SAUDER
School of Business

Real Estate Division

LESSON 6

Basics of Model Building

Note: Selected readings can be found under "Online Readings" on your Course Resources website

Assigned Reading

1. UBC Real Estate Division. 2014. *BUSI 344 Course Workbook*. Vancouver, BC: UBC Real Estate Division.
   Lesson 6: Basics of Model Building

Recommended Reading

1. UBC Real Estate Division. 2009. *Advanced Computer-Assisted Mass Appraisal*. Vancouver, BC: UBC Real Estate Division.
   Chapter 2: Multivariate Analysis
   Chapter 3: Mass Appraisal Model Building

Learning Objectives

After completing this lesson, the student should be able to:

1. develop a simple linear regression model for selling price of real estate against one other variable;

2. develop a simple additive multiple regression model for determining the value of the same set of properties;

3. examine the statistics of the models above to determine the usefulness of the models in estimating sale price;

4. interpret the various measures of regression results: the coefficient of determination (R^2), standard error of the estimate (SEE), coefficient of variation (COV), correlation coefficient (r), *t*-statistic, F-value, VIF, and Tolerance.

5. explain multicollinearity, its implications in regression analysis, and two methods for detecting and avoiding it; and

6. describe the initial steps in model application, including calculating and analyzing the model's predicted values and residuals.

Instructor's Comments

Lessons 4 and 5 focused on how statistical and computer applications can be applied in single property valuation uses. In this lesson, we switch gears and turn our attention to using regression analysis for the valuation of multiple properties. This is commonly known as "mass appraisal", and with the now standard use of computers, "computer-assisted mass appraisal" (CAMA). Increasingly, these applications are also called automated valuation models (AVMs).

The use of regression for mass appraisal model building has been common in real property assessment since computer use became widespread in business applications. Since then, these applications have evolved from use in

only odd or unusual circumstances, with results often incomprehensible to both the layperson and real estate professional, to common usage in everyday real estate business by the general public. Similarly, what was formerly the domain only of statisticians and technicians is now easily accessible by the typical real estate professional.

Roughly speaking, the process in valuation model building involves taking a database of property sales and applying regression procedures to create a model that is able to predict these sale prices. When the model is fine-tuned such that it can predict these sale prices accurately, then it is ready to be applied to properties outside the database: e.g., estimating the market value of all properties in a city for property assessment for taxation purposes or the lending values for properties for mortgage underwriting.

We will cover valuation model building in this and the following two lessons (Lessons 6 through 8). In this lesson, we will spend more time on understanding what regression is and how it works in a modelling context. In Lesson 7, we will explore further complexities in the model building process. In Lesson 8, we will carry out a more comprehensive model building exercise, incorporating full data screening and model testing, in other words, starting the process from scratch and ending with a model that could be applied in practice.

The lessons will introduce valuation modelling in a hands-on and realistic manner, using real data from various locations across Canada. The software instructions for Lessons 7 and 8 will be in SPSS only. For Excel instructions, we have provided brief instructions in footnotes. In Lesson 6, Excel can carry out many or most of the necessary calculations, although it cannot easily calculate some of the statistics. However, as the modelling procedures become more complicated in Lessons 7 and 8, Excel's capabilities for modelling become limited and we strongly recommend the use of SPSS for these lessons.

NOTE FROM THE TUTOR

This and the next two lessons focus on mass appraisal, using SPSS and real market data for a variety of locations. Lesson 6 will offer a short and simple introduction, while Lessons 7 and 8 will each introduce additional complexity. Project 2 is based on a very similar analysis to what will be done in Lesson 8.

Introduction to Model Building: What is Regression All About?

As discussed in Lessons 1 and 2, regression is a powerful tool used by many industries and research groups. Regression can determine if there is a relationship between one thing (called the *dependent variable*) and one or more other things (*called independent variables*). We see applications of this technique regularly reported in the media; e.g., "scientists have determined a relationship between smoking and lung cancer". In statistical terms, this relationship is called a *correlation* and can be measured by the *correlation coefficient* (called R).

A *positive correlation* between two variables, e.g., smoking and lung cancer, simply means that as one thing increases, the other also increases (or, alternatively, if one decreases, the other also decreases). A *negative correlation* means the two items move in opposite directions – i.e., as one thing increases, the other decreases. An example of a negative correlation would be the size of a car's engine versus the gas mileage – the bigger the engine, the lower the kilometres you can expect per litre of gas. The closer the correlation coefficient to +1 or -1, the stronger the relationship. A strong correlation exists when the correlation coefficient is between 0.8 and 1.0 (or -0.8 and -1.0). For example, there is a perfect positive correlation (+1) between age and year of birth and a perfect negative correlation (-1) between age and life expectancy (or, perhaps, age and enjoyment of rap music).

When analyzing correlations, you must be careful not to assume causal effects – in other words, just because someone has lung cancer does not mean they were ever a smoker. This assumption of causal relationships is one method of misleading with statistics. When you read or hear about statistical information, you need to review it carefully to truly understand what is being said and what might be hidden from the audience. Regression analysis builds on the correlation relationship and produces a formula that estimates the value of the dependent variable by using known value(s) for the independent variable(s).

Now we will illustrate the application of regression in a real estate context, first demonstrating simple linear regression (two variables) and second, multiple regression (three or more variables).

Building a Simple Linear Regression

We will now apply simple linear regression to determine if there is a relationship between two variables and then build a predictive regression model for the dependent variable. For data, we will use a database of 120 condominium (condo) sales from the south market area of Regina. For variables, we will focus on the selling prices of the condos and total living area (in square feet). Selling price is the dependent variable and living area is the independent variable. In other words, we are assuming that a condo's selling price is dependent (to some extent) on its total living area.

The following section will illustrate how a simple linear regression model can be built using computer software. The "Regina1" database can be downloaded from the Course Resources website under "Online Readings", in both formats: Excel and SPSS. This lesson will illustrate techniques using the SPSS software, but brief Excel instructions are provided in footnotes.

The following is an excerpt from the "Regina1" database:

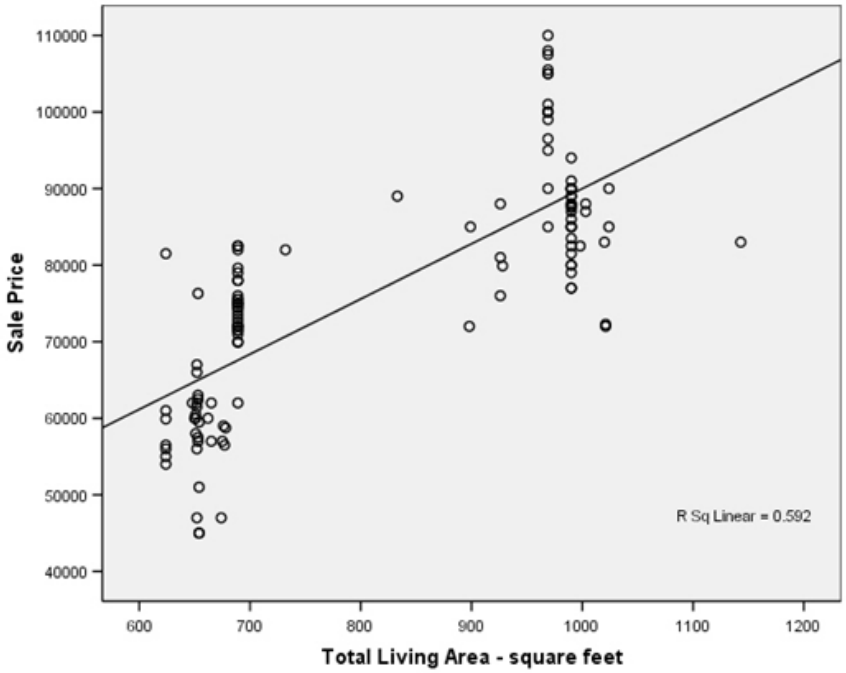
Condo #	Market	Unit #	Total Area	Sale Price
56	South	4	990	80000
56	South	9	990	77,000
56	South	10	990	85,000
56	South	11	990	89000
56	South	13	624	56500
56	South	20	990	94000
56	South	20	990	90000
56	South	22	624	56000
56	South	27	990	83500
56	South	28	990	88000
56	South	33	990	81500
56	South	37	990	90000
56	South	37	990	77000
56	South	41	990	88000
56	South	45	624	55000
56	South	48	990	89000
56	South	49	990	90000
56	South	52	990	85000
56	South	65	990	86000
56	South	79	624	61000
56	South	85	990	89900
56	South	85	990	87000
56	South	89	624	54000
56	South	92	990	79000
56	South	96	990	87700
56	South	99	990	82500
56	South	101	624	81500
56	South	113	624	59900
56	South	114	990	87500
56	South	124	990	85000
56	South	131	990	91000
56	South	137	990	80000
59	South	2	654	51000
59	South	3	654	45000
59	South	3	654	45000

Using this sample of 120 condo sales, we want to determine if there is a relationship between selling price and the total living area in the condo.

