SAUDER
School of Business

Real Estate Division

LESSON 7

Model Building Using Multiple Regression Analysis

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 7: Model Building Using Multiple Regression Analysis

Recommended Reading

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

Learning Objectives

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

1. describe the nine steps for building a multiple regression analysis model;
2. define a general linear regression model for predicting selling prices;
3. examine a database of variables and evaluate their suitability for use in the regression model;
4. apply a variety of transformation techniques to create variables that are appropriate for use in the regression model;
5. specify and calibrate an additive linear multiple regression model;
6. create and use a regression equation to predict the selling price of a property; and
7. test and evaluate the regression model using a variety of statistical methods.

Instructor's Comments

In Lesson 6, we built a simple model to estimate selling prices for condos in the south market area of Regina, using a multiple regression analysis for three independent variables. In this lesson, we will continue this exploration of multiple regression analysis (MRA), adding further complexity with more variables, data analysis, and testing.

In the first part of this lesson, we will expand the model from Lesson 6 to include more variables that contribute to selling price of condos. Towards creating an even more explanatory model for condos, we will define the general steps that should be taken in creating a model using MRA and follow those through.

In the second part of the lesson, we will further illustrate this process by starting with a database of raw data from the City of Regina's Assessment Department and performing the data analysis needed to create a model from scratch, including preliminary testing and analysis. In Lesson 8, we will once more complete a comprehensive model building process using new data, this time adding further complexities and realism.

Lessons 7 and 8 continue our hands-on approach with real data. The analysis will become a notch more complicated than Lesson 6 and therefore the instructions will be provided in SPSS only. We have produced an online supplement with instructions for NCSS, available in "Online Readings" on the Course Resources website. Because of the complexities of the modelling procedures in Lessons 7 and 8, we strongly recommend the use of either SPSS or NCSS for these lessons, because Excel's capabilities for this level of modelling become limited and problematic.

Steps for Building MRA Models

Building an MRA model requires a systematic step-by-step approach. The *Advanced Computer-Assisted Mass Appraisal* text listed as a recommended reading outlines this process in great detail in Chapter 3. There are two main phases in model building: (a) model specification and (b) model calibration. *Model specification* involves selecting the variables to be considered and defining their relationships to value and to each other. *Model calibration* means attaching numbers to the specified model, solving for the coefficients attached to the variables of interest. A part of model calibration is testing to ensure the model will create the estimated value with the desired accuracy. In this testing, the model's results are compared to real-world examples to see if the model can provide accurate estimates. If not, then the model must be re-specified and re-calibrated until it produces acceptable results.

In this lesson, the modeling process will be explained in nine steps:

- Steps 1 to 6 are model specification;
- Step 7 is model calibration; and
- Steps 8 and 9 are testing.

We will illustrate these steps through their practical application in developing a model. However, before this practical application, we will first briefly outline some background on these steps.

As outlined in Lesson 6, the first step is to describe an appropriate general model. The *Advanced Computer-Assisted Mass Appraisal* textbook describes several general model structures in Chapter 3. In selecting an appropriate model structure, you need to consider:

- the available data and its general format;
- the available software programs that can be used to calibrate the model – i.e., it does not make sense to specify a complex hybrid regression model when all that is available for calibration is a simple spreadsheet program; and
- good economic and appraisal theory.

For most of the uses in this course, we will apply the simple additive regression model as follows:

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

The dependent variable (y) is the variable you are trying to estimate – usually market value. The independent variables (x) represent the attributes or characteristics of the sold properties in the database (e.g., square footage, number of bedrooms). The coefficients (b) are the numbers assigned by the regression to each independent variable (e.g., dollars per square foot, dollars per bedroom).

Additive multiple regression is the most common form of regression as it conforms to the concepts of value and expectations of most market participants. It does not require a level of complexity beyond simple transformations and most statistical software packages (e.g., Excel, SPSS, NCSS) are capable of carrying out additive regression analysis. In terms of statistical procedures, additive regression is reasonably straightforward and the coefficients assigned to variables are understandable – in many cases, they can be explained in dollar terms, such as so many dollars per square foot of living area. Additive regression is the only method of model calibration that will be used in this lesson and the next. Multiplicative regression is another method for calibrating models which is more complex and less easy to explain. It is useful in some situations, but because of its limited application we will leave this topic to the more advanced model building discussed in BUSI 444.

The second step in the modeling process is to review the variables in the database in order to identify which may be suitable to use as independent variables for the model we are building. Up until now, we have used all of the variables provided in the databases, but in real applications you have to sort through many unsuitable variables and choose between alternative variables that may be all seem suitable but cannot be used together (e.g., multicollinear variables). In order to choose an appropriate model, you need to become familiar with the variables available in the database. For example:

- What property characteristics are represented?
- Do these characteristics fit better with an additive or multiplicative model structure?
- Are there several alternate forms for the same characteristic?
- If the form of the variable is not appropriate for the model structure, can a transformation be developed to put the variable into the correct form?

This is the data analysis phase, outlined in steps three and four below.

The third step, and one of the most important, is to examine the data in the database using the techniques described in Lessons 3, 4, and 6. Tools such as graphic analysis, crosstabs, compare means, and correlation analysis can help identify how variables are related to the dependent variable and also relationships between any independent variables. Multicollinearity was touched on in the previous lesson but will be described in more detail later in this lesson.

In the fourth step, we need to identify any variables that may, in our appraisal judgement, be useful in the model, but are not useable in their current form – these require transformation. For example, in the MRA model for selling price developed in Lesson 6, let's say we had a new variable that indicated the quality of each condo unit. This would likely be of use in the model, since purchasers tend to review quality in determining price. However, this new variable is coded as poor, fair, average, good, very good, and excellent, descriptions that are of no use in a mathematical regression model – e.g., what might 812.35 times "excellent" work out to? We need to translate these descriptions into numerical equivalents, so we transform poor to 0, fair to 1, average to 2, up to excellent as 5. Thus, 812.35 times 5 (for excellent) makes arithmetic sense and would add over $4,000 to the selling price of the condo.[1] This process is called *transformation* and is the fourth step in the process.

The fifth step is to repeat Step 3 with any transformed variables – re-evaluating the potential variables for the model to see which offer the best suitability. If necessary, variables can be transformed again.

The sixth step is listing the final group of potential independent variables.

[1] This type of substitution of 0, 1, 2, 3, 4, and 5 for poor, fair, average, good, very good, and excellent is provided as an illustration only. This specific transformation generally should not be used in practice because it creates the assumption that an excellent quality is five times "better" than a fair, which may not be correct. A better practice in this case would be to create linear coefficients to represent these values, with the factors based on market evidence. For example, if market research shows most properties are average, with good 15% more valuable, very good 20% more valuable, excellent 40% more valuable, fair 10% less valuable, and poor 25% less valuable, then this variable could be recoded as: poor=0.75, fair=0.9, average=1, good=1.15, very good=1.2, and excellent=1.4. This variable could then be used as a multiplicative adjustment. This process will be illustrated in Lesson 8.

The final three steps (7 through 9) are:

- create the model using the specified variables and statistical software, which will determine the coefficients for the regression equation;

- test and evaluate the model, finding how well it performs in terms of estimating the dependent variable or describing each variable's contribution towards explaining the value of the dependent variable; and finally,

- state your conclusions as to the quality of the model – in other words, describe (ideally in plain English that a layperson can understand!) what you did and why, and how well your model achieves its intended results.

For model calibration, in valuation work models are primarily calibrated to produce either predictive or explanatory results.

Predictive Model: developed to produce the highest quality overall prediction of market value – for example, achieve the best possible estimate of selling price, but not necessarily the most reliable estimates for the individual coefficients.

Explanatory Model: developed primarily to explain the value that each variable contributes to market value – for example, the value per square foot of living area. In other words, rather than focussing on the outcome of the model overall, this type of model focuses on developing the most accurate possible values for the coefficients.

In a good explanatory model, the variable coefficients could be used as adjustments in a direct comparison approach or as market-derived costs in a cost approach. In an explanatory model, the model builder wants to maximize the accuracy of the values of the coefficients. As discussed in Lesson 6, the significance of the coefficients is indicated by the *t*-statistic and its associated significance level. Higher *t*-statistics and lower significance levels increase the reliance the model builder can place on the statistical significance of the coefficients. A high *t*-statistic leads to the acceptance of the hypothesis that the coefficient is significantly different than zero, meaning you are confident the coefficient number is accurate. As mentioned in the previous lesson, the criteria for this is usually to have a *t*-statistic over 2 and a significance level of less than .05. This would indicate that the probability of this coefficient being equal to zero is 5% or less, meaning you are confident it is a reliable result.

On the other hand, a good predictive model can be used to directly estimate sales prices. In a predictive model, the model builder wants to ensure the R-square is high and the standard error of the estimate is minimized. This is normally achieved by including all variables that reduce the model's standard error, regardless of the *t*-statistics or significance levels for the variables. This leads to a model with the lowest overall error possible, but it does not necessarily produce reliable individual variable coefficients.

Depending on which goal is emphasized, the actions taken in the variable selection and calibration processes will vary. As this is an introductory course, the procedures described will tend to follow a middle course which allows for both a good prediction of market value and variable coefficients which are reasonable and rational from an appraisal perspective. During the modeling process illustrated in this lesson, we will briefly demonstrate how these different approaches vary in procedures and outcomes.

Illustration of MRA Model Building

Model Data

For this lesson, we will continue to focus on the Regina condominium sales data. The database we will use in this lesson is "Regina3". This contains the same 120 sales used in previous lessons, but with 15 additional variables. You can download this database from the course website under "Online Readings".[2]

There are 22 variables in the "Regina3" database. These are listed below:

Regina3 Database Variables

No.	Name	Label	Description
1	Condo#	Condo #	Condo Complex Identifier
2	Market	Market Area	Market Area – all are South
3	Unit#	Unit #	Unit Number
4	Topflr	On Top Floor	Located on the Top Floor
5	Floor#	Floor Number	Location of the unit – Floor
6	Directio	View Direction	Predominant view direction
7	Abutting	Abuting Influence	"Unit abuts specific influences (e.g., elevator, stairwell)"
8	Total_Area	Square Footage	Total Living Area – square feet
9	Bedrms	Bedrooms	Number of Bedrooms
10	Bath#	Number of Bathrooms	Number of Bathrooms
11	Fullbath	No. Full Bathrooms	Number of Full Bathrooms
12	Tqrbath	No. 3/4 Bathrooms	Number of Three Quarter Bathrooms
13	Halfbath	No. Half Bathrooms	Number of Half Bathrooms
14	Patiofl	Patios	Patio Flag – '1' unit has a patio
15	Balc#	Balconies	Number of Balconies
16	Parkug	Underground Parking	Number of Underground Parking Stalls
17	Parksurf	Surface Parking	Number of Surface Parking Stalls
18	Parkdgar	Det. Garage Parking	Number of Detached Garage Parking Stalls
19	SaleYear	Sale Year	Year of Sale
20	SaleMnth	Sale Month	Month of Sale
21	SalePrice	Sale Price	Sale Price
22	Pool	Pool	Pool Flag – Complex – '1' complex has a pool

To obtain a printout of the variables in SPSS, you can select the Variable View tab and click File → Print.

STEP 1: Specifying the Model

We will be developing an additive model to estimate the value of condominiums (condos) based on the variables given in the "Regina3" database. Again, the additive general model that is often applied to residential property is:

$$MV = LV + BV$$

[2] This data can be downloaded in all three formats: Excel, SPSS, and NCSS. Only the SPSS instructions will be shown in this lesson.

where

MV = estimated market value (or selling price);
LV = land value; and
BV = building value.

However, because we are dealing with condos, land value will be ignored. Given the list of variables available in our database, we will produce the following general model for the market value:

$$MV = b_0 + (b_1 \times LIVINGAREA) + \sum(b_i \times LOCATION\_VARIABLE_i) + \sum(b_j \times AMENITY_j)$$

where

MV = is the estimated condo value or the estimated market value (or selling price);
b_0 = constant;
b_1, b_i, b_j = coefficients of the independent variables;
LIVINGAREA = total living area in the condo;
$LOCATION\_VARIABLE_i$ = any variable associated with the location of the condo (e.g., floor number);
$AMENITY_j$ = any variable associated with condo amenities (e.g., underground parking or a pool in the complex).

STEP 2: Reviewing the Variables

The next step in the model development process is to review the variables available in the database and group the variables according to the factor the characteristic represents in the general model (e.g., in this case living area, location, or amenities). Very often, some characteristics to be included in the model will be represented by more than one variable and sometimes it is necessary to use a combination of more than one variable to correctly represent the characteristic or factor needed in the model.

For our database, the variables can be sorted as follows:

FACTOR	VARIABLES
Living Area	Total_Area
Location	Topflr, Floor#, Abutting
Amenities	Directio, Bedrms, Bath#, FullBath, Tqrbath, Halfbath, Patiofl, Balc#, Parkug, Parksurf, Parkdgar, Pool

Before making choices among variables or making changes to the database (Step 4: Transformations), you should familiarize yourself with the characteristics of the variables. In the next step, we will examine the variables to see which are best suited to be included in the modeling process (Step 6: List the Variables for Calibration).

STEP 3: Examining the Variables

In this section, we will examine the variables in our database, their relationship to sale price, and their relationship to each other. There are a number of reasons for testing these relationships:

- To get a sense of the important variables so we can get a feel for what to expect out of the final model.
- To exclude variables from the regression model that are of no use. For example, they may have little or no statistical relationship to the sale price (within the data being analyzed) or they have too few occurrences within the data to be accurately representative of the property characteristic (e.g., if only one property of 120 has a view, then the model would not be able to accurately determine what value a view might have).

