
    a. The “undo” button (select all records)

Hints For a Better Value

1. Don’t trim too much
2. Regression begins to mis-behave as you go below 30 sales
3. Check the coefficient unit values carefully
    a. Do they make sense?
    b. If they are strong and wrong-they will have a significant impact on adjustments
    c. BACK TO SITE VALUE EXAMPLE
4. Be careful to “analyze” NOT “tinker”
5. Leaving “well-enough” alone
6. Don’t be afraid to “turn-off” coefficients if appropriate
7. Be Focussed on COV/COD/Standard Error
    a. You can live with a lower $\mathbf{R^2}$ if you have to.
    b. 40% +/- might be fine
8. Ultimately, try to achieve a balance in model output; data quality/quantity; and reasonableness of the coefficients (Mark’s Zen of Regression)
9. You have four chances to come up with a supportable estimate of value:
    a. Neighborhood Predominate Value; range, mean and median sales prices
    b. Regression estimate of value
    c. Sales Estimate of Value
    d. Listing Estimate of Value

10. At the end of the day-this is a flexible application; when it’s a tough property to value, regression may have challenges as well!
11. Ultimately a **good appraiser** can come out of the process with a **good outcome!**

Things to Remember

The more your property differs from the predominant property value in a neighborhood (the norm/the average); the tougher job regression will have

BUT: you can always come up with a value with one of the alternative techniques and use regression as a supporting technique to support sales and listing data.

If regression doesn't work-**Don't Panic!!!** We have sales, listings, neighborhood data, and your expertise. We can come up with a credible value, no matter what!