The first thing we might do is plot an x-y scatter diagram to visualize what the data looks like. Scatterplots provide an efficient method of examining relationships among quantitative variables. You could graph the data by hand using graph paper and a sharp pencil, but it is much easier to do this using the computer. This scatterplot can be produced in SPSS or Excel. Lesson 2 provided instructions for these programs, but the SPSS instructions will be briefly reviewed below.

- Select Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define[1]
- Select SalePrice for the Y-axis and Total_Area for the X-axis. Select "Use chart specifications from:" and browse to the "RSQ1" template saved in Lesson 2.
- Click OK to produce the chart.

Your chart should look similar to the following:



You can clearly see what appears to be a general upward trend in the data. This is expected, as it makes intuitive sense that larger condos sell for more. To further illustrate this trend graphically, we have also added a straight (or linear) trend line to this chart.

The linear trend line illustrated is sometimes called the "line of best fit" – by eyeballing the data, you could pick out where the best line through the data would fit. In statistical terms, this is called the "least squares regression line". The software calculates the exact position of the line such that it minimizes the sum of the squares of the differences (or distance) between the actual observations and the regression line itself.

[1] Excel: select the two columns Total Area and Sale Price (from the titles to the final entries), click on the Insert tab, click Scatter, and select the top left sample. The scatterplot will appear in the worksheet. You can move and resize the chart as you see fit.

To add a Trendline: with the chart active, click on the Layout tab, click Trendline, and select More Trendline Options…. Then select Linear, Display Equation on chart, and Display R-squared value on chart, Close. You can reposition the equation and R^2 text box within the chart area if you wish.

We can also describe the regression line mathematically by finding the regression line equation. To do so, follow these instructions in SPSS:

- Analyze → Regression → Linear…
- Put Sale Price under Dependent, Total Living Area under Independent(s), choose Enter as the Method, and under Statistics select Estimates, Model Fit and Descriptives, then click Continue.
- Click OK.

Your output will include the following:

Descriptive Statistics

	Mean	Std. Deviation	N
Sale Price	76593.50	14903.185	120
Total Living Area - square feet	814.05	159.126	120

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.770(a)	.592	0.589	9556.241

a Predictors: (Constant), Total Living Area - square feet

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15654521176.799	1	15654521176.799	171.422	.000(a)
	Residual	10775965153.202	118	91321738.586		
	Total	26430486330.000	119			

a Predictors: (Constant), Total Living Area - square feet
b Dependent Variable: Sale Price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17918.029	4565.621		3.925	.000
	Total Living Area - square feet	72.078	5.505	.770	13.093	0

a Dependent Variable: Sale Price

At this point, we are only interested in the Coefficients table. It includes the y-intercept for the regression line equation (17,918.03) as well as the slope (72.08), so we can write our regression line equation for living area and selling price as:

$$\text{Selling Price} = \$17{,}918.03 + (\$72.08 \times \text{Total Living Area})$$

This regression equation is the mathematical form of the line we "eyeballed" in the graph of the selling price and total living area.

This equation can be used to calculate the estimated selling price for a condo given its size. For example, we can estimate that a condo with 850 total square feet of living area will sell for approximately:

$$\text{Selling Price} = \$17{,}918.03 + (\$72.08 \times \mathbf{850}) = \$17{,}918 + \$61{,}268 = \$79{,}186$$
(rounded to $79,200)

With very little work we have created a rough estimator of the selling price for condos in this market area based on the total square footage of living area. Given any total square footage of living area we can estimate the selling price. This is a very simple regression model. However, an important question to ask at this stage is: "how strong is the relationship between selling price and total square footage of living area?" That is, how good will our estimates be, statistically speaking?

Evaluating Regression Results

We will look at six key statistics used in evaluating regression results. Four are measures of *goodness of fit* and relate to evaluation of the predictive accuracy of the equation. They are the coefficient of determination (R^2), the standard error of the estimate (SEE), the coefficient of variation (COV), and the F-Statistic. In different ways, each indicates how well the equation succeeds in predicting sales prices and minimizing errors. The other two statistics, the correlation coefficient (r) and the *t*-statistic, relate to the importance of individual variables in the model. The statistics we need are in the tables produced above.[2]

Coefficient of Determination

There are a number of additional measures that can be used to determine how well our regression line predicts the selling price. One of the most common is R^2, called the *coefficient of determination* (the correlation coefficient squared). R^2 measures how much of the variability in the dependent variable (sale price) is accounted for (or explained) by the regression line. That is, essentially, how good are the estimates of selling price based on this expression involving total square footage of living area.

Possible values of R^2 range from 0 to 1. When $R^2 = 0$, none of the variation in sales prices is explained by the model. On the other hand, when $R^2 = 1$, all deviations from the average sale price are explained by the regression equation and the sum of the squared errors equals 0. In a one-variable model, this implies that all sales prices lie on a straight line.

In our example, we found an R^2 of 0.59 – this is displayed in the chart above and in the SPSS output.[3] The R^2 statistic measures the percentage of variation in the dependent variable (sale price) explained by the independent variable (living area). If the R^2 is 0.59, this means that the regression line is able to explain about 60% of the variation of the sales prices ("variation" refers to the squared differences between sales prices and the average sale price). In practice, this can be loosely interpreted to mean total living area accounts for about 60% of the purchaser's decision to buy a specific condo. Or, conversely, total living area determines 60% of the selling price set by the vendor, while 40% is explained by other characteristics or by random variations in price. These two statements make intuitive sense at the very least – an important result, as common sense is a key factor in analyzing regression results!

The use of R^2 has two shortcomings. First, as we add more regression variables, R^2 can only increase or stay the same, which can overstate goodness of fit when insignificant variables are included or the number of variables is large relative to the number of sales. Assume that we have regressed sales prices on eighteen independent variables and obtained an R^2 of 0.920. Now suppose we re-run the model with a nineteenth variable, number of windows. As long as number of windows has any correlation whatsoever with sale price, R^2 will increase to above 0.920.

[2] Excel: to calculate the correlation coefficient (R), use the CORREL function; e.g., in our sample Excel data, in cell G4 type: =CORREL(D4:D123,E4:E123)^2<press Enter>.

[3] The R^2 and statistics discussed in the following section can be generated in Excel using the LINEST function.

Fortunately, R^2 can be adjusted to account for the number of independent variables, resulting in its sister statistic, adjusted R^2 or $\overline{R^2}$. In the present example, the addition of number of windows as a nineteenth variable will cause adjusted R^2 to fall unless the variable makes some minimum contribution to the predictive power of the equation.

The second shortcoming of R^2 (shared also by $\overline{R^2}$) is more a matter of care in interpretation. There can be no specified universal critical value of R^2; i.e., you cannot say "acceptable results have an R^2 of 85%" or any other value. The critical value of the R^2 statistic will vary with several factors and there are several non-mathematical reasons for variations in R^2 which make setting a specific target for this statistic inadvisable.

In mass appraisal, we often divide properties into sub-groups and develop separate model equations for each, e.g., for each neighbourhood separately. This reduces the variance among sales prices in sub-group and therefore we should not expect MRA to explain as large a percentage as when one equation is fit to the entire jurisdiction. For example, if one model is developed to estimate sale price for all neighbourhoods in a sales database, there may be $300,000 in variation among the sales prices. A model that explains 80% of the variation, still leaves 20% or $60,000 unexplained. A model for a single neighbourhood, with only $50,000 variation in sale price may have an adjusted R^2 of only 60%, but will produce better estimates of sales prices in that neighbourhood because 40% of the variation is only $20,000. The standard error and COV (discussed later) will show this improvement.

In general in regression models, improving the standard error and COV is more important than increasing the adjusted R^2, but you should generally try to have the adjusted R^2 as high as possible and the standard error and COV as low as possible.

Standard Error of the Estimate

The analyst must not only be able to estimate the equation for the regression line, he or she must also be able to measure how well the regression line fits the points. The techniques provided so far enable the analyst to determine a best fit regression line and measure its overall goodness of fit using R^2. However, it is also desirable to find out how well the regression equation fits each individual observation. It may be that the best fit line is very accurate at representing the data, or alternatively, if the data points are highly dispersed, the best fit line may be very poor.

The standard error of the estimate (SEE)[4] is one measure of how good the best fit is, in terms of how large the differences are between the regression line and the actual sample observations. The SEE measures the *amount* of deviation between actual and predicted sales prices.[5] If the SEE is small, the observations are tightly scattered around the regression line. If the SEE is large, the observations are widely scattered around the regression line. The smaller the standard error, the better the fit.

In our example, we found an SEE of $9,556.24. Note that whereas R^2 is a percentage figure, the SEE is a dollar figure if the dependent variable is price. Similar to the standard deviation discussion in Lesson 1, assuming the regression errors are normally distributed, approximately 68% of the errors will be $9,556 or less and approximately 95% will be $19,112 or less (see Figure 2.1 in Lesson 2).

In general, you want a small SEE relative to the size of the dependent variable – in our case the selling price. Say, for example, you were running several different potential models for estimating sale price with a variety of

[4] This statistic is also known as the square root of the Mean Square Error (MSE or RMSE).