- To avoid multicollinearity, excluding any variable strongly correlated with another variable.

- To find variables that might be useful, but need to be changed into a useable format. For example, a variable that has values of TRUE or FALSE would need to be changed into a numerical form: e.g., 1 for TRUE and 0 for FALSE.

- To find variables that might be useful, but are multiplicative in nature. For example, a percent condition variable might range from 80% to 120% indicating 20% better or worse condition than the 100% average. If you wanted to include the effect of condition in the additive regression model, you could multiply this percentage by the total square footage of the home. This would mean homes in better condition would appear to the model to have more square footage and in this way the effect of condition would be brought into the model.

Many of the statistical tools for examining variables and their relationships were shown in previous lessons, including:

- descriptive statistics – frequency distributions and crosstabulation tables;
- charts or graphs, such as scatterplots and boxplots; and
- more advanced statistics such as correlation coefficients and simple linear regression.

Crosstabulation Tables

First, let's see how the data varies according to the floor number, examining the top floor variable against floor number.

- Select Analyze → Descriptive Statistics → Crosstabs
- Select TopFloor as the Column variable and Floor# as the Row variable
- Click OK to run the crosstab

Floor# * Topflr Crosstabulation

		Topflr		
		0	1	Total
Floor#	1	13	0	13
	2	13	0	13
	3	13	0	13
	4	11	0	11
	5	18	0	18
	6	6	9	15
	7	3	3	6
	8	5	0	5
	9	1	0	1
	11	2	0	2
	12	5	0	5
	13	2	0	2
	14	2	0	2
	15	3	0	3
	16	5	0	5
	17	2	0	2
	18	1	0	1
	19	2	0	2
	20	0	1	1
Total		107	13	120

We see that there are 13 sales of condos on the top floor of their respective buildings; nine are on the 6th floor, three on the 7th, and one on the 20th.

Next, let's look at the sale date range of the sales in our database. Since the sales dates are provided in two separate variables SaleYear and SaleMnth we will use a crosstab.

Use Analyze → Descriptive Statistics → Crosstabs (use month as the row and year as the column):

Crosstabulation: Month of Sale by Year of Sale

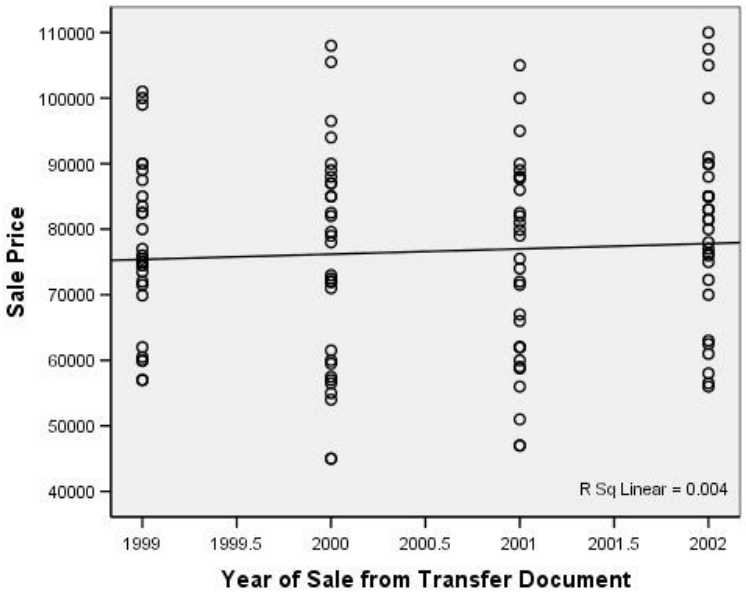
		Year of Sale from Transfer Document				
		1999	2000	2001	2002	Total
Month of Sale from Transfer Document#	1	0	2	2	2	6
	2	1	1	1	4	7
	3	5	2	0	1	8
	4	4	3	2	4	13
	5	3	4	2	0	9
	6	4	2	7	2	15
	7	3	1	3	5	12
	8	2	4	0	2	8
	9	3	4	3	4	14
	10	0	2	1	1	4
	11	3	6	6	2	17
	12	1	1	3	2	7
Total		29	32	30	29	120

The results indicate that the sale dates range from February 1999 to December 2002, a span of 47 months. Because 47 months is a relatively long time in most real estate markets, we are going to need to determine if the sale prices should be time adjusted. Ideally, we want to view prices against date of sale using a scatterplot. We will do so in the next section.

Graphical Analysis

A scatterplot between sale price and year of sale should tell us if a time adjustment is necessary.

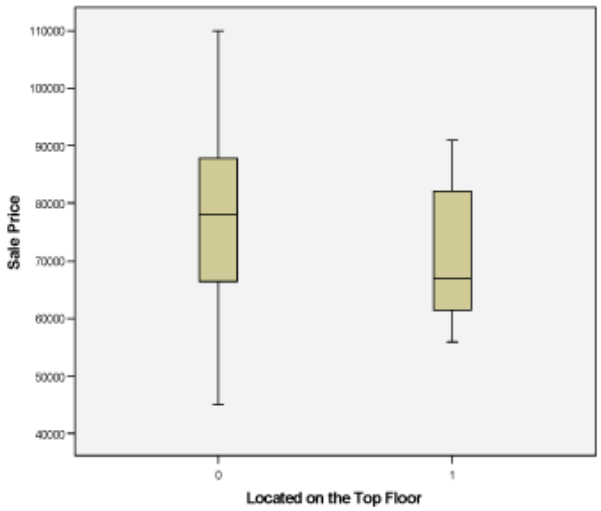
- Select Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → click Define
- Select SalePrice for the Y-axis and SaleYear as the X-axis → OK
- Under Template, select Use Chart Specifications from, and browse to the RSQ1 file created in Lesson 2. This will add the regression line and R-square

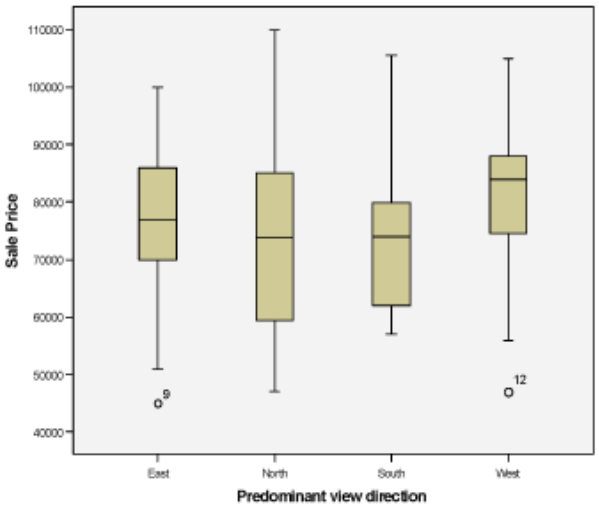


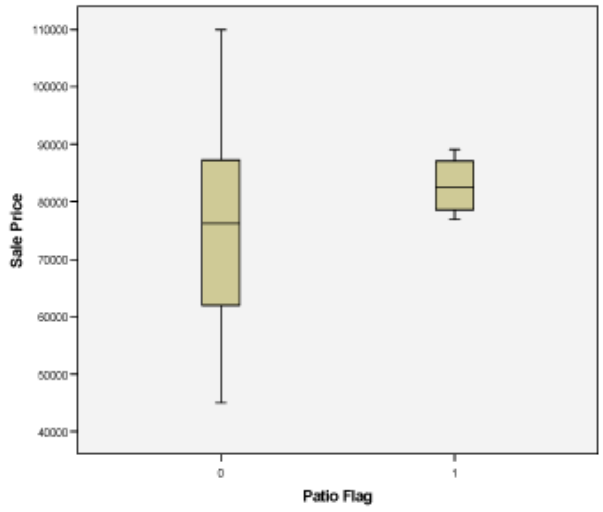
The graph shows the market has been flat for almost four years in this market area. The regression line is nearly flat and the correlation between sale price and year of sale is only 0.004. No time adjustment is necessary.

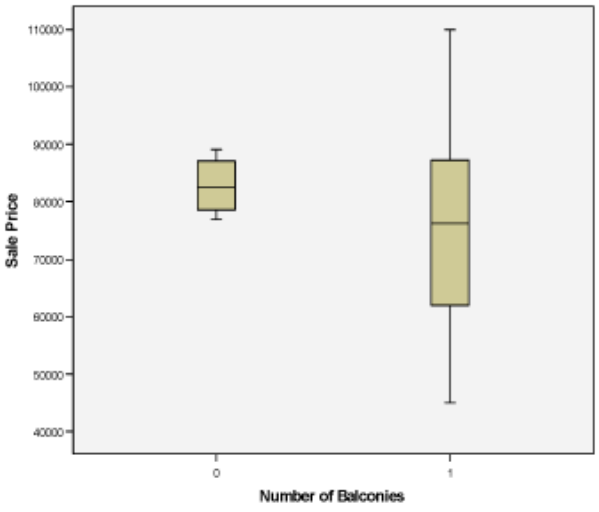
In Lesson 6, we already examined the relationship of sale price with total square footage of living area, floor height, and number of bathrooms. As such we will not reproduce those charts here, and will instead only examine the new variables. The variables that we will examine are: Topflr, Directio, Patiofl, Balc#, Pool, Bedrms, Halfbaths, and the three parking related variables. We are interested in their relationship with selling price – that is, do the selling prices differ significantly for condos with different values for each of the variables? For example, do properties with a pool sell for significantly more or less than those without a pool? As these are all discrete variables, we will use boxplots to show these relationships.

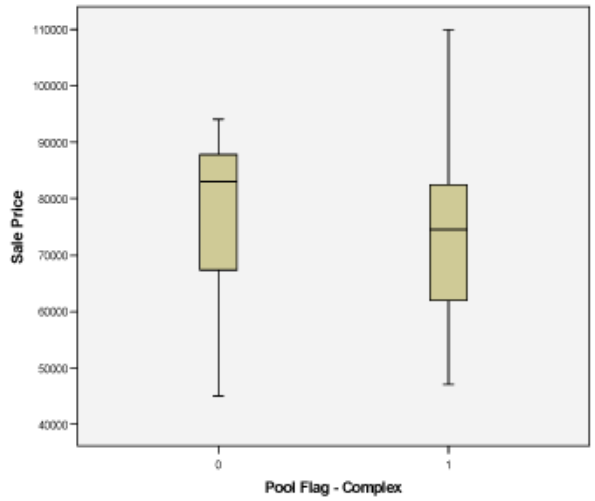
Use Graphs → Legacy Dialogs → Boxplots → Simple → Define:

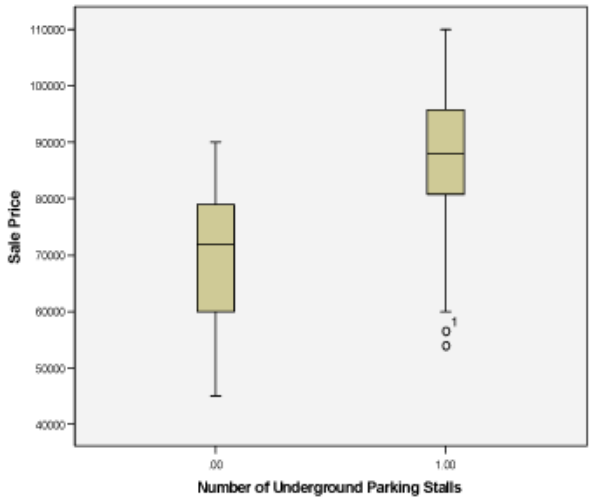


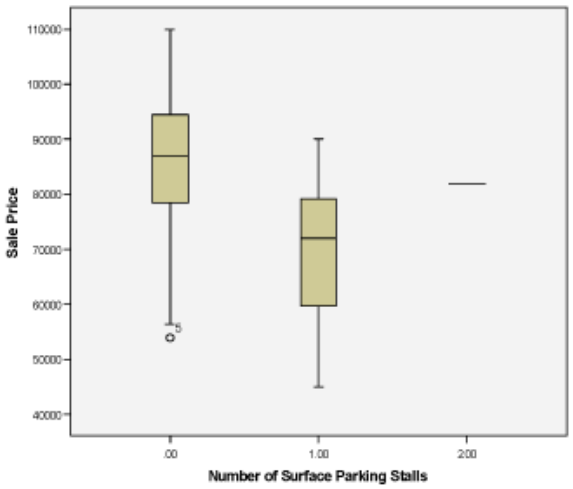


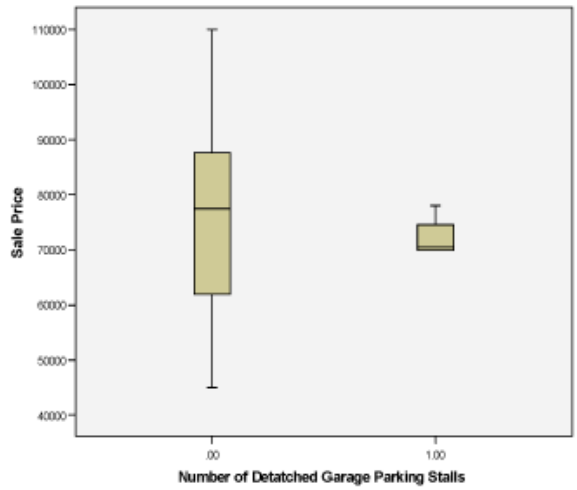


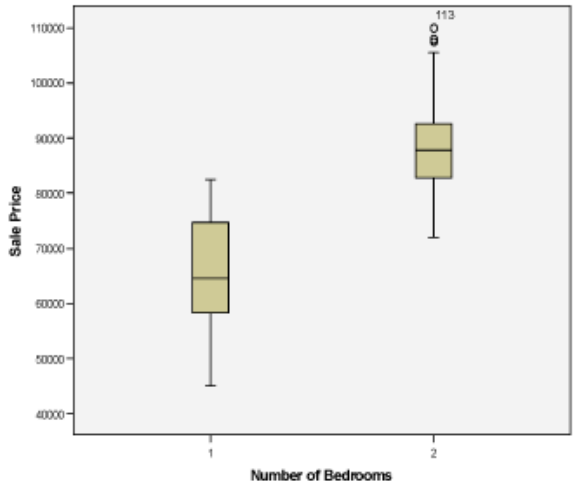


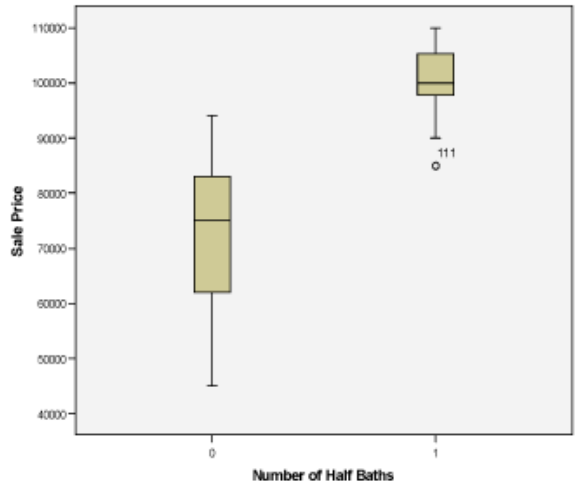












A number of things can be determined from these boxplots:

- There is little difference between the selling prices of condos on the top floor of their building and those not. However, it is interesting to note that the median selling price for condos not on the top floor is actually greater than the median of condos on the top floor.

- The predominant view direction shows some variation in selling price with west appearing to be the most valued view direction, followed by east, and north and south very close.

- The Patio Flag and the Number of Balconies appear to be mutually exclusive, with all sales having either a patio or balcony. This can be confirmed with a crosstab report, or, because we are only dealing with 120 sales, by simply looking at the data. As it turns out, four of the first floor sales have a patio and all the rest of the condos have a balcony (this means that Patio Flag and Balcony are perfectly correlated). Because only four properties out of 120 differ on these characteristics, these variables do not appear to be useful for our model.

- There is some variation in the Pool variable, but in this market it appears the median selling price for condos with a pool in the complex is actually less than those without.

- Parking seems varied, with surface and underground virtual opposites of each other, surface having only a single instance of two spaces and detached having only four occurrences.

- Bedrooms and Half Baths both show significant differences between their two possible values.

Descriptive Statistics

Next we will examine descriptive statistics for the variables selected. Select Analyze → Reports → Case Summaries to produce a statistical summary of the following variables (mean, maximum, and minimum are good statistics to review at this stage):

Total_Area	SalePrice
Bedrms	Bath#

The report indicates:

- total living area ranges from 624 to 1,143 square feet with a mean of 814;
- sale prices range from $45,000 to $110,000 with a mean of $76,593; and
- number of bedrooms and bathrooms are both either one or two.

Use Analyze → Descriptive Statistics → Frequencies to examine the following variables:

Topflr	Parkug
Directio	Parksurf
Abutting	Parkdgar
Patiofl	Pool
Balc#	

A number of things can be seen from these frequencies:

- 13 of the 120 sales are for condos on the top floor;
- the floor number varies a lot, ranging from 1 to 20 with the majority ranging from 1 to 6;
- the predominant view direction is fairly evenly distributed;
- the abutting influences are elevator (11 occurrences), laundry (3), lobby (2), and stairwell (7), while 97 condos have no abutting influences;
- the Patio Flag and the Balcony Flag definitely appear to be mutually exclusive;
- parking seems varied, with only four sales having detached parking and only one with two surface parking stalls; and
- 73 of 120 condos have access to a pool in their complex.

The final few variables of interest that have yet to be investigated are those related to bathrooms.

Use Analyze → Descriptive Statistics → Frequencies to examine these variables:

Bath#	Tqrbath
Fullbath	Halfbath

It can been seen here that there are no three-quarter bathrooms and condos with two bathrooms have one full bathroom and one half bathroom. This means the variables FullBath and Tqrbath will be of no use for our model as they are the same for all sales. We have a perfect correlation between Bath# and HalfBath, because when Bath# is 1, HalfBath is 0, and when Bath# is 2, HalfBath is 1. Therefore, based on our earlier discussion about multicollinearity, we may conclude that only one of these two variables should be included in our model.

Correlation Analysis

In the previous sections, we have seen how some of the variables are related to SalePrice. Correlation analysis will be the final step in this variable examination process, identifying variables that are correlated with each other. We must ensure that our model only includes one of any correlated variables in order to avoid multicollinearity. Some correlations have already been determined: Patiofl and Balco#, Bath# and HalfBath.

Go to Analyze → Correlate → Bivariate…. Select the following variables: SalePrice, Topflr, Floor#, Total_Area, Bedrms, Halfbath, Parkug, Parksurf, Parkdgar, and Pool. Ensure that Pearson is the Correlation Coefficient selected. We have removed the N and Sig lines for clarity and brevity (using the instructions in Lesson 3).

		Bedrms	Floor#	Halfbath	Parkdgar	Parksurf	Parkug	Pool	SalePrice	Topflr	Total_Area
Bedrooms	Pearson Correlation	1	-.136	.404(**)	-.174	-.481(**)	.550(**)	-.413(**)	.759(**)	-.057	.979(**)
Floor#	Pearson Correlation	-.136	1	.411(**)	.062	-.066	.094	.477(**)	.329(**)	.067	-.090
Halfbath	Pearson Correlation	.404(**)	.411(**)	1	-.070	-.433(**)	.471(**)	.303(**)	.609(**)	-.132	.370(**)
Parkdgar	Pearson Correlation	-.174	.062	-.070	1	-.213(*)	-.149	.149	-.055	-.065	-.147
Parksurf	Pearson Correlation	-.481(**)	-.066	-.433(**)	-.213(*)	1	-.918(**)	.282(**)	-.520(**)	.180(*)	-.465(**)
Parkug	Pearson Correlation	.550(**)	.094	.471(**)	-.149	-.918(**)	1	-.335(**)	.570(**)	-.115	.530(**)
Pool	Pearson Correlation	-.413(**)	.477(**)	.303(**)	.149	.282(**)	-.335(**)	1	-.018	-.050	-.435(**)
SalePrice	Pearson Correlation	.759(**)	.329(**)	.609(**)	-.055	-.520(**)	.570(**)	-.018	1	-.129	.770(**)
Topflr	Pearson Correlation	-.057	.067	-.132	-.065	.180(*)	-.115	-.050	-.129	1	-.041
Total_Area	Pearson Correlation	.979(**)	-.090	.370(**)	-.147	-.465(**)	.530(**)	-.435(**)	.770(**)	-.041	1

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

The first set of correlations that are of interest are the ones with Sale Price. The matrix confirms our previous analysis of how variables are related to Sale Price: Total Living Area has the strongest correlation with Sale Price (0.770) and will clearly be a significant variable in the final model.

Of more interest at this point in this analysis, however, are the correlations between the potential independent variables. Where variables are highly correlated (e.g., over 0.8), only one should appear in the model – otherwise, multicollinearity problems will surface in the model.

The correlation between Bedrooms and Total Living Area is very strong (0.979) and therefore one of these will have to be removed from the model. This correlation is not unexpected, because as the number of bedrooms increases in residential property so does the living area, especially in condos. As a result, any effect that number of bedrooms has on value should also be addressed through the Total Living Area variable. Therefore, we will exclude Bedrms from the model. If they were both included, it would lead to multicollinearity, which is a major problem in MRA models. Multicollinearity creates errors in the coefficients and generates meaningless results. **BE CAREFUL of multicollinear variables!**

One other strong (albeit negative) correlation is between Surface Parking Stalls and Underground Parking Stalls (-0.918). It appears that if a condo has underground parking it will NOT have surface parking, and *vice versa*. This was suspected from the results of the boxplot analysis. Therefore, to avoid multicollinearity, only one can be included in the model. We will include Parkug. Remember that Parksurf has that single occurrence of 2 parking stalls – using Parkug removes that issue.

No other correlations are of concern.

Summary of Data Analysis

Our data examination has provided a solid foundation for selecting the variables that should be considered in the model. We have found:

- some variables can be ignored, e.g., Fullbath and Tqrbath which are constant for all sales;
- some variable combinations should be avoided to eliminate multicollinearity issues later, e.g., Bedrms and Total_Area; and
- some variables will need to be modified or transformed into a suitable form to use in an additive regression model, e.g., Abutting and Directio.

In the next step we will carry out these necessary transformations.

STEP 4: Transformations

In the previous step we examined potential independent variables for use in a model to estimate value. We have gained an understanding of how these property characteristics are related to value. However, we need to do two things further before we can create the model. We need to transform any variables that are not in a useable format and then examine these new variables to ensure that they should be included in the model.

We have identified two variables that require transformation: Abutting and Directio. As explained earlier, these variables cannot be used in a regression calculation, because their values will not make sense in a mathematical equation. For example, it is impossible to calculate a coefficient that can multiply by the values "Elevator" or "Lobby".

In order to use these variables in an additive regression equation, the best way to transform them is to create a series of binary variables. In other words, yes/no variables for each of their categories – for example, a new variable for "West View" which equals 1 if there is a west view and 0 if any other view direction.

In creating these binary variables, we must ensure that each category (e.g., West, East, North, South in Directio) has at least five occurrences. If there are too few observations in a category, there is a tendency for the regression process to "chase" these sales. "Chasing" sales in modeling means the model assigns all residual value not explained by the main variables in the equation to the unusual characteristic. As an example, assume that only one condo in the analysis has a fireplace and the property has a sale price of $130,000. If the regression equation creates a value estimate of $105,000 for this property, before adding fireplaces, when the fireplace variable is added to the equation it would be assigned a coefficient of $25,000. This variable would be used to explain all of the remaining value needed to exactly match the sale price, thus "chasing" the sale. A general rule of thumb is to have at least 5 observations in each category, and preferably 5% of the total (or 6 observations for a variable in our 120 sale database).

For Directio, all four view directions have enough occurrences to create binary variables. For Abutting, only Elevator and Stairwell have enough.

Open a new syntax file, type and run the transformations below, and then name the syntax file "Lesson7.SPS" (File → Save As):[3]

```
RECODE Directio ('North'=1) (ELSE=0) INTO NorthView.
RECODE Directio ('South'=1) (ELSE=0) INTO SouthView.
RECODE Directio ('East'=1) (ELSE=0) INTO EastView.
RECODE Abutting ('Elevator'=1) (ELSE=0) INTO Elevator.
RECODE Abutting ('Stairwel'=1) (ELSE=0) INTO Stairwell.
EXECUTE.
```

Helpful Hint!

The commands in syntax file must be typed in EXACTLY as shown, with matching quotes around each value and periods at the end of each line. There are no spaces in the new variable names NorthView, SouthView and EastView. SPSS does not allow spaces in variable names – if you want a variable name to "look like" it contains a space use the "_" (under-score) character in place of a space.

The first transformation creates a new variable called NorthView which will have a value of 1 if the value of Directio is "North" and otherwise will have a value of zero. In essence it tells SPSS to recode (or translate) the variable Directio into a new variable called NorthView – if Directio is equal to "North", then NorthView is given the value of 1; if Directio is anything else NorthView is given the value 0. The rest of the transformations do the same thing with the other values of Directio and the two values of Abutting. Recoding a variable into a series of binary variables is a very useful and commonly used transformation in regression modeling; it is important that you understand what is being done here.

A common question at this point is: "why did we only create three view direction variables, why not create a WestView variable as well?" The short answer is that in NOT creating a WestView variable, we can better interpret the model. Because we have only created binary variables for North, South, and East views, the reference value for View Direction is West. Since all condos have a View Direction, keeping West as the reference value means that any coefficient for a North, South, or East view will be relative to a West view. In other words, West is the base view for the regression equation and is automatically considered in the estimated selling price – the other views, if added into the equation, will either add or subtract from the selling price for a west view unit. For example, if the coefficient for a South view was calculated to be $5,000, that would represent the dollar amount a South view is worth over a West view. If the coefficient for a North view was -$2,500, this implies a North view is considered $2,500 inferior to a West view to purchasers in this market. Often, we aim to have the reference be the most prevalent value – in our case, West is the most common value in Directio.

Before continuing, you should check the results of these transformations. As they are fairly straightforward, you can simply review the variables in the Data View window to ensure the new binaries are correct in comparison to the source variable. Alternatively, you can use Analyze → Descriptive Statistics → Crosstabs. Two crosstab examples are given below: Abutting and the new binary Elevator; and Directio and the new binary NorthView. Your output should look similar to the following tables:

[3] To run the command in the syntax file, select the entire command and click Run (Play arrow icon). The Execute command will create the new variable with the values specified. If you did not include the Execute command, then you need to select Transform → Run Pending Transformations. Alternatively, you may create these transformations using Transform → Recode. The advantage of using syntax files for transformations is that you have a record of all transformations carried out. This can help you later in reviewing transformations made. Also, if you ever have a data problem or lose your data (e.g., computer theft, hard drive crash, fire), you can quickly and easily restore your data from the original database. This is a good time to remind you: **backup your data frequently**! You should save it periodically to a new file in your hard drive (so you can go back a few steps if you do something wrong later) and you should also save it to a disk or other computer, ideally in a different location, in case of a computer problem.