[5] Lessons 1 and 2 explained that the standard deviation describes how spread out the data points are around the mean of those data points. SEE is a type of standard deviation: it measures how spread out the data points are around the regression line. Since the regression line is an estimate of the position of the data points, we call this type of standard deviation *the standard error of the estimate*. It is important to keep these two concepts separate in your mind. Remember, too, that the SEE statistic is distinct from the Standard Error of the Mean (often referred to as just the Standard Error), which we learned about in Lesson 2.

variables. You could then compare the R^2 and SEE for each to see which predicts the most variation in selling price with the least associated error.

The SEE is free from the second interpretive shortcoming of R^2 mentioned above. In other words, whereas R^2 evaluates the seriousness of the errors indirectly by comparing them with the variation of the sales prices, the SEE evaluates them directly in dollar terms. The problem with SEE is that it is an absolute measure, meaning its size alone does not tell you much in itself, and thus it can only be used in comparison to other similar models. However, you can create a further statistic from it that tells you how well you are doing in relative terms in your particular model. By dividing the SEE by the mean of the dependent variable, you get a relative measure called the coefficient of variation or COV.

Coefficient of Variation

In our example, the SEE is $9,556. This would indicate a good predictive model when mean property values are high, but not when they are low. Expressing the SEE as a percentage of the mean sale price removes this source of confusion.

In regression analysis, the coefficient of variation (COV) is the SEE expressed as a percentage of the mean sale price and multiplied by 100. The formula is the same as that described in Lesson 1, except that the SEE notation replaces the σ (standard deviation) notation.

Most regression software reports the SEE but not the COV, so we have to calculate it manually. Here, the COV is calculated by dividing the SEE (9,556.24) by the mean of the sale prices (76,593.50), yielding 12.48%. In general, for residential models which have sale price as the dependent variable, a COV of approximately 20% is acceptable, while a COV of approximately 10% indicates a very good result. At 12.5%, our model's COV is acceptably small, but not fantastic. This tells us that total square footage of living area does a fairly good job of predicting sale price, but there is more to sale price than just this one variable (as we would expect!).

Our COV implies that, given a normal distribution, roughly two-thirds of sales prices lie within 12.5% of their MRA-predicted values.

Correlation Coefficient

The correlation coefficient (r) is the first of two statistics that relate to individual regression variables. As explained in Lesson 1, the correlation coefficient is a measure that indicates the strength of the relationship between two variables. It can take on values from -1.0 to +1.0, ranging from very strong negative correlation to very strong positive correlation, or somewhere in between.

In our example, the correlation between sale price and living area is 0.7696 (rounded to 0.77). This is a moderate level of correlation getting close to being strong (0.80 is considered strong). So, it seems that our simple estimate is doing a pretty good job (based on this sample data of course). There is a strong positive linear relationship between square feet and sale price. Given the regression coefficient of $72.08, as the number of square feet increases by 1, the estimated sale price increases by $72.08.

MRA software usually includes an optional correlation matrix showing the correlation coefficient between each pair of variables. In analyzing correlations with the dependent variable, remember that the correlation coefficient is a dimensionless figure or percentage, indicating only whether two variables are linearly related. The correlation coefficient measures how strongly two variables have a straight line relation to each other, but does not give the exact relationship. Two sets of data (x,y) yielding exactly the same regression equation (straight line) may have very different correlation coefficients between x and y. Regression coefficients, on the other hand, indicate how variables are related; that is, how many units (dollars) the dependent variable changes when the independent variable changes by one unit (for example, one square foot) with other variables in the equation held constant.

t–Statistic

The *t*-statistic is a measure of the significance or importance of a regression variable in explaining differences in the dependent variable (sale price). It tests whether the slope of the regression line is equal to zero. Put simply, the *t*-statistics provides information as to the "goodness" of the regression. It helps answer the question: "does living area provide information in estimating selling price for condos in this market area?".

The *t*-values and their associated significance levels indicate the degree of confidence one can place on the regression coefficients. The significance of the *t*-values varies with the number of observations, so the significance level is more useful for determining the relevance of the variables. Higher *t*-values and lower significance levels increase the reliance the model builder can place on the statistical significance of the coefficients. A high *t*-value leads to the acceptance of the hypothesis that the coefficient is significantly different than zero.

What constitutes "high values" of *t*? Statistical tables provide the answer. These tables are based on the amount of confidence you would like in your answer and the number of *degrees of freedom* your data provides you (degrees of freedom was defined in Lesson 2). Generally, if you have plenty of data and want to have a statistical confidence of 95% in your answer, the critical value that the *t*-statistic must exceed is ± 1.96. A *t*-statistic in excess of $\pm$ 2.58 indicates that one can be 99% confident that the independent variable is significant in the prediction of sale price.

The *t*-statistic is dependent on the number of observations and therefore we cannot specify a universal value for acceptance or rejection. However, as a rough rule-of-thumb, modelers often use critical levels of *t*-statistic over 1.6 (90% confidence) or 2.0 (95% confidence). A significance level of .10 suggests that one can be at least 90% confident that the variable coefficient is significantly different from 0 – or, in other words, less than 10% probability that the coefficient is equal to zero. If the probability is high that the coefficient is equal to zero, this would indicate that the variable provides no useful information to the model. A significance level of less than .05 would indicate that the probability of the coefficient being equal to zero is 5% or less, which indicates a reliable result. Normally in mass appraisal work, a significance level of less than .10 is desired, and often .05 or less.[6]

Our *t*-statistic for living area is 13.093. If we were to refer to a *t*-table, we would find that when *t* is outside of $\pm$ 3.767, one can be 99.9% confident that the coefficient does not equal 0.[7] Therefore, in this case we can conclude with confidence that square feet of living area is significant in estimating residential values.

Contrast these results with another regression model estimating sale price using the independent variables, number of bedrooms and family rooms. If the variables had low *t*-statistics indicating significance values of 0.842 and 0.919 respectively, this indicates an 84% and 92% probability that the coefficients for these variables are actually equal to zero. Or, in other words, a high probability that neither of these variables are useful in the model.

F-Statistic

The F-Statistic (F-value or F-ratio) also provides information as to the "goodness" of the regression. It also helps answer the question: "does living area provide information in estimating selling price for condos in this market area?". The F-Statistic shows the overall quality of the regression, as opposed to the usefulness of the individual variables as reported by the *t*-statistic.

[6] Note that the choice of significance level depends on a number of factors and no one limit is applicable to all situations; for example in some mass appraisal work, a significance level of 20% to 25% may be appropriate.

[7] Statistical packages such as SPSS have *t*-tables built in and will usually provide the 95% significance level automatically. As such, in this course students do not have to manually reference *t*-tables.

The F-value is related to the correlation coefficient (r). It measures whether the overall regression relationship is significant; that is, it tests whether the model is useful in representing the sample data. The F-value is a ratio showing the portion of the total variation of the dependent variable that is explained by the regression divided by the remaining variation that is left unexplained by the model.

$$F = \frac{\text{variance explained by the regression}}{\text{variance unexplained}}$$

If explained variation is small relative to unexplained variation, the regression equation does not fit the data well and the regression results are not considered statistically significant. A small value of F (generally less than 4) leads to acceptance of the hypothesis that the regression relationship is not significant. If F is large, the hypothesis that the derived regression model is not significant is rejected and it is concluded that the overall regression results are statistically significant.

Similar to the *t*-statistics, critical values for F are found in statistical tables. The F-statistic is simply the *t*-statistic squared, so the critical values for F are 3.842 (95% confidence) and 6.636 (99% confidence) respectively. For a rough rule-of-thumb, modelers often use a critical level of F-statistic > 4 to indicate a statistically significant relationship.

Continuing with our example, the F-ratio of 171.422 is quite a bit larger than 4. This indicates that the estimates produced by the regression model provide a better representation of the sample data than the mean of the observations. In other words, the regression estimates fit the data well and the results are statistically significant. The size of the F-ratio above the critical value of 4 must be viewed with caution. At larger magnitudes, the F-ratio is useful mostly as a relative measure; for example, if two models are identical in all respects other than their F-ratios, the model with the larger F-ratio is probably the better one. The absolute measure of the F-ratio is less meaningful because F-ratios are sensitive to the number of observations and the number of variables in the model. Few observations, together with a relatively large number of variables, will generally produce a low F-ratio. The large F-ratio in our example is greater than the critical value of 4 and indicates that the estimates produced by the model are better predictors of value than the mean. However, the large number of observations and few variables in the model would be expected to produce a very high F-ratio.

Summary: Evaluating Regression Results

In evaluating regression models, it is important to evaluate both how well the regression model captures the observed variation in the dependent variable (price), as well as the error generated from the model.

When checking the regression output, the following points are important:

- the coefficients have the expected sign (positive or negative);
- the *t*-statistics are significant, i.e., greater than 1.64 (significance level less than .10);
- the F-statistic is "large" and the probability provided with the F-statistic should be less than .05;
- the standard error of the estimate or SEE (also termed the "root mean square error" or RMSE) should be small;
- the Coefficient of Variation (COV = SEE ÷ Mean Sale Price) should be small; and
- the adjusted R^2 should be large.

Note that these are just general guidelines and cannot be applied universally in all cases. Regression analysis is extremely complex and there are many interrelated factors that can affect results. Because of this complexity, the analyst must be very careful about not relying on universal measures or "cookbook" procedures.