Unit abuts specific influences * Elevator Crosstabulation

		Elevator		
		.00	1.00	Total
Unit abuts specific influences	Elevator	97	0	97
	Laundry	0	11	11
	Lobby	3	0	3
	Stairwel	2	0	2
Total		7	0	7
		109	11	120

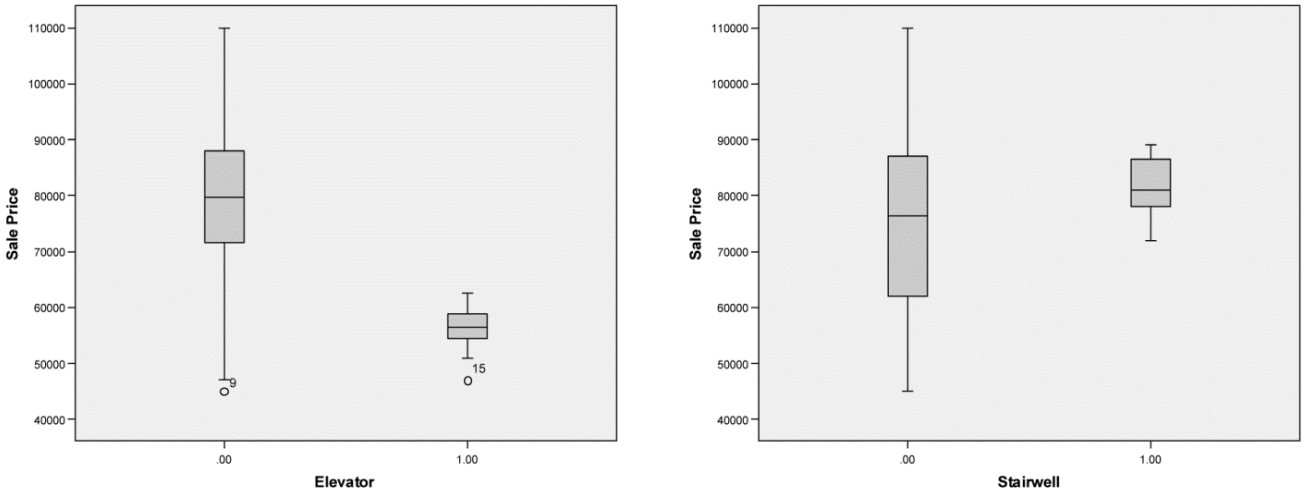
Predominant view direction * NorthView Crosstabulation

		NorthView		
		.00	1.00	Total
Predominant view direction	East	33	0	33
	North	0	24	24
	South	29	0	29
	West	34	0	34
Total		96	24	120

The binaries all appear correct. We can now proceed to examine the new variables for use in the model.

STEP 5: Examining the Transformed Variables

We now need to examine these new variables using the techniques applied in Step 3. We will use boxplots and a correlation matrix. Because we already examined View Direction in Step 3 with a boxplot, we will only look at the boxplots for Elevator and Stairwell here.



The boxplots show that condos abutting an elevator sell for quite a bit less than those that do not. Proximity to a stairwell does not seem to influence the selling price much at all.

A new correlation matrix with the five new binary variables added and Bedrms, Parksurf, and Parkdgar removed shows no correlation concerns.

The new variables have been examined and we are ready to proceed to Step 6. The table following the correlation matrix summarizes the results of our variable analysis process.

Correlations

	Sale Price	Locate d Top Floor	Location of unit - Floor	Total Living Area -square feet	No. of Half Baths	No. of u/g Parking Stalls	Pool Flag - Complex	NorthView	SouthView	EastView	Elevator	Stairwell
Sale Price	1	-.129	.329(**)	.770(**)	.609(**)	.570(**)	-.018	-.046	-.075	-.047	-.440(**)	.083
Located on the Top Floor	-.129	1	.067	-.041	-.132	-.115	-.050	-.040	.116	-.035	-.111	-.087
Location of the unit - Floor	.329(**)	.067	1	-.090	.411(**)	.094	.477(**)	.086	.089	-.104	-.219(*)	-.150
Total Living Area - square ft	.770(**)	-.041	-.090	1	.370(**)	.530(**)	-.435(**)	-.159	-.211(*)	.119	-.323(**)	.143
Number of Half Baths	.609(**)	-.132	.411(**)	.370(**)	1	.471(**)	.303(**)	.126	.081	-.120	-.120	-.094
Number of Underground Parking Stalls	.570(**)	-.115	.094	.530(**)	.471(**)	1	-.335(**)	-.102	-.254(**)	.118	-.137	-.200(*)
Pool Flag - Complex	-.018	-.050	.477(**)	-.435(**)	.303(**)	-.335(**)	1	.401(**)	.453(**)	-.385(**)	-.100	.200(*)
NorthView	-.046	-.040	.086	-.159	.126	-.102	.401(**)	1	-.282(**)	-.308(**)	.202(*)	.142
SouthView	-.075	.116	.089	-.211(*)	.081	-.254(**)	.453(**)	-.282(**)	1	-.348(**)	-.179	.192(*)
EastView	-.047	-.035	-.104	.119	-.120	.118	-.385(**)	-.308(**)	-.348(**)	1	.128	-.153
Elevator	-.440(**)	-.111	-.219(*)	-.323(**)	-.120	-.137	-.100	.202(*)	-.179	.128	1	-.079
Stairwell	.083	-.087	-.150	.143	-.094	-.200(*)	.200(*)	.142	.192(*)	-.153	-.079	1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

No.	Name	Description	Action
1	Condo#	Condo Complex ID	will not use–information only
2	Market	Market Area–all are South	will not use–same for all sales
3	Unit#	Unit Number	will not use–information only
4	Topflr	Located on the Top Floor	use as is
5	Floor#	Location of the unit - Floor	use as is
6	Directio	Predominant view direction	transformed into three binary variables
7	Abutting	Abutting influence	transformed into two binary variables
8	Total_Area	Total Living Area - sqft	use as is
9	Bedrms	Number of Bedrooms	will not use–correlated with Total Area
10	Bath#	Number of Bathrooms	will not use–use Halfbath instead
11	Fullbath	Number of Full Bathrooms	will not use–all are '1'
12	Tqrbath	Number of Three Quarter Bathrooms	will not use–all are '0'
13	Halfbath	Number of Half Bathrooms	use as is–proxy for number of bathrooms
14	Patiofl	Patio Flag - '1' unit has a patio	will not use–only four are '1'
15	Balc#	Number of Balconies	will not use–only four are '0'
16	Parkug	Num. of Underground Parking Stalls	use as is
17	Parksurf	Num. of Surface Parking Stalls	will not use–correlated with Parkug
18	Parkdgar	Num. of Detatched Garage Parking Stalls	will not use–only four are '1'
19	SaleYear	Year of Sale	do not use–time adjustment unnecessary
20	SaleMnth	Month of Sale	do not use–time adjustment unnecessary
21	SalePrice	Sale Price	dependent variable
22	Pool	Pool Flag	use as is
23	NorthView	North View	use as is–from Directio
24	SouthView	South view	use as is–from Directio
25	EastView	East view	use as is–from Directio
26	Elevator	Next to Elevator	use as is–from Abutting
27	Stairwell	Next to Stairwell	use as is–from Abutting

STEP 6: List the Variables for Calibration

The final step in the variable selection process is to provide the final list of variables for use in calibrating the model. To determine which final variables to focus on, we will use regression to evaluate the candidate variables. From the preceding analysis we can determine the following list of candidate variables:

Topflr	Total_Area	Parkug	NorthView	EastView	Stairwell
Floor#	Halfbath	Pool	SouthView	Elevator	

We will run a regression for these variables with the dependent variable SalePrice. We will then run further iterations of this model with different groupings of variables to see which produces the best results.

- Select Analyze → Regression → Linear and enter the above variables as Independent variables and SalePrice as the Dependent variable. Ensure the Method is set to Enter.
- Click the Statistics button, and select: Estimates, Model Fit, Descriptives, and Collinearity Diagnostics.
- Click Continue to close Statistics window and OK to run the model.

Helpful Hint!

In any SPSS procedure, the Paste function will take whatever you have just set up through mouse clicks and paste it into a syntax file. In this case, having your regression criteria in a syntax file allows for easy changes, such as adding or removing a variable. Once in the syntax file you simply block select the regression section of the file and run it. This syntax file procedure for running regressions can save you time when you need to run the same or similar regressions in a number of different models, since SPSS will save the input criteria after the program has been closed and re-opened. You may find this helpful in practical applications of these procedures in your real estate work. However, if you prefer to instead use the "point and click" features in the Linear Regression window, you will obtain exactly the same output.

Your SPSS output should include the following:

Descriptive Statistics

	Mean	Std. Deviation	N
Sale Price	76593.50	14903.185	120
Located on the Top Floor	.11	.312	120
Location of the unit - Floor	6.37	4.894	120
Total Living Area - square feet	814.05	159.126	120
Number of Half Baths	.13	.332	120
Number of Underground Parking Stalls	.3917	.49017	120
Pool Flag - Complex	.61	.490	120
NorthView	.2000	.40168	120
SouthView	.2417	.42989	120
EastView	.2750	.44839	120
Elevator	.0917	.28976	120
Stairwell	.0583	.23536	120

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet, Pool Flag - Complex(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Sale Price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.913(a)	.834	.817	6375.058

a Predictors: (Constant), Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet, Pool Flag - Complex

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22041218615.200	11	2003747146.836	49.303	.000(a)
	Residual	4389267714.801	108	40641367.730		
	Total	26430486330.001	119			

a Predictors: (Constant), Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet, Pool Flag - Complex

b Dependent Variable: Sale Price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	13410.238	6033.541		2.223	.028		
	Located on the Top Floor	-3896.922	1995.883	-.082	-1.952	.053	.880	1.136
	Location of the unit - Floor	674.245	159.813	.221	4.219	.000	.558	1.791
	Total Living Area - square feet	66.884	6.384	.714	10.477	.000	.331	3.022
	Number of Half Baths	3913.247	2938.689	.087	1.332	.186	.359	2.789
	Number of Underground Parking Stalls	5150.981	1709.934	.169	3.012	.003	.486	2.057
	Pool Flag - Complex	6768.240	2676.125	.223	2.529	.013	.198	5.038
	NorthView	-1764.232	2271.773	-.048	-.777	.439	.410	2.438
	SouthView	-1774.640	2163.589	-.051	-.820	.414	.395	2.533
	EastView	-1800.250	1581.447	-.054	-1.138	.257	.679	1.472
	Elevator	-5442.038	2493.891	-.106	-2.182	.031	.654	1.529
	Stairwell	297.423	3030.415	.005	.098	.922	.671	1.489

a Dependent Variable: Sale Price

The first report shown, **Descriptive Statistics**, provides the mean, standard deviation, and N for each of the variables selected. This information is not needed at this point.

The **Variables Entered/Removed** report shows all of the variables entered into the model. In the Linear Regression window, we used "Enter" as the Method, which means that all variables selected are entered into the model. This differs from an alternative method called "Stepwise", which selects the variables in sequence or steps based on the importance they have in explaining the dependent variable. This method will be used later, when we want to ensure that the variables in the model contribute significantly to explaining the variation in sale price. At this point in the analysis, we only want to determine which group of variables in total best explains sale price. In other words, we want to find which group of variables produces the highest R^2 value for explaining variations in Sale Price.

The **Model Summary** report displays the relevant statistics for the whole model. The important statistics here are R^2, adjusted R^2, and the Standard Error of the Estimate (SEE). In comparing models, the most important statistics are adjusted R^2 and SEE. At this stage of the analysis, the best group of variables is the one that produces the highest adjusted R^2 and lowest SEE.

The **ANOVA** report is the **AN**alysis **O**f **VA**riance report. It provides the data necessary for calculating the F-statistic. The F-statistic measures performance of the model overall when compared to the result that would be

obtained by estimating the sale price by simply using the mean sale price. With the number of sales used here, and the relatively small number of variables in the model, the F value of 49 is about what would be expected. The Sig. for F is .000, which is less than the critical value of .05. These are positive results for this model.

The **Coefficients** report shows several statistics relevant to the selection of the variables. The first is the unstandardized coefficients. These are the estimated values for each of the variables that would be used to estimate the sale price; e.g., the extra half bathroom in a condo is shown to add $3,913 to value. At this stage of the analysis the exact values of these coefficients are not very useful, but they should be checked to see if they have the expected signs – that is, whether they are positive, adding to value, or negative, subtracting from value, as well as the expected magnitude. For example, Living Area adds $66.88 per square foot, which makes sense in sign (positive, bigger size adds to value) and magnitude ($67 per square foot is not an unreasonable number in appraisal terms). However, a few of the coefficients are puzzling here:

- the negative influence of being on the top floor (our boxplot earlier also showed this);
- the availability of a pool seems to be worth a great deal; and
- all the view variables have virtually the same negative value (remember this is in comparison to the West view).

These issues will be discussed further after the other statistics have been reviewed.