Overall, our model appears to reasonably approximate sale price:

- the coefficient is +$72.08, which makes intuitive sense – as living area increases, price increases;
- the *t*-statistic is good at 13.093 (significance level is .000);
- the F-statistic is large at 171 and the associated significance is .000;
- the SEE of $9,556 is small relative to the $76,593 mean;
- the COV is 12.48%, which is acceptable but larger than optimal; and
- the adjusted R^2 at 0.59 is reasonably large, but not great, as 40% of variation in sale price remains unexplained.

Our simple model appears to be a reasonable start, but could be improved. We probably need to consider adding further explanatory variables to the model.

Our example has illustrated simple linear regression analysis – although the analysis of the statistics may not seem all that simple! Consider the box on the next page, which provides another example of evaluating regression results.

In our example, we used linear regression to estimate the value of one variable using one other. In essence, we created a simple mathematical model to describe sales price in terms of living area for condominiums in this market area based on a sample of 120 sales. A basic premise underlying modelling is that we use a sample to create our model, and once created we can then apply this model to the larger population the sample was taken from. For example, we could use the 120 condos sold in Regina's south market area to then estimate the selling price (or market value) for the other thousands of condos that did not sell. There are in-depth statistical methods for determining when you have an appropriate sample and what level of confidence you can place on generalizing the sample statistics to the entire population, but these go beyond the depth of coverage in this course.

So far in this lesson, we have used Excel and SPSS to demonstrate how multiple tools can achieve the same results. From here on, however, we will be using SPSS to illustrate the statistical procedures necessary. Where Excel can provide equivalent output, these steps will be explained as well.

With the simple form of regression under our belts, we will now broaden our investigation to multiple regression analysis.

Further Example of Evaluating Regression Results

Consider a regression model which has carport market value as the dependent variable and quality of carport as the independent variable. This model attempts to use carport quality to predict carport market value. Carport quality in this model is described using three binary variables: low, average, and above average quality. A binary variable has two values, 0 for no or 1 for yes. For example, a property with a low quality carport would have these three variables: LOW=1, AVERAGE=0, ABOVE_AVERAGE=0. In the regression formula shown, average quality is the base condition, shown to have a value of $2,419.51. This value is adjusted up or down for higher or lower quality.

According to the regression equation calculated, the estimated value of a low quality carport is:

Yi = 2,419.51 - 1,415.29 (1) + 1,915.83 (0)
= $1,004.22

The estimated value of an average quality carport is:

Yi = 2,419.51 - 1,415.29 (0) + 1,915.83 (0)
= $2,419.51

The estimated value of an above average quality carport is:

Yi = 2,419.51 - 1,415.29 (0) + 1,915.83 (1)
= $4,335.34

The coefficients appear reasonable; e.g., a high quality carport is shown to add $1,915 to value, while a low quality carport deducts $1,415. The t-statistics were large (12.8 and 13.6) and the significance levels for both were .0000, indicating that one can be virtually 100% confident that carport quality affects carport value.

The R^2 for this model was only .291, which indicates that only 29% of the variation in carport market value was explained by the regression model. This indicates that there are factors other than quality which influence carport market value, but have not been included in this analysis.

The SEE was calculated as $707.79. Given the mean carport value of $2,342, the COV is 30.2% (707.79 ÷ 2,342). This relatively large value indicates that the observations are widely dispersed along the estimated regression line and that carport quality, while definitely related to carport value, is not by itself a good predictor of carport market value.

The F-statistic of 2.84 indicates the model is not doing significantly better than simply estimating the carport value by simply using the mean carport value. Combining this with the low R^2 and high COV, it appears carport quality does not "explain" the variation in value well. It is likely that the carport value is more dependent on carport size than on quality. If information on carport size was available, it could potentially form part of the final model in addition to carport quality.

Model Building using Multiple Regression Analysis – Simple MRA

For our simple linear regression analysis example, we determined the following general expression:

$$y = a + bx$$

where

- y = the **Estimated Sale Price** (the dependent variable);
- x = the **Total Square Footage of Living Area** (the independent variable);
- a = the intercept or constant ($17,918.03 in our previous example); and,
- b = the slope of the line or (in more mathematically specific terms) the coefficient for the independent variable **Total Area** ($72.08 in our previous example).

Our analysis above showed that living area was a pretty good indicator of selling price, but that a lot of variation was left unexplained. As market participants, we all know that people shopping for real estate look at more than just size of the property – e.g., they might also consider number of bedrooms, bathrooms, view, neighbourhood attributes, building quality, and numerous other potential characteristics. The next step in our exploration of regression analysis is to create a model that can account for multiple variables.

In multiple regression analysis (MRA), there is still one dependent variable y and one intercept (or constant) a (often called b_0), but there are multiple independent variables x_1 to x_n and each has its own coefficient b_1 to b_n. So it follows that the general expression for an additive multiple regression analysis model is:

$$\mathbf{y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \ldots + b_nx_n}$$

OR

$$\mathbf{y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + \ldots + b_nx_n}$$

In simple linear regression, with only two variables, we "eyeballed" a line of best fit through the data and called that the regression line. The computer software then gave us the mathematical equation for this line. In multiple regression, we will add further variables to this equation. Say, for example, we added number of bedrooms as a third variable to our simple linear equation. For our graph, this now adds a third dimension to the drawing, with selling price still showing the height (left axis), total living area still showing width (right axis), but now with bedrooms showing depth. Our simple line graph is now a cube, and our condo sales are found floating somewhere within the cube depending on their selling price, living area, and number of bedrooms. Now, if you wanted to once again eyeball a line of best fit, you would have to draw this line through the cube, analogous to shooting an arrow through it with a line attached. This line would be the new multiple regression equation. However, figuring out the equation for this line becomes complicated mathematically, which is why we have computers to help us.

This graphical analogy helps to explain what is happening when a simple two variable regression becomes a three variable multiple regression, simply going from a 2-D diagram to a 3-D diagram. However, when further variables are added, you get into dimensions that are difficult to conceptualize and this simple graphical analogy no longer works. At this point, you will have to trust that the computer software you are using has the underlying mathematical equations programmed correctly – providing you with the line of best fit for an arrow shot into multiple dimensions. Now…clear as mud? We will provide an example to illustrate how straightforward and easy multiple regression can be!

In this section we will expand our sale price example (**y** as the dependent variable) to include x_1 as total square footage of living area, x_2 as floor number, and x_3 as number of bathrooms to create a simple model using multiple regression analysis. Using the notation above, and substituting market value (MV) for y, the equation would be specified as follows:

$$\mathbf{MV = b_0 + (b_1 \times LIVING\_AREA) + (b_2 \times FLOOR\_NUMBER) + (b_3 \times BATHROOMS)}$$

where

MV	=	estimated condo value (and hence market value or selling price);
$\mathbf{b_0}$	=	constant; and,
$\mathbf{b_1}$, $\mathbf{b_2}$, $\mathbf{b_3}$	=	the coefficients of living area, floor number, and number of bathrooms respectively.

This initial step is called model specification – setting out our expectations for what we think our multiple regression equation will look like at the end of our analysis. This is called an *additive model*, in that the terms are added together.[8] In a *multiplicative model* the characteristics are instead multiplied against an estimate of market value (analogous to percentage adjustments in the direct comparison approach). We will examine multiplicative models in a later course, BUSI 444.

Gathering Data

A key issue in all valuation work is the need to gather sufficient useful data. For example, when preparing an appraisal report, you must analyze a number of sales in order to choose good comparables and to support adjustments to their sale prices. Appraisals may require a minimum of three to six comparable properties, with the depth of analysis often depending on the availability of *bona fide* sales information. In some cases, limited data will necessitate a lesser depth of analysis.

In larger urban areas where many sales occur, a different problem may present itself, picking the best three (or six) can be time-consuming, and determining which are "best" may be more a matter of "best guesses" and time constraints than of solid opinion or of objective scientific methods. In evaluating comparables, the valuer may focus on the same small number of variables (e.g., lot size, age, square footage of house, condition, number of bathrooms, garage/carport type and size) regardless of subtly changing purchaser tastes.

Multiple regression analysis techniques can be used to more accurately choose the characteristics on which to compare properties. Multiple regression analysis allows a valuer to make full use of all data available rather than a simple stratification of the properties through selecting the "best" comparables. Because of these advantages, MRA can in many cases offer more accurate value estimates for properties than the traditional methods used for single property appraisal.

As seen in some of the earlier case studies, data gathering is critical in all facets of valuation. In fact, part of being an effective appraiser over the long-term comes from collecting and cataloguing data – this can be a competitive advantage for many appraisers. Sales and property data can be gathered (or perhaps purchased) from a variety of sources including:

- a land titles or registry office;
- a local assessment office;
- the local municipality, regional district, or county;
- a local realty company;

[8] The common specification for an additive model for residential property is: $MV = LV + BV$, where,
MV = market value;
LV = land value; and
BV = condominium value
Because our example deals with condos, land value will be ignored, so $MV = BV$.

- local listings; and
- a multiple listing service.

The data we use in this course comes from assessment offices and from the databanks of fee appraisers.

Data Analysis

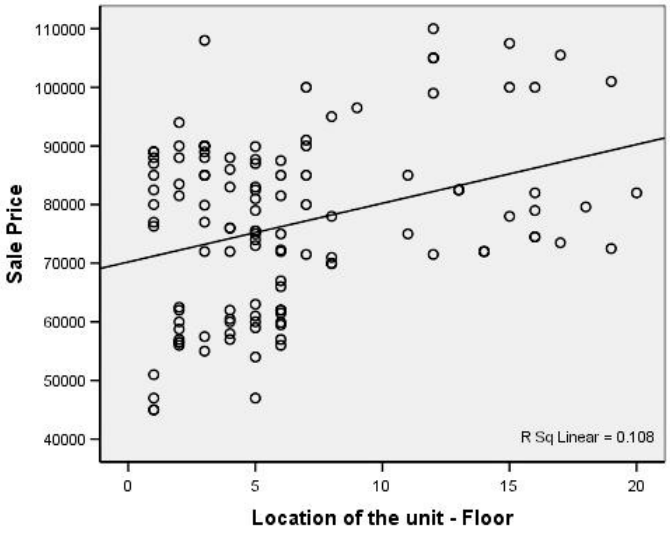
The database entitled "Regina2" contains property information on 120 stratified condominium sales from Regina, Saskatchewan.[9] These are the same sales we looked at in the previous simple linear regression example, but with two additional characteristics. There are now seven variables in the database. These are listed below:

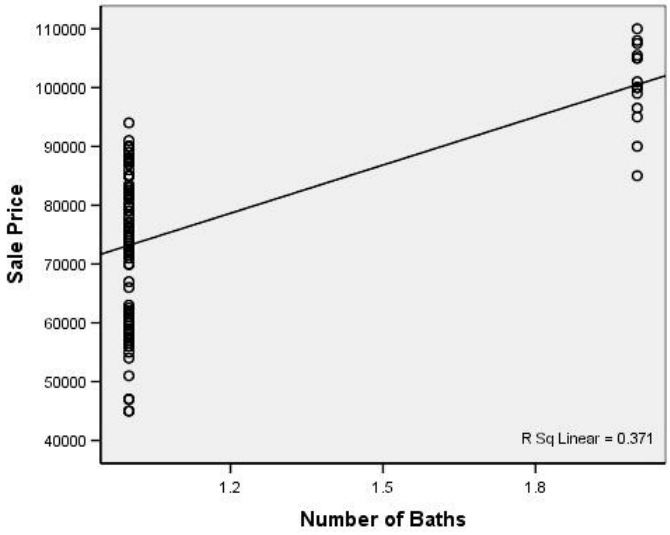
Regina2 Database

Condo #	Market	Unit #	Floor Number	Total Area	Number of Bathrooms	Sale Price
56	South	13	2	624	1	56,500
56	South	22	2	624	1	56,000
56	South	45	3	624	1	55,000
56	South	79	5	624	1	61,000
56	South	89	5	624	1	54,000
56	South	101	6	624	1	81,500
56	South	113	6	624	1	59,900
59	South	2	1	654	1	51,000
59	South	3	1	654	1	45,000
59	South	3	1	654	1	45,000
59	South	9	2	653	1	62,500
59	South	37	5	652	1	47,000
59	South	45	6	652	1	56,000
60	South	1	1	653	1	76,300
60	South	7	1	674	1	47,000
60	South	11	2	648	1	62,000
60	South	14	2	653	1	57,000
60	South	18	2	678	1	58,750
60	South	20	2	651	1	60,000
60	South	23	3	653	1	57,500
60	South	38	4	652	1	62,000
60	South	42	4	675	1	57,000
60	South	48	5	662	1	60,000
60	South	54	5	676	1	59,000
60	South	59	6	654	1	59,500
60	South	60	6	665	1	62,000
60	South	60	6	665	1	57,000
60	South	88	2	677	1	56,500
60	South	105	4	651	1	60,500
60	South	111	4	650	1	60,000
60	South	114	4	651	1	58,000
60	South	117	5	653	1	63,000
60	South	132	6	652	1	61,500
60	South	132	6	652	1	67,000
60	South	138	6	652	1	66000

[9] The "Online Readings" webpage provides this data in three formats: Excel, SPSS, and NCSS. Only SPSS will be illustrated here.

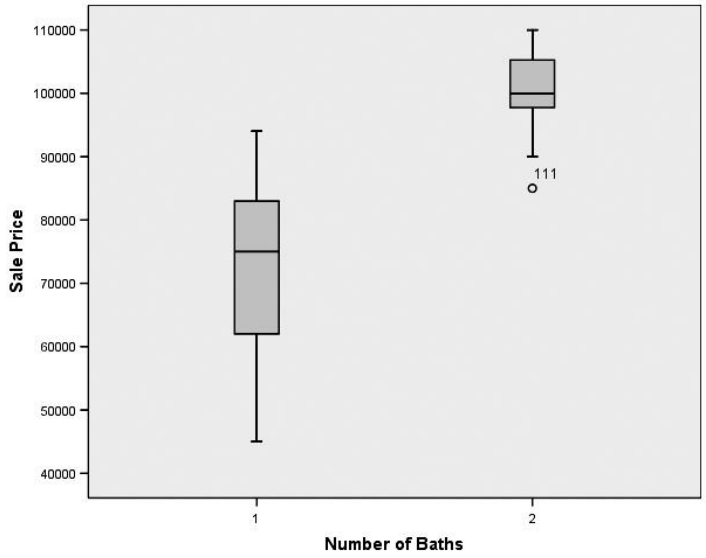
We will now examine the two new variables with scatterplots, to better understand their relationship to sale price.





These scatterplots show a slight upward trend for both variables in relation to sale price, although not as strong as what we saw with living area in the previous example. The R^2 statistics are only 0.108 for Floor Number and 0.371 for number of bathrooms, which are fairly low correlations. However, we believe they should improve our simple model, so we will try adding these variables to see what effect they have.

Because number of bathrooms is a discrete variable, it is helpful to view a boxplot to see if there is a significant difference between the selling prices of one bathroom condos versus two bathroom condos.



The boxplot shows a significant difference between selling prices of the one bathroom condos compared to the two bathroom condos, with their medians approximately $25,000 apart. This further emphasizes that number of bathrooms is likely a significant factor in determining sale price. You could use a Compare Means test here to accurately determine the difference:

- Select Analyze → Compare Means → Means…
- Select Sale_Price and place it in the Dependent List.
- Select Number of Baths and place it in the Independent List.

- Click Options
- Select Median from the Statistics list and add it to the Cell Statistics list
- Click Continue.
- Click OK.

You will see that there is a difference between the Means and Medians of $27,300 and $25,000 respectively.

Regression Analysis

We will now run the regression in SPSS and analyze the results.[10]

- Select Analyze → Regression → Linear…
- Select Sale_Price as the Dependent Variable.
- Select Total_Area, Floor_Number, and Bathrooms as the Independent Variables.
- Select Method as Enter.
- Under Statistics, select Estimates, Covariance Matrix, Model Fit, Descriptives, and Collinearity Diagnostics, click Continue.
- Click OK to run the regression.

The output should include the following:

Descriptive Statistics

	Mean	Std. Deviation	N
Sale Price	76593.50	14903.185	120
Location of the unit - Floor	6.37	4.894	120
Total Living Area - square feet	814.05	159.126	120
Number of Baths	1.13	.332	120

Correlations

		Sale Price	Location of the unit - Floor	Total Living Area - square feet	Number of Baths
Pearson Correlation	Sale Price	1.000	0.329	0.770	0.609
	Location of the unit - Floor	0.329	1.000	-0.090	0.411
	Total Living Area - square feet	0.770	-0.090	1.000	0.370
	Number of Baths	0.609	0.411	0.370	1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.886(a)	.784	.779	7010.071

a Predictors: (Constant), Number of Baths, Total Living Area - square feet, Location of the unit - Floor

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20730119265.208	3	6910039755.070	140.616	.000(a)
	Residual	5700367064.793	116	49141095.386		
	Total	26430486330.001	119			

a Predictors: (Constant), Number of Baths, Total Living Area - square feet, Location of the unit - Floor
b Dependent Variable: Sale Price

[10] Excel's LINEST function will calculate the coefficients and some of the statistics, but does not calculate all of them.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	5054.052	3580.636		1.411	.161		
	Location of the unit - Floor	923.313	150.319	.303	6.142	.000	.763	1.310
	Total Living Area - square feet	67.036	4.535	.716	14.781	.000	.793	1.261
	Number of Baths	9857.911	2374.074	.220	4.152	.000	.664	1.505

a Dependent Variable: Sale Price

From these results, we can create the regression equation for sale price using total living area, floor number, and number of bathrooms:

Selling Price = \$5,054.05
+ \$67.04 × Total Living Area
+ \$923.31 × Floor Number
+ \$9,857.91× Number of Bathrooms

Similar to the simple linear regression example, we could use this formula to manually calculate an estimated selling price for a given condo. For example, a condo with 950 total square feet of living area, on the 8th floor, and with one bathroom will sell for approximately \$86,000 in this market area. The calculation is below:

Selling Price = \$5,054.05
+ \$67.04 × 950
+ \$923.31 × 8
+ \$9,857.91× 1
= \$5,054.05 + \$63,688.00 + \$7,386.48 + \$9,857.91
= **\$85,986.44** (rounded to \$86,000)

Similarly, the estimated selling price for a condo with 1,100 total square feet of living area, on the 15th floor, and with two bathrooms is approximately \$112,400.