Next is the standard error of the coefficient. If the value assigned to the coefficient is significant, then the standard error will be low relative to the coefficient itself. This value and the coefficient value are used to compute the *t*-statistic located in the fifth column. The next column to the right of the *t*-statistic is the significance value of the *t*-statistic. Because the *t*-statistic is dependent on the number of observations, a specific value for acceptance or rejection that works in all situations cannot be given. However, a value of approximately 2 generally indicates significance at the 95% confidence level, and a value approximately 1.6 indicates significance at the 90% confidence level. These translate into values of .05 and .10 respectively for the Sig. value. Interpreting this further, Sig. values less than .05 indicate that the coefficient is significantly different from zero at the 95% confidence level and a value less than .10 means that the coefficient is significantly different from zero at the 90% confidence level. Another interpretation is that the Sig. value is the probability that the coefficient is equal to zero, indicating that the variable provides no useful information to the model. In the cases of Stairwell, North View, South View, and East View, the Sig. values are 0.922, 0.439, 0.414 and 0.257 respectively, showing a high probability that none of these variables are useful in the model. In the next lesson, we will investigate the use of this statistic more completely in the variable selection process.

The next statistic to be discussed is the Standardized Coefficient or Beta value. This statistic indicates the relative importance of the variable in the model. High numeric or absolute values show that the variable is very important, while low values show that the variable does little to contribute to the estimate. The highest value is for Total_Area. This result is what would be expected. The lowest values are those for Stairwell, the View variables, and Half Baths; these also have high Sig. values, indicating they may not be providing useful information to the model.

The last two columns show two measures of multicollinearity: Tolerance and VIF (variance inflation factor). These two measures are inversely related, since VIF = 1 ÷ Tolerance. The tolerance is calculated by creating a model which assumes that the variable in question is the dependent variable and the remaining variables are the independent variables, and then determining the R^2 of this model. This R^2 will indicate how well the other variables explain any given variable, and thus indicate the degree of multicollinearity. Tolerance is 1 minus this R^2 and, therefore, the higher the tolerance figure (closer to 1), the lower the multicollinearity of that variable. Tolerance values less than approximately 0.3 are considered to show a degree of multicollinearity that can have a serious effect on the value of the coefficients of the variable in question (in other words, a correlation of 70% or more with the other variables). As the tolerance and VIF are directly inversely related, values of VIF over 3.3 are a serious concern.

Pool, with a tolerance of 0.198 and a VIF of 5.038, appears to have a serious multicollinearity problem, which may be causing some of the other issues mentioned above.

The last report produced by the system is labelled Collinearity Diagnostics (this report is not included above). As the values in this report have already been summarized by the tolerance and VIF statistics in the coefficients section, this report is not required. It is automatically produced when Collinearity Diagnostics are requested, and while we ignore this table, you must request Collinearity Diagnostics in order to obtain the tolerance and VIF statistics we do need.

We will now re-run the regression without Pool to eliminate the multicollinearity issue uncovered by its VIF statistic. Your SPSS output should include the following:

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Sale Price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.908(a)	.824	.808	6530.962

a Predictors: (Constant), Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21781258428.719	10	2178125842.872	51.066	.000(a)
	Residual	4649227901.282	109	42653466.984		
	Total	26430486330.001	119			

a Predictors: (Constant), Stairwell, Elevator, Located on the Top Floor, EastView, Number of Half Baths, SouthView, Location of the unit - Floor, Number of Underground Parking Stalls, NorthView, Total Living Area - square feet
b Dependent Variable: Sale Price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	22149.956	5066.881		4.372	.000		
	Located on the Top Floor	-4828.659	2009.559	-.101	-2.403	.018	.911	1.097
	Location of the unit - Floor	842.282	148.901	.277	5.657	.000	.675	1.481
	Total Living Area - square feet	58.675	5.632	.626	10.419	.000	.446	2.240
	Number of Half Baths	7470.506	2643.331	.166	2.826	.006	.465	2.150
	Number of Underground Parking Stalls	3772.799	1660.417	.124	2.272	.025	.541	1.848
	NorthView	1197.615	1994.265	.032	.601	.549	.559	1.790
	SouthView	734.319	1969.779	.021	.373	.710	.500	2.001
	EastView	-1650.236	1618.982	-.050	-1.019	.310	.680	1.470
	Elevator	-7412.777	2426.967	-.144	-3.054	.003	.725	1.380
	Stairwell	2477.143	2976.326	.039	.832	.407	.730	1.369

a Dependent Variable: Sale Price

The results are reasonable in relation to the previous model calculation. The adjusted R^2 is slightly lower at 0.808 and the SEE is slightly higher at 6,531 (the COV is 8.53%, or 6,531 ÷ 76,593.5, which is a very good result, below the 10% target level). As you can see, the multicollinearity introduced by the Pool variable gave us artificially good results. This is the unfortunate allure and trap of multicollinearity – pay close attention to your VIF (or tolerance). The F statistic is slightly improved at 51.1, again with a Significance of 0.000.

In general, the coefficients look reasonable. Purchasers prefer a North view, followed by a South view, then West (base view), and finally East. An extra bathroom is worth almost $7,500 and underground parking is worth over $3,700. Proximity to a Stairwell is worth an additional $2,477, while being next to an Elevator is a negative influence at -$7,413. Each square foot of living area adds $58.68, while each floor adds $842.28. Perhaps the only unusual coefficient is that being on the top floor of the complex is a negative influence of almost $5,000.

All of the VIF (and tolerance) statistics look fine, with VIF < 3.333 and tolerance > 0.3. No variables even approach these critical values, indicating no evidence of multicollinearity in the model.

If we were building a predictive model we might stop here. The aim of a predictive model is to ensure that the adjusted R^2 is as high as possible and the SEE is as low as possible. This is usually achieved by including more variables in the model than is warranted by the *t*-statistics. For example, Northview, Southview, Eastview, and Stairwell all have *t*-statistics below the critical value of ± 1.6, implying their coefficients are not statistically significant (we are not 95% confident they are any different than zero). Southview, in particular, is extremely low and the addition of this and the other questionable variables may mean that the coefficients for all variables are unreliable – e.g., we cannot rely upon $58 being an accurate per square foot indicator size for living area. However, in a predictive model the individual coefficients do not matter; all that is important is the overall high R^2 and low SEE. Shortly, we will continue with developing a compromise predictive and explanatory model, but for now we will pause to briefly examine this model as a predictor.[4]

Assume we were using this model to predict the selling price of a condo in the South market area with the following characteristics:

- 1,010 square feet;
- a view to the North;
- on the fifth floor of seven; and
- an underground parking stall.

Its estimated selling price would be $90,600 (rounded), calculated as:

$$22{,}149.96 - 4{,}828.66 \times (0) + 842.28 \times (5) + 58.68 \times (1{,}010) + 7{,}470.51 \times (0) + 3{,}772.80 \times (1) + 1{,}197.62 \times (1) + 734.32 \times (0) - 1{,}650.24 \times (0) - 7{,}412.78 \times (0) + 2{,}477.14 \times (0) = \$90{,}598.58$$

Consider another condo in the same area with following characteristics:

- 900 square feet;
- a view to the West;
- an extra half bathroom;
- is next to the stairwell;
- on the tenth floor of ten; and
- a surface parking stall.

Its estimated selling price would be $88,500 (rounded).

[4] We will see in the continuation of Step 6 that this group of variables does not actually provide the best combination of R^2 and SEE, and therefore would not be the best predictive model either. We have stopped here for demonstration purposes only.

Appraisal judgement and local knowledge would then be applied to determine if these values make sense for the market area.

We will now carry on beyond this predictive model to create a model with more explanatory power. A pure explanatory model would focus on creating the highest quality coefficients possible, at the expense of maybe not estimating selling price as accurately. Instead, we are going to follow a middle course, developing a compromise predictive/explanatory model which allows for both a good prediction of market value and variable coefficients that are reasonable and rational from an appraisal perspective. In order to achieve this dual goal, when evaluating variables for inclusion in the model we will generally apply the following criteria:

- *t*-statistics outside the interval of $\pm$ 1.6 (approximately);
- Significance levels of approximately .10 or lower (90% confidence interval); and,
- Meaningful reduction in the standard error (say .1%).

These criteria are not clear-cut in all cases and occasionally must be tempered by appraisal judgement. Occasionally there will be conflicts in achieving all three of these standards for some variables.

Our next step is to remove extraneous or unnecessary variables from our model. We could do this manually, as we have been already doing in this lesson, by removing a variable, re-running the results, and comparing to see which group is best. Or, alternatively, we can use a procedure called *stepwise regression*. This introduces one independent variable at a time into the model, starting with the variable with the most impact and then adding further variables in order of importance. Stepwise regression tests the model at each "step" and removes a variable if it no longer passes the test to remain in the model (due to other variables being added to the model). The modeller provides the critical values the program will apply to determine whether variables should be entered into the model or removed from it. The critical values are called probabilities of F, based on the F-statistic. In the stepwise regression module in SPSS, these are called Entry or "PIN" and Removal or "POUT" – in other words, the probability level to be brought into the model and probability level to be left out of (or removed from) the model.

Looking at the model produced thus far, the Coefficients table shows several of the variables having a low *t*-statistic, with their absolute value less than 1.6. These variables are likely not helping the model much, as the statistics show they have a minimal influence (not significantly different than zero). In fact, their inclusion may be clouding the results from the variables that really do matter in estimating sale price. The view variables and stairwell all have very high Significance values. If we set the Entry probability (PIN) as .15 and the Removal at .20, this will eliminate any variables with low *t*-statistics.

- Analyze → Regression → Linear.
- Enter SalePrice as the Dependent variable.
- Enter Topflr, Floor#, Total_Area, HalfBath, Parkug, NorthView, SouthView, EastView, Elevator, and Stairwell as Independent variables.
- Set Method to Stepwise.
- Click the Statistics button, and select: Estimates, Model Fit, Descriptives, R squared Change, and Collinearity Diagnostics. Click Continue.
- Click the Options button, and select Use probability of F. Enter .15 for Entry and .20 for Removal. Click Continue.
- Click Paste or Click OK to run. If you used Paste, you should go to the syntax file, block and select the regression portion, then run it.

Helpful Hint!

If you saved the regression instructions from your syntax file from earlier, you may modify this instead of the steps above. Change the METHOD from ENTER to STEPWISE, change PIN to .15 and POUT to .20, and add CHANGE after TOL. Then block select the regression section and run the procedure.

Your SPSS output should include the following:

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	Total Living Area - square feet	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
2	Location of the unit - Floor	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
3	Number of Half Baths	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
4	Elevator	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
5	Located on the Top Floor	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
6	Number of Underground Parking Stalls	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).
7	EastView	.	Stepwise (Criteria: Probability-of-F-to-enter <= .150, Probability-of-F-to-remove >= .200).

a Dependent Variable: Sale Price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.770(a)	.592	.589	9556.241	.592	171.422	1	118	.000
2	.867(b)	.752	.748	7480.828	.160	75.556	1	117	.000
3	.886(c)	.784	.779	7010.071	.032	17.242	1	116	.000
4	.895(d)	.801	.794	6767.888	.016	9.450	1	115	.003
5	.901(e)	.812	.804	6605.369	.011	6.729	1	114	.011
6	.904(f)	.817	.807	6541.008	.005	3.254	1	113	.074
7	.906(g)	.822	.810	6487.823	.005	2.860	1	112	.094

a Predictors: (Constant), Total Living Area - square feet
b Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor
c Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths
d Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator
e Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor
f Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls
g Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls, EastView

ANOVA(h)