We have now created a simple multiple regression model! Our next step is to analyze how good a model it is. For this, we examine the statistics generated in the SPSS Regression report.

Analysis of Regression Results

The correlation (R) among these four variables is 0.886. This is a strong correlation. The R^2 or the correlation coefficient squared is 0.784, meaning more than 78% of the variability in the sales price is accounted for (or explained) by the regression line based on total square footage of living area, floor number, and number of bathrooms. Comparing to our simple linear regression model, floor number and number of bathrooms help explain approximately 18% more of the variation in sale price over living area alone.

So, can we conclude from this that, in this market area, the combination of total living area, floor number, and number of bathrooms determine 78% of a purchaser's decision to buy or that of a vendor when the price is set? Answer: maybe / maybe not. The tool (SPSS) has simply created a multiple regression model based on its computational ability and the data provided. This *exact* number cannot likely be relied upon, as it could actually vary significantly either way. The sample (120 sales) may not be representative of the population (thousands of unsold properties); there may be other problems with the model, some of which are discussed in the following paragraphs. Roughly speaking, our model indicates that these three characteristics appear to be important determinants of selling price. But the model must be analyzed critically and its application must be accompanied by appraisal judgment. While we are obviously keen on emphasizing regression as a powerful tool for valuation, we must caution you that it is not a "black box", where you put the data into the mystery machine and it magically comes out with the correct answer. A computer is a powerful tool for carrying out complex

mathematical calculations, but the human brain is much more powerful because it can apply reasoned judgment. A tool as powerful as regression applied uncritically is like a child playing with a loaded gun! Be careful and make sure you use that super-computer on your shoulders wisely.

Beyond correlation, we must also examine the "error" in the model. Keep in mind that error does not mean the model has made a mistake – in a statistical sense, error refers to the variation between the predicted value (market value) from the actual value (sale price), offering an indication of how accurate the prediction was. The SEE in our model is 7,010.071, but this alone does not tell us much. Dividing the SEE by the mean selling price results in a COV of 9.15% (7,010.017 ÷ 76,593.50). The target COV for a good model is less than 10%, and this model achieves the target. This means total square footage of living area, floor number, and number of bathrooms together provide a good predictor of sale price in this market area.

Further statistics to review in examining the goodness of the model are the F-statistic and *t*-statistics. The F-statistic is 140.616 which is quite large. The significance associated with it is zero, meaning the computer is telling us there is 0% probability that the results are by "chance". The target for the significance of the F-statistic is less than 5% (0.05), so this result means there is a good relationship here. If the F-statistic was small, say less than 4, it would indicate the overall model's significance was in question, meaning we cannot trust it to accurately estimate sale price.

In a model built through MRA methods, each variable in the model will have its own *t*-statistic. For each variable, the larger the *t*-statistic the more significant its contribution to the overall model. The critical value for a *t*-statistic is 2 or more – here, all three of the variables in the model have high *t*-statistics, indicating they are all significant. If any of the variables had a *t*-statistic less than 2, then we would have to consider if they were providing any benefit to the model. In statistical terms, a *t*-statistic of less than 2 implies we cannot be confident at a 95% level that the coefficient for the variable is different than zero, meaning we are not confident that its value is actually correct or if the variable could even be eliminated without affecting the model. This is not the case for our model.

Multicollinearity

With all of these statistics indicating positive results, it appears we can conclude this is a good model to estimate the selling price of condominiums in this market area. However, there is one more element of the model that needs to be checked before we can make this claim; we must examine for multicollinearity.

When we created the simple regression model between sale price and living area we were only concerned about the one relationship between sales price and the living area. In creating our more complex model, we must consider the relationship between sale price and each of living area, floor number, and bathrooms, but also the relationships among the independent variables – that is, how living area and bathrooms, living area and floor number, and bathrooms and floor number may be related. If any of these three combinations show any significant correlation, then we have multicollinearity in our model. The existence of high multicollinearity can invalidate an MRA model. This is because the overlap in the variables will cause the MRA process to become "confused" and the values of the coefficients will be inaccurate.

The first part of multicollinearity testing should be done during data screening, prior to running the regression (as will be seen in the following lessons). The second part of this testing should be done after the model is generated. The part that can be done beforehand is the examination of the correlation matrix. As can be seen in the Correlation table included in the regression results, our three variables have the following correlations:

- living area and bathrooms 0.370
- living area and floor number -0.090
- bathrooms and floor number 0.411

Variables with correlations over ±0.500 should be closely examined, although generally only those over ±0.800 will cause problems in an MRA model. At the outset of specifying a model, variables with correlations over ±0.800 should not be placed in the same model. In our case, the correlations are all low enough not to be of concern.

The second measure for multicollinearity is generated when the model is created, in the Tolerance and VIF (variance inflation factor) statistics. These two statistics measure the same thing, as they are inversely related; that is, Tolerance = 1 ÷ VIF. The Tolerance should be greater than 0.3 (and the VIF less than 3.333). A variable that has a tolerance value less than the target of 0.3 is considered to show a degree of multicollinearity which can have a serious effect on the value of its coefficient. A modeler must be wary to watch for low tolerance (high VIF) as the coefficients may be inaccurate.

In our case, the tolerances are:

- living area 0.793
- floor number 0.763
- bathrooms 0.664

All are greater than the critical value and indicate no multicollinearity. We can safely conclude that we have produced a good model to estimate the selling price of condominiums in this market area.

Applying the Model: Predicting Values

We manually calculated an estimated selling price for one condo, using the model coefficients, the characteristics of one condo, and the model regression equation. However, SPSS can calculate these values as well within the Linear Regression module:

- Click the Dialog Recall icon to return to the Linear Regression window (or select Analyze → Regression → Linear… again).
- Keep all entries the same, but click Save and check Unstandardized under Predicted Values and Unstandardized under Residuals, Continue. This will calculate the predicted values from the model and save them into a new variable (PRE_1) in the database. The residuals, or differences between the predicted and actual selling price, will also be calculated and saved into a new variable named RES_1. Click Continue.
- Click OK to run the regression.

You will find the output file includes the same output as before. However, if you view the data, you will now see that the predicted values and residuals were added to the end of the data file when the model was run. These are saved with the default names PRE_1 and RES_1. If you run further models in this database, they will save the values with further increments, e.g., PRE_2, PRE_3, and so on. Note that in other statistical literature, these predicted values are sometimes referred to as $\hat{Y}$ (or Y hat).

Helpful Hint!

Like any variable, these variables can be renamed as desired. You may click on the Variable View tab at the bottom left of the window, click on the PRE_1 box (under the Name column), and enter the desired variable name. Variables can also be deleted here. For example, if you run many models and save the results, when finished, you may wish to delete unneeded PRE and RES variables.

If you wanted to verify the predicted values were calculated correctly, you could manually calculate one or two of the PRE_1 values using that property's characteristics. For example, let's confirm the results for the first sale in the database, a 624 square foot condo, on the 2nd floor, and with 1 bathroom:

Selling Price	=	\$5,054.05
	+	\$67.04 × 624
	+	\$923.31 × 2
	+	\$9,857.91× 1
	=	\$5,054.05 + \$41,832.96 + \$1,846.62 + \$9,857.91
	=	\$58,591.54

This confirms PRE_1 is calculated correctly (note the slight difference is due to rounding in the manual calculation above).

RES_1 is the difference between the predicted value of \$58,591.54 and the actual selling price of \$56,500 – equal to -2,091.54 here (again, slight difference due to rounding). Knowing the residuals helps us in testing the model. Testing focuses on both how well the model predicts values, but also on the residuals (or errors) it creates. By analyzing the errors, you can often determine ways to improve the accuracy of the model.

Consider this model's residual statistics and the descriptive statistics for PRE_1, RES_1, and Sale Price:

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	58589.22	107270.96	76593.50	13198.586	120
Residual	-16625.869	19217.529	.000	6921.145	120
Std. Predicted Value	-1.364	2.324	.000	1.000	120
Std. Residual	-2.372	2.741	.000	.987	120

a Dependent Variable: SalePrice

Statistics

		PRE_1	RES_1	SalePrice
N	Valid	120	120	120
	Missing	0	0	0
Mean		76593.5000000	.0000000	76593.50
Median		75872.9569924	-336.0090117	77000.00
Std. Deviation		13198.58639783	6921.14472049	14903.185
Minimum		58589.21902	-16625.86838	45000
Maximum		107270.96130	19217.52936	110000

The mean predicted value is \$76,593.50, the same as Sale Price. However, the median predicted value is slightly lower, indicating some skew in the model results. The range of sale prices is wider than the predicted range, indicating some possible outliers in actual sale prices. The residuals are centred on zero, which is a sensible result given the regression model's function focuses on finding the line of best fit that limits each observation's residual from this line.

Let's return to the Regression window and produce the histogram for the residuals and the Casewise Diagnostics report for outliers:

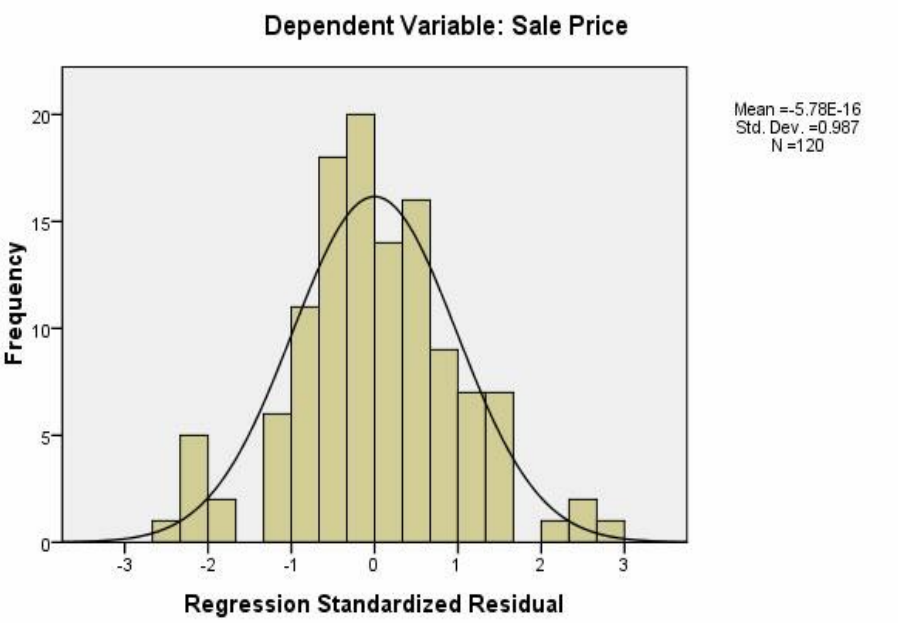
- Click the Dialog Recall icon to return to the Linear Regression window (or select Analyze → Regression → Linear again).
- Keep all entries the same, but click Statistics and select Casewise Diagnostics. Change the Outliers outside field to 2 standard deviations. Click Continue.
- Click Plots, and select Histogram. Click Continue.
- Click OK to run the regression.

The results now include the Casewise Diagnostics report and histogram of residual values:

Casewise Diagnostics(a)

Case Number	Std. Residual	SalePrice	Predicted Value	Residual
6	2.741	81500	62282.47	19217.529
9	-2.094	45000	59676.99	-14676.994
10	-2.094	45000	59676.99	-14676.994
12	-2.316	47000	63236.17	-16236.173
14	2.381	76300	59609.96	16690.042
93	-2.015	72000	86125.93	-14125.930
97	-2.372	72270	88895.87	-16625.868
99	2.471	89000	71676.49	17323.514
106	2.211	108000	92497.95	15502.045
111	-2.123	85000	99884.46	-14884.458

a Dependent Variable: SalePrice



The histogram shows the model's residuals. As expected, the model error is clumped around zero, with the number of larger errors decreasing as you move outwards from zero. The normal curve imposed on the histogram shows the distribution of residuals you would expect normally – we want to see a normal bell-shaped curve. The distribution here appears tighter than normal in the centre, a good result, but with a few possible outliers on either end. We will examine these further.

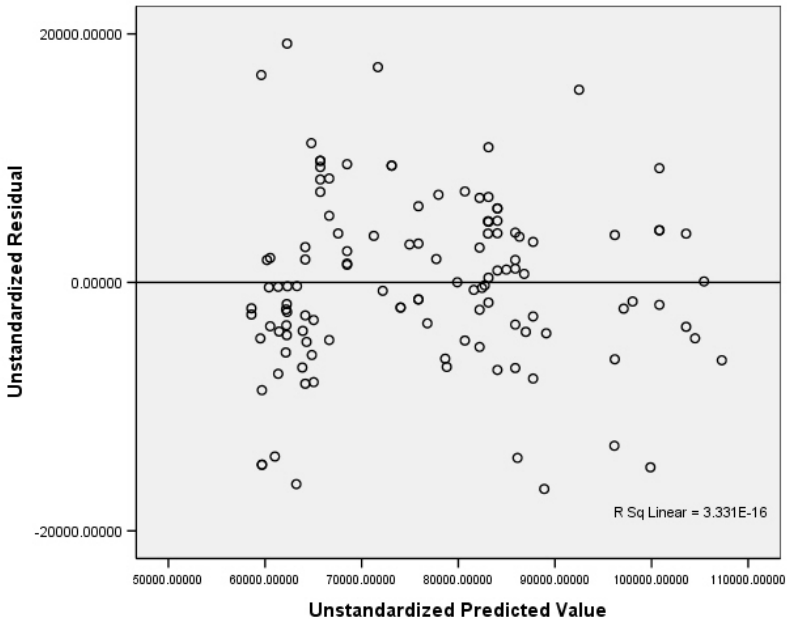
The Casewise Diagnostics report shows all predicted values with residuals more than two standard deviations away from zero. The default setting for this is three standard deviations, as at that point you are far enough from the mean that the result is questionable and should be examined further. Here, Case 6 is approaching that threshold, at 2.7 standard deviations from the mean. Examining this variable under the Data View tab shows it sold for $81,500, but the model is predicting its value as only $62,282 – a residual of over $19,000. This result is probably close enough to acceptable to leave it in the model. However, if we found more extreme results, we could consider setting a filter to remove this and any other outliers based on the RES_1 value (Data → Select Cases).[11] We will illustrate this in Lesson 8.

[11] In this model, if this one high outlier was removed, the model results improve slightly, with adjusted R-square increasing from .779 to .804 and SEE decreasing from 7,010 to 6,630.

We also want to know the relationship between residuals and predicted values: regression analysis assumes that these two variables are unrelated.

- Select Graphs → Legacy Dialog → Scatter/Dot → Simple Scatter → Define
- Select RES_1 for the Y-axis and PRE_1 for the X-axis → OK
- Double-click on the chart, click "Add Fit Line at Total" → Close the Properties window, Close the Chart Editor window

The plot should look as follows:



This plot shows there is no pattern to the residuals against predicted values, as all appear randomly centred on 0. The regression line and R-square show no significant correlation (the R-square is in scientific notation[12] – in percent it would be a correlation of 0.000000000000003%, certainly close to zero!). If the residuals and predicted values did appear to be correlated, this could imply a systematic over- or under-valuation by the model. For example, if properties become increasingly inaccurate as their prices increased, this would indicate heteroskedasticity. This could indicate a non-linear size relationship, as discussed in Lesson 3, which would necessitate some adjustments to the model.

Testing the Model

Once you are happy with your model, then you would begin testing it – compare value estimates from the model against the real sale prices, to see how well your model did. We started this process above, in analyzing the residuals.

Ideally, you want to carry out this testing with a group of sales that were not used in creating the model, meaning you would hold back a certain number of sales out of the model creation phase and then after the model was created separately apply the model to this group to estimate their sale prices. This allows you to test how

[12] For those unfamiliar with scientific notation it is a short form of writing very large or very small numbers. The number after the E represents powers of 10. So, 1.2E2 would be 1.2 times 10^2 or 1.2 times 100 or 120; 1.2E-2 would be 1.2 times 10^{-2} or 1.2 times 0.01 or 0.012. Further examples: 1.704E10 is equal to 17,040,000,000 (the decimal moves over 10 positions to the right); 1.704E-6 is equal to 0.000001704 (the decimal place moves over 6 positions to the left). In the example above where R Sq Linear equals 3.331E-16 that is the same as 0.0000000000000003331 or 0.00000000000003331%.

well the model did in estimating the sales prices of a group of properties not used to create the model. The model was created using a sample from a larger population – i.e., the properties that sold in our database are a sample, while the entire population is all properties in the neighbourhood, including those that did not sell. Because the model will eventually be applied to properties outside of the database, this testing method ensures that generalizing the results outside of the sample database will produce accurate results.

We will leave detailed testing to the examples in the next two lessons, where we will comprehensively evaluate the soundness of the model.

Summary

In this lesson you have:

- created a simple linear regression model;
- described a general additive model for a database of condominium property;
- examined the variables in the database;
- created a final model using multiple regression analysis;
- examined the statistics produced by the final model;
- produced the predicted values and residuals from the model; and
- analyzed the predicted value and residuals.

One of the key aspects of modeling is interpreting regression results. When checking the regression output, the following points are important:

- the coefficients have the expected sign (positive or negative);
- the *t*-statistics are significant, i.e., greater than 1.64 (significance level less than .10);
- the F-statistic is "large" and the probability provided with the F-statistic should be less than .05;
- the standard error of the estimate or SEE (also termed the root mean square error or RMSE) should be small;
- the Coefficient of Variation (COV = SEE ÷ Mean Sale Price) should be small; and
- the adjusted R^2 should be large.