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	26430486330.001	119			
2	Regression	19882840560.910	2	9941420280.455	177.643	.000(b)
	Residual	6547645769.091	117	55962784.351		
	Total	26430486330.001	119			
3	Regression	20730119265.208	3	6910039755.070	140.616	.000(c)
	Residual	5700367064.793	116	49141095.386		
	Total	26430486330.001	119			
4	Regression	21162991344.877	4	5290747836.220	115.508	.000(d)
	Residual	5267494985.124	115	45804304.218		
	Total	26430486330.001	119			
5	Regression	21456563557.011	5	4291312711.402	98.355	.000(e)
	Residual	4973922772.990	114	43630901.517		
	Total	26430486330.001	119			
6	Regression	21595804819.399	6	3599300803.233	84.126	.000(f)
	Residual	4834681510.602	113	42784792.129		
	Total	26430486330.001	119			
7	Regression	21716200067.155	7	3102314295.308	73.703	.000(g)
	Residual	4714286262.846	112	42091841.633		
	Total	26430486330.001	119			

a Predictors: (Constant), Total Living Area - square feet
b Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor
c Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths
d Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator
e Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor
f Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls
g Predictors: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls, EastView
h Dependent Variable: Sale Price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	17918.029	4565.621		3.925	.000		
	Total Living Area - square feet	72.078	5.505	.770	13.093	.000	1.000	1.000
2	(Constant)	7368.849	3774.496		1.952	.053		
	Total Living Area - square feet	75.472	4.327	.806	17.441	.000	.992	1.008
	Location of the unit - Floor	1223.056	140.706	.402	8.692	.000	.992	1.008
3	(Constant)	14911.963	3976.207		3.750	.000		
	Total Living Area - square feet	67.036	4.535	.716	14.781	.000	.793	1.261
	Location of the unit - Floor	923.313	150.319	.303	6.142	.000	.763	1.310
	Number of Half Baths	9857.911	2374.074	.220	4.152	.000	.664	1.505
4	(Constant)	20813.611	4292.103		4.849	.000		
	Total Living Area - square feet	61.543	4.729	.657	13.013	.000	.680	1.471
	Location of the unit - Floor	782.901	152.143	.257	5.146	.000	.694	1.440
	Number of Half Baths	10917.391	2317.820	.243	4.710	.000	.650	1.539
	Elevator	-7290.794	2371.637	-.142	-3.074	.003	.815	1.227
5	(Constant)	21183.259	4191.459		5.054	.000		
	Total Living Area - square feet	61.659	4.616	.658	13.358	.000	.680	1.471
	Location of the unit - Floor	825.096	149.378	.271	5.524	.000	.686	1.457
	Number of Half Baths	9943.518	2293.105	.222	4.336	.000	.632	1.582
	Elevator	-7862.467	2325.154	-.153	-3.381	.001	.808	1.238
	Located on the Top Floor	-5152.607	1986.400	-.108	-2.594	.011	.954	1.048
6	(Constant)	23283.337	4310.777		5.401	.000		
	Total Living Area - square feet	57.961	5.009	.619	11.570	.000	.566	1.767
	Location of the unit - Floor	819.252	147.958	.269	5.537	.000	.686	1.458
	Number of Half Baths	8707.404	2371.890	.194	3.671	.000	.579	1.726
	Elevator	-8037.283	2304.537	-.156	-3.488	.001	.806	1.240
	Located on the Top Floor	-4912.497	1971.543	-.103	-2.492	.014	.950	1.053
	Number of Underground Parking Stalls	2787.104	1544.948	.092	1.804	.074	.627	1.595
7	(Constant)	22682.446	4290.462		5.287	.000		
	Total Living Area - square feet	59.307	5.032	.633	11.786	.000	.552	1.813
	Location of the unit - Floor	831.171	146.924	.273	5.657	.000	.684	1.462
	Number of Half Baths	7843.257	2407.451	.175	3.258	.001	.553	1.807
	Elevator	-7340.456	2322.636	-.143	-3.160	.002	.781	1.281
	Located on the Top Floor	-5002.567	1956.237	-.105	-2.557	.012	.949	1.054
	Number of Underground Parking Stalls	3123.350	1545.230	.103	2.021	.046	.617	1.622
	EastView	-2357.608	1394.010	-.071	-1.691	.094	.905	1.105

a Dependent Variable: Sale Price

Excluded Variables(h)

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Located on the Top Floor	-.098(a)	-1.670	.098	-.153	.998	1.002	.998
	Location of the unit - Floor	.402(a)	8.692	.000	.626	.992	1.008	.992
	Number of Half Baths	.376(a)	7.064	.000	.547	.863	1.158	.863
	Number of Underground Parking Stalls	.225(a)	3.382	.001	.298	.719	1.391	.719
	NorthView	.079(a)	1.332	.185	.122	.975	1.026	.975
	SouthView	.092(a)	1.538	.127	.141	.955	1.047	.955
	EastView	-.141(a)	-2.431	.017	-.219	.986	1.014	.986
	Elevator	-.213(a)	-3.605	.000	-.316	.896	1.116	.896
	Stairwell	-.027(a)	-.460	.646	-.043	.980	1.021	.980
2	Located on the Top Floor	-.124(b)	-2.753	.007	-.248	.994	1.006	.988
	Number of Half Baths	.220(b)	4.152	.000	.360	.664	1.505	.664
	Number of Underground Parking Stalls	.150(b)	2.800	.006	.252	.699	1.431	.699
	NorthView	.050(b)	1.063	.290	.098	.969	1.032	.969
	SouthView	.063(b)	1.334	.185	.123	.950	1.052	.950
	EastView	-.104(b)	-2.275	.025	-.207	.977	1.024	.977
	Elevator	-.109(b)	-2.207	.029	-.201	.833	1.200	.833
	Stairwell	.029(b)	.624	.534	.058	.961	1.041	.961
3	Located on the Top Floor	-.095(c)	-2.187	.031	-.200	.963	1.039	.643
	Number of Underground Parking Stalls	.092(c)	1.712	.090	.158	.631	1.584	.600
	NorthView	.016(c)	.349	.728	.033	.934	1.070	.640
	SouthView	.034(c)	.761	.448	.071	.926	1.080	.647
	EastView	-.078(c)	-1.792	.076	-.165	.954	1.048	.649
	Elevator	-.142(c)	-3.074	.003	-.276	.815	1.227	.650
	Stairwell	.049(c)	1.120	.265	.104	.950	1.053	.657
4	Located on the Top Floor	-.108(d)	-2.594	.011	-.236	.954	1.048	.632
	Number of Underground Parking Stalls	.100(d)	1.933	.056	.178	.630	1.588	.567
	NorthView	.038(d)	.879	.381	.082	.909	1.100	.631
	SouthView	-.005(d)	-.103	.918	-.010	.848	1.179	.599
	EastView	-.056(d)	-1.296	.198	-.120	.922	1.085	.629
	Stairwell	.042(d)	.975	.332	.091	.946	1.057	.643
5	Number of Underground Parking Stalls	.092(e)	1.804	.074	.167	.627	1.595	.566
	NorthView	.038(e)	.891	.375	.084	.909	1.100	.615
	SouthView	.009(e)	.197	.844	.019	.837	1.195	.597
	EastView	-.060(e)	-1.424	.157	-.133	.921	1.086	.611
	Stairwell	.031(e)	.739	.462	.069	.936	1.068	.624
6	NorthView	.047(f)	1.107	.270	.104	.898	1.114	.559
	SouthView	.030(f)	.656	.513	.062	.788	1.270	.532
	EastView	-.071(f)	-1.691	.094	-.158	.905	1.105	.552
	Stairwell	.059(f)	1.362	.176	.128	.851	1.175	.530
7	NorthView	.028(g)	.637	.526	.060	.818	1.223	.544
	SouthView	.012(g)	.247	.805	.023	.738	1.354	.521
	Stairwell	.048(g)	1.086	.280	.103	.824	1.214	.510

a Predictors in the Model: (Constant), Total Living Area - square feet
b Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor
c Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths
d Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator
e Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor
f Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls
g Predictors in the Model: (Constant), Total Living Area - square feet, Location of the unit - Floor, Number of Half Baths, Elevator, Located on the Top Floor, Number of Underground Parking Stalls, EastView
h Dependent Variable: Sale Price

The output shows the seven steps taken by the Stepwise procedure and the variables introduced at each step.

The key part of the Model Summary is the Adjusted R-square and SEE at each of the seven steps. Note how the R-square increases and the SEE decreases as variables are added, and then levels off near the end. If further variables had been added beyond the seventh step, they would have little or no improvement in these two statistics.

The Excluded Variables table shows which variables were omitted at each step, with the final seventh step excluding NorthView, SouthView, and Stairwell, all of which had low *t*-statistics in our earlier regression.

The adjusted R^2 is 0.810 and the SEE is 6,487.8, which means the COV is 8.47% (6,487.8 ÷ 76,593.5). The F statistic is 73.7 with a Significance of 0.000. This is a better result than our earlier one, indicating that adding NorthView, SouthView, and Stairwell did cause some clouding of the model.

All of the VIF (and tolerance) statistics look fine – no variables approach the critical values of VIF $>$ 3.333 or tolerance $<$ 0.3, showing no evidence of multicollinearity in the model.

From these results, it appears we have found our "best" model, using seven variables to estimate Sale Price: Topflr, Floor#, Total_Area, HalfBath, Parkug, EastView, and Elevator.

As all desired or required variables have now been entered, this completes the variable selection process.

STEP 7: Model Calibration

Now that we have found a final list of variables, the model is ready to be calibrated; or, in other words, we can now determine the coefficients for each of the variables to be included in the regression model. The Stepwise Regression module displays the model coefficients in its final step. However, we will run an Enter method too, because it produces more compact reports with only one step, rather than a separate report for each step for included variables.

We will now calibrate the regression model using the variables selected above. To run another model using the Enter method, proceed as follows:

- Select Analyze → Regression → Linear
- Enter SalePrice as the Dependent variable and Topflr, Floor#, Total_Area, HalfBath, Parkug, EastView, and Elevator as Independent variables.
- Set the Method to Enter.
- Click the Statistics button, and select Estimates, Model Fit, Descriptives, Collinearity Diagnostics and Casewise Diagnostics (use 3 standard deviations). Click Continue.
- Click Plots, and check Histogram and Normal Probability Plot. Click Continue.
- Click the Options button and select Use Probability of F. Enter .05 for Entry and .10 for Removal.
- Click Save, and check Unstandardized for Predicted Values and Unstandardized for Residuals to save the predicted values and residuals from the model into new variables (PRE_1 and RES_1). Click Continue.
- Click Paste. This pastes the procedure into syntax file (note: you must press Paste before running the regression). Now run the regression from the syntax file.

OR

- Click OK to run the regression.

If you used Paste to save the regression to the syntax file, it should appear as below:

```
REGRESSION
 /DESCRIPTIVES MEAN STDDEV CORR SIG N
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT SalePrice
 /METHOD=ENTER Topflr Floor# Total_Area Halfbath Parkug EastView Elevator
 /RESIDUALS HIST(ZRESID) NORM(ZRESID)
 /CASEWISE PLOT(ZRESID) OUTLIERS(3)
 /SAVE PRED RESID.
```

You can highlight this procedure and run it (or click OK in the regression module). Your SPSS output should include the following:

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.906(a)	.822	.810	6487.823

a Predictors: (Constant), Elevator, Topflr, EastView, Halfbath, Floor#, Parkug, Total_Area
b Dependent Variable: SalePrice

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	21716200067.155	7	3102314295.308	73.703	.000(a)
	Residual	4714286262.846	112	42091841.633		
	Total	26430486330.001	119			

a Predictors: (Constant), Elevator, Located on the Top Floor, EastView, Number of Half Baths, Location of the unit - Floor, Number of Underground Parking Stalls, Total Living Area - square feet
b Dependent Variable: Sale Price

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	B	Std. Error
1	(Constant)	22,682.446	4,290.462		5.287	0.000		
	Located on the Top Floor	-5,002.567	1,956.237	-0.105	-2.557	0.012	0.949	1.054
	Location of the unit - Floor	831.171	146.924	0.273	5.657	0.000	0.684	1.462
	Total Living Area - square feet	59.307	5.032	0.633	11.786	0.000	0.552	1.813
	Number of Half Baths	7,843.257	2,407.451	0.175	3.258	0.001	0.553	1.807
	Number of Underground Parking Stalls	3,123.350	1,545.230	0.103	2.021	0.046	0.617	1.622
	EastView	-2,357.608	1,394.010	-0.071	-1.691	0.094	0.905	1.105
	Elevator	-7,340.456	2,322.636	-0.143	-3.160	0.002	0.781	1.281

a Dependent Variable: Sale Price

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	52485.34	106909.60	76593.50	13508.852	120
Residual	-18506.340	16083.813	.000	6294.112	120
Std. Predicted Value	-1.785	2.244	.000	1.000	120
Std. Residual	-2.852	2.479	.000	.970	120

a Dependent Variable: Sale Price

When checking regression output, the following points are important:

- the coefficients should all have the expected sign (positive or negative) and appear reasonable in magnitude;
- the *t*-statistics should be significant, i.e., absolute value should be greater than 1.64 (significance level less than .10);
- the *F*-statistic should be "large" and the Significance 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.

In general, the coefficients look reasonable.

- Each square foot of living area adds $59.31, while each floor adds $831.17.
- An East view is a negative influence of $2,358 while an extra bathroom is worth over $7,800.
- Underground parking is worth over $3,100 and proximity to an elevator is a negative influence of $7,340.
- The only unusual coefficient is that being on the top floor of the complex is a negative influence of just over $5,000 – perhaps this might be due to long waits for the elevator or long treks up and down the stairs when the elevator is not in service.

All of the *t*-statistics are significant and the F-statistic is large with a significance of 0.000. The adjusted R^2 is 0.810 and the SEE is 6,487.5, which means the COV is 8.47% (6,487.5 ÷ 76,593.5).

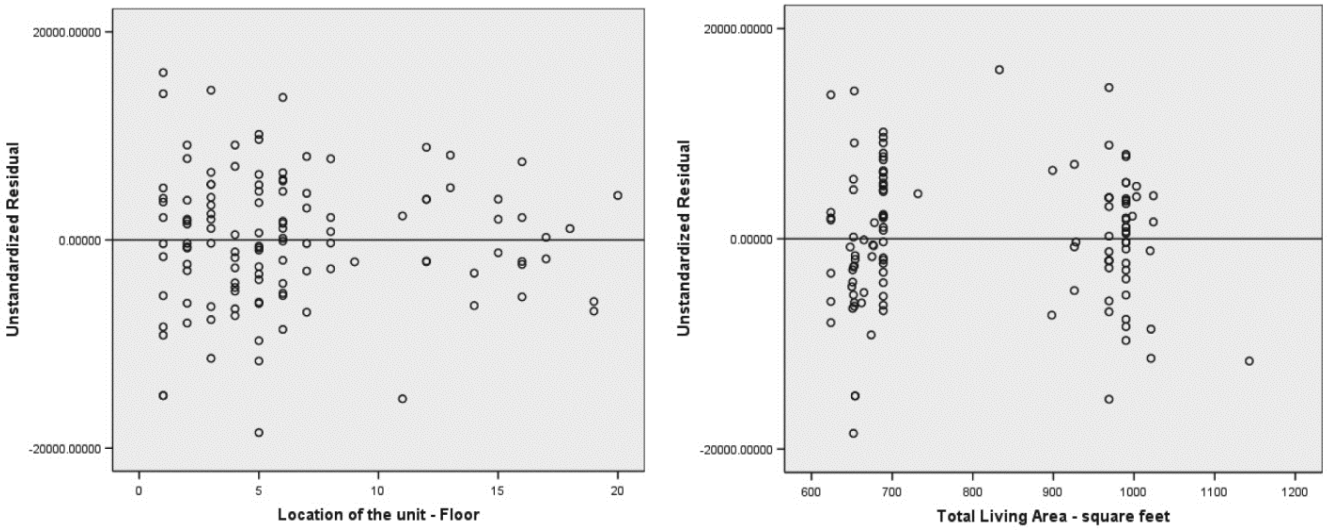
The Casewise Diagnostics report was not included in the output as there were no sales with a residual (error) of more than three standard errors from the mean. The histogram confirms there are none. As well, the histogram of the residuals should appear as a normal bell-shaped curve, and it does. This means the residuals are approximately normally distributed, which is one of the assumptions of regression analysis. The normal P-P plot examines the same distribution in another way: the closer the plot is to a straight line, the more normal the distribution.

The predicted values and residuals were added to the end of the data file when the model was run, with the default names PRE_1 and RES_1. The predicted values were calculated using our new regression equation:

Estimated Selling Price = 22,682.45	−	5,002.57	×	TopFloor
	+	831.17	×	Floor#
	+	59.31	×	Total Living Area
	+	7,843.26	×	Half Bathroom
	+	3,123.35	×	Underground Parking
	−	2,357.61	×	East View
	−	7,340.46	×	Elevator

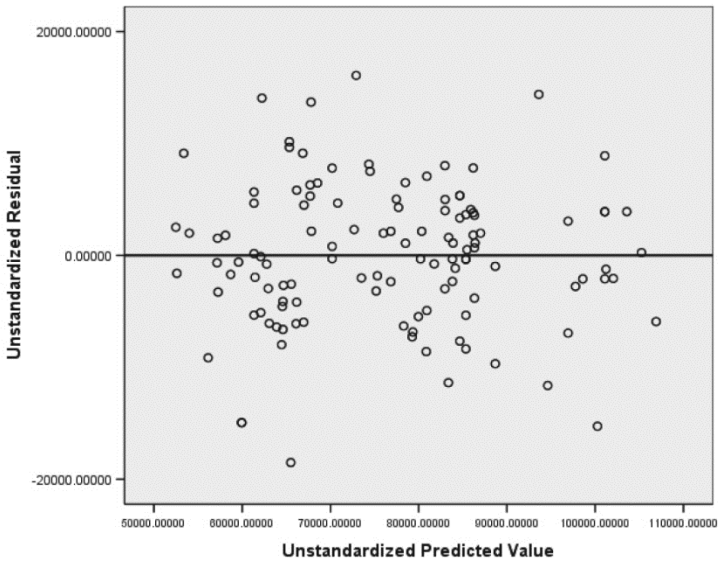
You want to ensure that the residuals (or the errors in predicted values) are not related to any specific value characteristics. If they were, then this would indicate a systematic under- or over-valuation that may require adjusting the model. For example, if you discover that all properties on the top floor are consistently under-valued by the model, then this might warrant a separate adjustment being made.

We will plot the residuals against several key property characteristics to ensure that the plots show a horizontal distribution randomly centred on 0. Run a scatterplot of RES_1 against property characteristics, with a horizontal reference line at 0. We will illustrate this with Floor# and Total_Area.



These plots show there is no pattern to the residuals against property characteristics, as all appear randomly centred on 0. If they are re-run with a regression line and R-square, these show no significant correlation.

We will review a plot of the residuals against predicted values to confirm that these two variables are unrelated. Run a scatterplot of RES_1 with PRE_1 and fit a horizontal reference line at 0 (in the Chart Editor, click Options → Y Axis Reference Line → Enter 0 in the Position field → Apply → Close → Close the Chart Editor). The plot should look as follows:



The residuals appear randomly distributed about 0, as desired.

STEP 8: Test and Evaluate the Model

We are happy with our model and now must test it, calculating value estimates from the model and comparing them against the real sale prices, to see how well the model did.

To test the model's effectiveness in estimating selling prices, we will create a ratio of predicted price to actual price ratio (PAR) by dividing the model's predicted values by the actual selling prices. We can then examine statistics for this ratio. For example, if we find the mean or median PAR is greater than 1.000, then we can say

that generally the predicted values are above actual sale prices. If the mean or median PAR is less than 1.000, then the predictions have under-valued the properties in comparison to their actual sale prices.

Before we create the predicted to actual price ratio, we will first rename PRE_1 to Predval (short for predicted values). Go to the Variable View tab and type Predval over PRE_1 in the Name column. Then, run the following transformation:

```
COMPUTE PAR    =    Predval/SalePrice.
```

If we have done a good job of predicting the selling price of the houses in our sample database, the mean and median PAR should be close to 1.000 – in other words, predicted values are equal to sale prices. If we find the PAR is not close to 1.000, then we may be able to make an adjustment to compensate: e.g., if PAR is 1.200, this means predicted values are approximately 20% higher than sale prices, and we should adjust predicted values down by 20%. Alternatively, this may tell us we need to go back to modeling and try to estimate a better model.

First, we will examine various descriptive statistics for PAR. To more precisely quantify the distribution of the ratios, we will use Frequencies in Descriptive Statistics to calculate quartiles and cut points for 10 equal groups:

- Analyze → Descriptive Statistics → Frequencies…
- Select PAR as the variable, ensure Display frequency tables is NOT selected.
- Click the Statistics button. Check Mean, Median, Standard Deviation, Minimum, Maximum, Quartiles and Cut Points (10).
- Continue → OK.

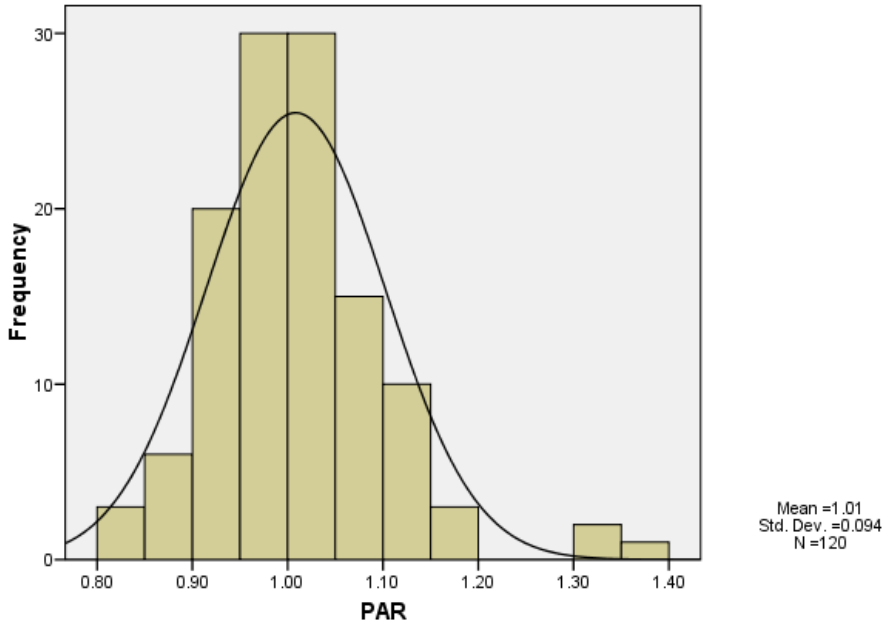
PAR

N	Valid	120
	Missing	0
Mean		1.0083
Median		1.0028
Std. Deviation		0.09396
Minimum		0.82
Maximum		1.39
Percentiles	10	0.9120
	20	0.9383
	25	0.9541
	30	0.9622
	40	0.9758
	50	1.0028
	60	1.0179
	70	1.0397
	75	1.0601
	80	1.0754
	90	1.1111

The model is performing reasonably well, with a median PAR of 1.003 and a mean PAR of 1.008, both very close to the *target* of 1.000, meaning predicted values are generally equal to sale prices. This shows our prediction is quite good, with predicted results very close to actual sales prices (within 0.3% to 0.8%).

The range of PARs is 0.82 to 1.39. Based on the first quartile (25th percentile) and third quartile (75th percentile), 50% of the observations have a PAR between 0.954 and 1.060. In addition, 80% of the observations fall between 0.912 and 1.111. These measures indicate that the PARs are quite tightly bunched in the centre. A tight distribution will give less dispersion and greater uniformity, whereas a wide distribution indicates more dispersion and less uniformity.

We will visually confirm these results with a histogram. Re-open the Frequencies window and click Charts. Under Chart Type, select Histogram and Normal Curve:



The chart shows a nice even distribution, but it does show larger than normal values near the 1.30 to 1.40 mark. When examining ratio results for outliers, you should take a closer look at any results more than two standard deviations away from the mean and definitely those more than three standard deviations.

Probability theory says that if the PARs are normally distributed, 68% of them will be within one standard deviation, 95% within 2 standard deviations, and 99% within 3 standard deviations (see Figure 2.1 in Lesson 2). Given a mean PAR of 100.8%, the standard deviation of 9.4% would provide the following intervals if the data was normally distributed:

68% between [1.008 − 0.094] and [1.008 + 0.094]
0.914 and 1.102

95% between [1.008 − 2×(0.094)] and [1.008 + 2×(0.094)]
0.820 and 1.196

99% between [1.008 − 3×(0.094)] and [1.008 + 3×(0.094)]
0.726 and 1.290

There are three properties with PARs greater than 130. The model seems to have over-valued all of these, with predicted values around $60,000 and actual sales prices around $45,000. It is possible these are data entry errors or perhaps have a condition of sale that caused their sale price to be lower than normal (e.g., not an arm's-length sale, related parties, atypical motivation). An appraiser may wish to fully investigate the conditions surrounding the sale and verify the prices. They may be simply an anomaly. One course of action at this point would be to eliminate the sales from the database and rebuild the model.[5]

We will also examine the PARs for various property characteristics. For example, if we examine PAR by direction of view, we can confirm the PARs are equally distributed. If we find that some view directions are over- or under-predicted, then we can consider adjustments for properties affected. If this model was being

[5] If these outliers were removed, the Adjusted R-square improves to .828 from .810, and the SEE improves to 5,899 from 6,487.

created for property tax assessment, this would be an important adjustment to preserve equity among condominiums throughout the jurisdiction – the principle being that property owners with one view direction should not be under-valued, while those facing other directions are over-valued.

We will use the Ratio Statistics module to test the model by property characteristics:

- Analyze → Descriptive Statistics → Ratio…
- Select Predval as the Numerator, SalePrice as the Denominator, and Directio as the Group Variable.
- Click Statistics. Under Central Tendency, select Median, Mean, and Confidence Intervals (95%). Under Dispersion, select COD, Range, Minimum, and Maximum.
- Continue → OK.

The table below has been altered to achieve a compact width. To replicate this chart, open the Chart Editor, select Pivot → Transpose Rows and Columns.

Ratio Statistics for Predval / SalePrice

		Group				
		East	North	South	West	Overall
Mean		1.009	1.002	1.002	1.017	1.008
95% Confidence Interval for Mean	Lower Bound	.971	.961	.976	.983	.991
	Upper Bound	1.048	1.043	1.028	1.051	1.025
Median		.994	.992	1.004	.988	1.003
95% Confidence Interval for Median	Lower Bound	.960	.927	.975	.963	.978
	Upper Bound	1.029	1.076	1.033	1.044	1.014
	Actual Coverage	96.5%	97.7%	97.6%	97.6%	96.5%
Minimum		.854	.819	.816	.832	.816
Maximum		1.332	1.195	1.111	1.394	1.394
Range		.478	.375	.296	.562	.578
Coefficient of Dispersion		.075	.082	.049	.068	.068

The overall results are similar to the Frequencies report. The coefficient of dispersion (COD) is 6.8% overall, which is within the desired 10% limit. The coefficient of variation (COV) and coefficient of dispersion (COD) allow comparisons to be made between models because they express the dispersion as a percentage and are therefore independent of the unit of measure or its mean. The coefficient of variation (COV) is rarely used in appraisal work. The more commonly used statistic to measure dispersion is the coefficient of dispersion (COD). The coefficient of dispersion expresses the average absolute deviation as a percentage of the median, measuring the spread of the values around the median. For this kind of study a COD of less than 15% is a good result; less than 10% is an excellent result.

The report also shows acceptable levels for all view directions. The lowest median PAR is 0.988 and the highest is 1.004. The means range from 1.002 to 1.017. All of the 95% confidence intervals contain 1.000, meaning there is less than 5% probability the mean/median PAR is not equal to 100%, or we can be 95% confidence that predicted values are statistically equal to sale prices across the view directions in this database. We can conclude no adjustments are needed for any of the view directions.

Boxplots could also be used as a visual test, highlighting differences in PAR by direction or other variables. We could also confirm our conclusion with a Kruskal-Wallis or Mann-Whitney test. However, we will leave these more advanced tests to Lesson 8. As well, comprehensive testing requires withholding some sales in order to test the model against a group of sales not used in creating the model. We will carry this out in Lesson 8 as well.

STEP 9: State Conclusions on Model Quality

Based on the calibration statistical results, this model seems to offer a reasonably good prediction of sale prices for condominium units in the south market area of Regina. This conclusion can be verified after we introduce more methods for model testing in the next lesson.