Note that these are just general guidelines and cannot be applied universally in all cases. Regression analysis is extremely complex and there are many interrelated factors that can affect results. Because of this complexity, the analyst must be very careful about not relying on universal measures or "cookbook" procedures.

In the next two lessons, we will expand this simple coverage of regression modeling to include more variables and more complex analysis, to further develop the model building skills introduced in this lesson.

Review and Discussion Questions

1. After running a regression, you find that the model yields an SEE of 5,000. Is this a good result? What are the problems with using SEE as a measure of "goodness of fit"?

2. Explain the difference between simple linear regression and multiple regression.

3. Based on your regression analysis of condominium sales in Nanaimo, you determine:

   Market Value = \$42,000 + \$70(living area in sqft) + \$5,000(number of bathrooms)

   What do the coefficients in the equation represent? How many bathrooms would you expect a 1,000 square-foot condo that sells for \$119,500 to have?

4. Additive multiple regression includes a major assumption that the impact of the coefficient for a specific independent variable x_i is independent of the impact of other variables, e.g., x_2, x_3, x_4, etc. In other words, the impact of one independent variable on the dependent variable Y, is assumed to not be related to changes in another independent variable. When these assumptions turn out to be false, what problem do we have? How can this issue be overcome?

5. You conduct a regression analysis of detached single family housing prices in Langley, and then use the regression formula to calculate predicted values for your data set and the residuals (actual sales price - predicted value). What kind of results should you expect when you analyze the descriptive statistics for the residuals?

6. The first step in testing for multicollinearity is conducted during data-screening where the correlation of each of the independent variables is determined. What other steps can be taken to ensure that multicollinearity is not present in your model?

ASSIGNMENT 6

LESSON 6: Basics of Model Building

Marks: 1 mark per question.

1. A high SEE indicates:

   (1) a better result.
   (2) a worse result.
   (3) that multiple regression analysis is not a viable option.
   (4) multicollinearity exists.

2. A high VIF indicates:

   (1) multicollinearity is not present.
   (2) multicollinearity is present.
   (3) the Tolerance is also high.
   (4) the correlation coefficient is significant.

3. A COV under 10% indicates:

   (1) a good result.
   (2) a poor result.
   (3) that multiple regression analysis is not a viable option.
   (4) multicollinearity exists.

4. Consider the following statistics for two samples of apartment rents versus suite size for rental apartments in Waterloo:

   Dataset A, luxury high-rise concrete construction – R^2 of .732 and SEE of 6,000
   Dataset B, older 3-storey frame walk-up construction – R^2 of .635 and SEE of 8,673

   Rents are much higher in Dataset A than Dataset B. In which dataset would a regression equation more accurately predict the apartment rent?

   (1) Dataset A since the R^2 is higher and SEE lower than dataset B.
   (2) Dataset B since the R^2 is lower and SEE higher than dataset B.
   (3) Both datasets would have equal statistical reliability.
   (4) Impossible to determine because the R^2 and SEE are based on absolute values, so relative comparisons are not possible.

Assignment 6 continues on next page

THE FOLLOWING FOUR (4) QUESTIONS REFER TO THE "REGINA1" DATASET FROM LESSON 6:

5. Refer to the regression equation for the Regina1 database. What is the predicted value for sale price when total living area is 850 sq ft?

   (1) $77,384
   (2) Not possible to calculate since this data point does not lie on the regression line
   (3) $82,790
   (4) $79,186

6. What problem might you encounter if you use the regression equation for Regina1 data to predict the sale price of a 1,800 square foot rental apartment?

   (1) The regression does not account for land size.
   (2) The regression line is only based on data up to 1,143 square feet and the relationship may change above this range.
   (3) There may not be a causal relationship between the two variables.
   (4) There are too many outliers.

7. In the Regina1 example, the regression equation contained a constant of 17,918. What is another way of expressing this constant?

   (1) Minimum condominium price.
   (2) If the regression line is graphed, it will intercept the X-axis at square feet = 17,918.
   (3) The mean difference between the regression line and all observations.
   (4) If the regression line is graphed, it will intercept the Y-axis at $17,918.

8. Run a linear regression of Sale Price against Unit#. What can you conclude about the outcome?

   (1) The Adjusted R-Squared of 0.026 indicates 97.4% of the variation in sale price is explained by unit number.
   (2) A one unit increase in unit number is worth $75 in value.
   (3) The small F-statistic provides confidence that the model results are significant.
   (4) None of the above.

9. The standard error of the estimate is a good statistical tool for measuring:

   (1) a mathematical expression of the best fit of ordered pairs.
   (2) the percentage of variation in Y that can be explained by the regression line.
   (3) the amount of dispersion of the observed data around the regression line.
   (4) None of the above.

Assignment 6 continues on next page

10. Consider a model where the dependent variable is sale price and the independent variable is age of building. This resulted in the following regression equation: $Y = 100{,}500 - 960X$ and an R^2 of 0.8. What can you conclude about these results?

(1) Each year adds $960 to value.
(2) Weak negative correlation with 64% of the variation in sale price explained by building age.
(3) Strong negative correlation with 80% of the variation in sale price explained by building age.
(4) A 1 year old building is worth $101,460.

11. What is the advantage of multiple regression over simple linear regression?

(1) Helps deal with non-linear relationships.
(2) Provides the analyst an opportunity to account for additional sources of predictive error.
(3) Accounts for the economic reality that many variables may affect the dependent variable.
(4) Multicollinearity becomes increasingly possible.

THE FOLLOWING SIX (6) QUESTIONS REFER TO THE "REGINA2" DATASET FROM LESSON 6:

12. Assume a 920 square foot condo on the 15th floor was predicted to sell for $105,000. Based on the regression equation for Regina2, how many bathrooms must this condo have? Round your answer.

(1) 3
(2) 3.5
(3) 1
(4) 2.5

13. The Regina2 dataset has high correlation among all the variables. Does this mean that it will always have a high likelihood of predicting the sale price for any combination of the variables within the model parameters?

(1) Yes, the regression has accounted for virtually all the variation in the dependent variable.
(2) Not always, since there are other factors such as sampling technique, sample size, and COV which should be considered.
(3) Yes, since there is no longer any residual error.
(4) No, since only two of the variables are highly correlated.

14. Consider the regression statistics for the Regina2 dataset. What would your reaction be if the bathroom variable had a t-statistic of 0.105 and all other statistics for the remaining variables were unchanged?

(1) The bathrooms variable may offer no benefit to the model.
(2) We can no longer be confident that the bathroom coefficient value is correct.
(3) Our confidence in the significance of the bathrooms variable is improved.
(4) Both (1) and (2).

Assignment 6 continues on next page

15. Run a boxplot of sale price versus total living area. What can you conclude?

(1) It is difficult to see any clear relationship due to the number of box entries.
(2) A clear correlation between sale price and floor height is evident.
(3) There are sufficient observations for each occurrence of sale price versus total living area.
(4) All of the above.

16. What is median for the model's residuals (RES_1)?

(1) 75,873
(2) 0
(3) -336
(4) 6,921

17. Remove total living area from the model and review the new regression results. Which of the following statements is FALSE?

(1) The Adjusted R-Square decreases to 0.368.
(2) The SEE increases to 11,852.
(3) Floor number's t-statistic improves, increasing to 1.183.
(4) Number of baths' t-statistic increases, an improved result.

18. Consider a dataset with 4 variables: rent, gross rentable area (square feet), useable area (square feet), and floor level. A regression equation has been developed to predict rent using the other three independent variables. The R^2 value for the relationship of two independent variables, gross rentable area versus useable area is .832. Would you rely on this model?

(1) No, the model is suspect since it does not contain multicollinearity.
(2) No, since useable area is poorly correlated with rent.
(3) Yes, since the R^2 is quite high.
(4) No, variables which demonstrate multicollinearity should not be placed in the same model.

19. Consider the unique scenario in which the sale price of single family detached homes is predicted using four variables: total finished area (square feet), lot size (acres), number of fireplaces, and number of bathrooms. Multiple regression analysis can be used to determine the coefficients for each independent variable. Which of the following statements is TRUE?

(1) The independent variable with the largest coefficient will always have the greatest effect on sale price.
(2) The independent variable with the smallest coefficient will always have the least effect on sale price.
(3) The effect of an independent variable's coefficient depends on its size, but also on the nature of the variable and its unit of measurement.
(4) Total finished area will always have the greatest effect on sale price.

Assignment 6 continues on next page

20. Which of the following statements is TRUE?

(1) You should never remove outliers, as this compromises model results.
(2) Model testing is best carried out on the same sales used in creating the model, for consistency of results.
(3) A high correlation of a model's predicted values and residuals is a poor result.
(4) All of the above are true.

____
20 Marks

Planning Ahead

Project 2 requires you to build an MRA model to predict the selling prices of single family residential properties. This process is based on the analysis illustrated in Lessons 6, 7, and 8. You should read Project 2 now, so you have a sense of what will be expected. You can also begin investigating the data for Project 2. Try using the techniques outlined in this lesson to run some simple multiple regressions, and gauge your results. Although Lesson 8 offers a step-by-step approach for Project 2, it won't hurt to experiment with the data at this stage.

End of Assignment 6