If this were a real-world application of modeling, at this point you would state your conclusions as to the quality of the model, describing in clear, non-technical terms what you did and why. You would provide your opinion as to how well the model achieves its intended results and note any problems, issues to be aware of, or constraints or limitations on its use.

Summary

In this lesson you have:

- described a general additive model for a database of condominium property;
- examined the variables in the database;
- created the required transformations for these variables;
- selected variables for the model using multiple regression analysis;
- created a final model using multiple regression analysis;
- examined the statistics produced by the final model;
- created a predicted to actual price ratio and analyzed its statistics, for the model as a whole and for property characteristics.

In Lessons 6 and 7, we have explained how regression can be used to create mass appraisal models and illustrated several examples, with each building slightly in complexity. In Lesson 3, we will illustrate one further model building example, this time adding the preliminary data screening and more comprehensive testing.

Review and Discussion Questions

1. Your multiple regression model results show a large F value, but a low R^2 value. What can you conclude about this result?

2. For a mass appraisal model, what is the importance of the variable coefficients? How can you explain these coefficients in valuation terms?

3. To include an ordinal variable for property characteristics (e.g., view) in a regression model, you may transform the variable into separate binary variables. What is a disadvantage of using binary variables versus another re-coding approach?

4. What is the VIF statistic useful for?

5. Assume you need to develop a regression model to explain the impact of view on high-rise condo sales in Burnaby. What type of model would you develop, predictive or explanatory? Why?

6. You want to use a multiple regression model to predict rents for a suburban industrial park. However, you can find only 5 or 6 rent comparables over a 2 month period. What action could you take?

7. Alex has purchased a property sales data-set from a Nova Scotia assessment organization to support his real estate appraisal business. The data includes information on a large number of data variables for residential neighbourhoods. After his initial data exploration, Alex concludes some variables are not very helpful in building a regression model. How could Alex possibly arrive at this conclusion?

8. A real estate analyst is developing a regression model to predict the rent which can be achieved for different types and sizes of office tenancies in Kanata suburban office parks. Two of the variables in her proposed model are square feet of *Rentable Area* (reflects "grossed up" area which includes tenant's share of common area) and office *Useable Area* (actual area occupied by the tenant – usually smaller than rentable area). Can you see a potential problem that the analyst may encounter?

9. Assume you are analyzing resort condo sales in the resort community of Whistler as a part of building a regression model to predict values for 1 and 2 bedroom units in a large strata complex. Most of these condos are included in rental pools for part of the year. Local real estate brokers have told you that the prime characteristics that drive sales are size of unit, number of bedrooms, floor height, view, amenities, ability to "lock-off" units (bedroom in a condo that can be rented separately), strata fees and taxes, and quality of finish. You would like to develop an adjustment for lock-off suites, but you only have 10 sales with this feature, some of which appear to be outliers. The dataset has 90 cases in total. What should you do?

10. In Step 3 of a regression analysis, an appraiser has developed the following correlation analysis for single family dwellings. What can we learn from this table? Note: Condition Rank and Quality Rank have been transformed from ordinal variables showing ranks (e.g., 1 to 10) into linearized variables ranging from 0 to 1.

Pearson Correlations - Single Family Dwelling Analysis

	Fin Area sq ft	Bedrms count	Stories count	Age	Condition Rank	Quality Rank	Lot size sq ft	Sale Price
Fin Area sq ft	1	.895	.674	-.45	.185	.001	.629	.891
Bedrms count	.895	1	.563	-.231	.022	.320	-.070	.921
Stories count	.674	.563	1	.567	.234	.397	-.105	.769
Age	-.45	-.231	.567	1	.764	.830	.392	.562
Cond Rank	.185	.022	.234	.764	1	.932	-.021	.852
Quality Rank	.001	.320	.397	.830	.303	1	.041	.732
Lot Size sq ft	.629	-.070	-.105	.392	-.021	.041	1	.331
Sale Price	.891	.921	.769	.562	.852	.732	.331	1

11. If one of the variables in your predictive regression model had the following statistics, what could you conclude?

    t-value = .095
    sig. = .732
    VIF = 4.107

12. What would happen if you changed the parameters in a step-wise linear regression analysis, as follows:

	Entry Probability	Removal Probability
From	.15	.20
To	.05	.10

    Test the outcomes with the Regina3 model. How do you determine the best thresholds for step-wise regression analysis?

13. In plotting unstandardized predicted dependent values against unstandardized residuals, what type of outcome are we looking for?

ASSIGNMENT 7

LESSON 7: Basics of Model Building

Marks: 1 mark per question.

1. How is model specification different from model calibration?

   (1) Specification involves solving for the model regression coefficients, while calibration mainly involves testing the model.
   (2) The focus of specification is understanding data relationships and selecting variables for modelling, while the focus of calibration is building and testing the model.
   (3) Both names refer to the same process.
   (4) Specification involves testing and refining the model, while calibration involves attaching variables to a model and solving for the coefficients.

2. Your valuation assignment is to develop a regression model to predict property values for August this year. You will use a sample data-set of residential single family property sales in Scarborough, Ontario over a 12 month period ending in October this year. Why might time adjusting these sales be important before model calibration?

   (1) To confirm that no time adjustment is necessary.
   (2) To remove the variance in sale price accounted for by changes in the sale date.
   (3) To ensure that market movement is not accounted for in some other variable.
   (4) Both (2) and (3).

3. You need to understand the contribution of various property characteristics to predicted rent for industrial properties in Richmond, BC. Which type of model would be required?

   (1) Explanatory model, since the goal is to explain the value that each variable contributes to the dependent variable, rent.
   (2) Predictive model, since the goal is to predict the rent outcome for industrial properties.
   (3) Explanatory model, since the goal is to understand and maximize the accuracy of the coefficient values.
   (4) Both (1) and (3) are correct.

Assignment 7 continues on next page

4. You are developing a model to predict new housing starts for Brandon, Manitoba. In step 3 of the model building process you notice that two independent variables, finished floor space and number of floors, have strong correlation. What should you do?

   (1) This is a desirable outcome, no action is required.
   (2) Remove one of the variables from the regression analysis to avoid multicollinearity.
   (3) Add one more independent variable to improve R^2.
   (4) Calibrate the model for a slightly larger acceptable standard error of estimate.

5. If you were reviewing a data-set of office building sales, which of the following potential independent variables is the most likely candidate for transformation?

   (1) Strength of tenant's covenant or credit risk (good, average, poor).
   (2) Rent per square foot.
   (3) Floor level.
   (4) Number of parking stalls per rentable square foot.

6. During data exploration, you found it was necessary to transform two variables, one into a binary variable (waterfront amenity feature) and the second into a unit of comparison (price per front foot). What should you do next?

   (1) Develop the regression model with the new variables.
   (2) Re-test the correlation between the new variables and dependent variable.
   (3) Test for multicollinearity.
   (4) Both (2) and (3).

7. Once a model is developed, it is important to test the results to ensure they are sufficiently accurate. One method is to calculate the ratio of predicted values to actual sale prices. Which of the following is TRUE with respect to evaluating these ratios?

   (1) View descriptive statistics for the ratios, evaluating central tendency and dispersion.
   (2) View a histogram for the ratios to see if the results are normally distributed and to identify outliers.
   (3) View ratios by property characteristics, to confirm the groups are evenly predicted.
   (4) All of the above are true.

Assignment 7 continues on next page

8. Consider the following statements regarding transformation of variables for use in modelling:

   A. A variable called CARPORT, with values of Yes and No, should be converted to a binary variable with values of one (1) for Yes and zero (0) for No.
   B. A variable called QUALITY, with values of Poor, Fair, and Average, can probably be used as is.
   C. A variable called LOTSIZE, with some values in square feet and others in acres, can probably be used as is.
   D. A variable called ATTIC, with attic area in square feet, can probably be used as is.

   Which of the above statements are TRUE?

   (1) B and C only
   (2) A, C, and D only
   (3) A and D only
   (4) All of the statements are true.

9. In addition to the correlation statistic, what additional measures would you consider to determine if a variable should be included in a multiple regression model?

   (1) t-statistic higher than 1.6.
   (2) Drop in R^2 and increase in SEE.
   (3) Sig. greater than 1.
   (4) All of the above.

10. Which of the following statements explains the key difference(s) between explanatory models and predictive models?

    (1) A goal of the predictive model is to obtain as high an R^2 as possible, indicating the variables explain as much variation in the dependent value as possible.
    (2) A goal of the explanatory model is to maximize the accuracy of each coefficient so that they can accurately explain the impact of each variable.
    (3) A goal of the predictive model is to forecast a value for the dependent variable with the lowest possible overall SEE.
    (4) All of the above.

11. If you were building an explanatory regression model using step-wise regression, which of the following tests could you reasonably perform to determine whether to remove or keep a variable?

    (1) Exclude the variable if the *t*-statistic was greater than 2.25.
    (2) Include the variable if the F-Statistic was greater than 4.
    (3) Include the variable if the *t*-statistic was less than 0.9.
    (4) Both (2) and (3).

Assignment 7 continues on next page

12. In the step-wise regression example in Lesson 7, the following independent variables were tested: *Topflr, Floor#, Total_Area, HalfBath, Parkug, NorthView, SouthView, EastView, Elevator,* and *Stairwell*. Which variables would be dropped from the regression if the probability of F was set at .05 for entry and .1 for removal?

(1) All the view variables and the parking variable would be excluded.
(2) The stairwell variable would be excluded.
(3) All the view variables, parking, and stairwell variables would be excluded.
(4) All the view variables, parking, elevator, and stairwell variables would be excluded.

13. Assume you have completed step 7 of the model building process and found all t-statistics significant and a large R^2 value. Should you be satisfied with these results and proceed to step 8, model testing?

(1) No, these outcomes indicate problems with the SEE and issues with variable relationships.
(2) Yes, assuming the coefficients are reasonable and the COV is small.
(3) No, since the high *t*-statistic and R^2 values mean the COV is likely a concern and more analysis is required.
(4) Both (1) and (3) are correct.

14. Consider the following property in Regina's South Market area:

1,000 square feet;
a extra half bathroom;
a view to the South;
on the 12th floor of 15;
next to the elevator; and
one underground parking stall.

Using the final regression equation from Lesson 7, what is this property's estimated selling price? (rounded to the nearest $100)

(1) $95,600
(2) $100,602
(3) $102,900
(4) $88,300

Assignment 7 continues on next page

15. Consider the following property in Regina's South Market area:

    700 square feet;
    a extra half bathroom;
    a view to the West;
    on the 1st floor of 10;
    not beside an elevator; and
    surface parking.

    Using the final regression equation from Lesson 7, what is this property's estimated selling price? (rounded to the nearest $100)

    (1) $75,000
    (2) $74,400
    (3) $72,900
    (4) $73,800

16. In your role as a resort marketing team member, you need to predict the selling price of ski resort condos. You plan to build a regression equation based on similar recent sales at the resort. Others on your team have suggested what you believe to be an excessive number of variables for the model. What strategy might you use to demonstrate the regression analysis can be simplified?

    (1) Refer the team members to the correlation matrix and highlight the variables which have little correlation with the dependent variable, sale price.
    (2) Develop scatter-plots of each dependent and independent variable to illustrate relationships and strength of the R^2 value
    (3) Use step-wise regression to demonstrate the impact of adding each new variable.
    (4) All of the above.

17. You have learned that the Regina3 database has an error. The values for total living area are consistently high by 40 square feet since the balcony area (limited common area) has been mistakenly included in condo total area. How does this error change the regression coefficients for the updated equation? *Hint:* transform living area into a new variable without this 40 square feet (for properties with balconies), and then re-run the regression: Newarea = Total_Area – (40×Balc#).

    (1) *Half Baths* coefficient is increased by 50%.
    (2) *Constant* is increased by approximately $3,000.
    (3) There were no changes in any aspect of the updated equation since total area differences were too small.
    (4) The living area coefficient increased slightly, by less than $1.

Assignment 7 continues on next page

18. After a subsequent field investigation you have learned the Regina3 database may have an error in the number of half baths for each suite. Rather than re-inventory all the sales, you decide to simplify the regression equation and eliminate the bath variable (note: you should continue using the Total_Area, not the revised area calculated in the question above). For this revised regression, what would happen to the estimated market value of a condo with the following characteristics? (Round to nearest $100)

800 square feet;
a extra half bathroom;
a view to the West;
on the 3rd floor of 11;
not beside an elevator; and
surface parking

(1) The updated model decreases value by approximately $8,000
(2) The updated model increases value by approximately $5,000
(3) No significant difference in value noted (less than $1,000)
(4) The updated model decreases value by approximately $9,000

19. What is the COV (mean-centred) for the PAR in the final model in the Regina3 database?

(1) 10.7%
(2) 6.8%
(3) 9.3%
(4) Cannot be calculated

20. Examine boxplots of PAR by a variety of property characteristics. Which of the following is TRUE?

(1) There appears to be some variation in PAR by number of surface parking stalls. This may require an adjustment.
(2) There appears to be some variation in PAR by month of sale. This may indicate the need for a time adjustment.
(3) The variation in PAR on floors 17 and above clearly indicates an adjustment is needed.
(4) There appears to be significant variation in PAR by number of bedrooms. This may require an adjustment.

___
20 Marks

Planning Ahead

Project 2 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 will note that Project 2 follows the steps in Lesson 8 quite closely, but you can certainly experiment with the Project's data now using the techniques covered in this Lesson, so that you can get a sense of what to expect. While reading Lesson 8, keep Project 2's requirements in mind, and consider how these steps may be applied in completing this project.

End of Assignment 7