SAUDER
School of Business

Real Estate Division

LESSON 8

Comprehensive Model Building – Data Screening and Testing

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: UBC Real Estate Division.
   Lesson 8: Comprehensive Model Building – Data Screening and Testing

Recommended Reading

1. UBC Real Estate Division. 2009. *Advanced Computer-Assisted Mass Appraisal*. Vancouver: UBC Real Estate Division.
   Chapter 11: Sales Analysis And Mass Appraisal Performance Evaluation
   Chapter 12: Statistical Procedures and Performance Evaluation I
   Chapter 13: Statistical Procedures and Performance Evaluation II
   These chapters explain valuation model testing in detail, for students wishing more background.

2. UBC Real Estate Division. 2003. "Time Adjustment Illustration". Vancouver: UBC Real Estate Division.

Learning Objectives

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

1. optimize raw data for modeling use by eliminating data not of interest, identifying and eliminating duplicate cases, and ensuring all data is well-understood and documented;

2. critically evaluate outliers in deciding whether or not to exclude them from the model, considering both the pros and cons of this decision;

3. examine the variables in the database for model usefulness, using descriptive statistics, scatterplots, boxplots, and correlations;

4. transform variables for use in the model, including linearizing, recoding binary variables, mathematical transformations, and string transformations (converting descriptive words to numbers that can be used in a model);

5. review variables for inclusion in an additive multiple regression model, excluding inappropriate variables on the basis of number of observations, variable type, relationship to the dependent variable, and multicollinearity;

6. specify a final group of variables for inclusion in a model, using stepwise regression;

7. separate a database into model and test components and explain the reasons for doing so;

8. calibrate an additive multiple regression model, analyzing the variable coefficients for reasonableness;

9. create a ratio variable for testing the performance of the model and apply statistical tools to examine the model quality;

10. test the model's uniformity in valuing property characteristics and make adjustments as necessary; and

11. formulate a conclusion on the appropriateness of the model based on testing outcomes..

Instructor's Comments

In this lesson, we will build on the material presented in Lessons 6 and 7. Lesson 6 introduced the use of regression in model building and Lesson 7 furthered this by showing examples of how these principles are applied. This lesson will complete our coverage of model building by illustrating a comprehensive model building application.

We will work with a database named "Ontario", which can be downloaded from the "Online Readings" section of the Course Resources webpage. This data was provided by the Municipal Property Assessment Corporation (MPAC), Ontario's property assessment agency. We will use this data to build a price predictive model for single family detached dwellings in a residential area of Southern Ontario.

Our approach will be similar to that in Lesson 7, although we will add a preliminary step for data screening to the start and more advanced testing at the end.

Preliminary Data Screening

The first step in any data analysis project, once the data is received, is to do a quick preliminary examination of the data to get a sense of what you have to work with. You will need to determine:

- how much data is in the file – number of records and number of variables; and
- if the data is what you want and need for your project.

A quick scan of the Ontario file shows that there are 589 records and 62 variables. Looking at the Missing column in the Variable View tab shows there are no missing values for any of the records.

In the Variable View tab, you will also find that many of the variables have a table of possible codes in the Values column. For example, for the propcode variable, click on the Values cell, and you will see a small button with … displayed in the cell. Click on the … button and you will see a list of 32 possible values for propcode. The propcode 301, "Single-Family Detached" is of most interest to us, since we want to build a model for single family detached dwellings only.

We have found our first item to examine: how many records are there for each propcode? More specifically, how many records are there with a propcode of 301 in the file? Use Analyze → Descriptive Statistics → Frequencies to examine the propcode variable.

Frequency Table for propcode variable

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Single-family detached	525	89.1	89.1	89.1
	Semi-detached residential use	43	7.3	7.3	96.4
	Duplex	11	1.9	1.9	98.3
	333	5	0.8	0.8	99.2
	335	1	0.2	0.2	99.3
	336	4	0.7	0.7	100
	Total	589	100	100	

We see from the table that there are 525 records of interest to us. So the next step is to eliminate the records we do not want; or, put another way, to keep the ones we do want, those with propcode 301.
To pare down this file, use the following steps:

- Data → Select Cases → If condition is satisfied → If…
- In the condition box enter: propcode = 301 → Continue.
- Click Delete unselected cases → OK.
- This should leave you with 525 records.
- File → Save As (and provide a new name for the file with 525 SFD records; we will call it "Ontario525").

Now that we have pared the file down to only the SFDs, we should see if there are any properties with more than one sale in the file. These are called *repeat sales* and can be useful for analyzing time trends. However, if left in for modeling they can introduce errors into a multiple regression analysis (MRA) because the model will try to predict the value of the same property based on two different records. In this case, we will examine the data for these duplicates and only keep the most recent sale.

To find and eliminate the earlier duplicates use the following steps:

- Data → Identify Duplicate Cases.
- Put Property Number Identifier (prop_num) in the Define matching cases by box.
- Put Sale Date in the Sort within matching groups by box, select Sort Ascending.
- In the Variables to Create area, click Indicator of primary cases, and click "Last case in each group is primary" (a new variable will be created called PrimaryLast).
- Ensure Move matching cases to the top of the file is selected.
- Ensure Display frequencies for created variables is not selected.
- OK.

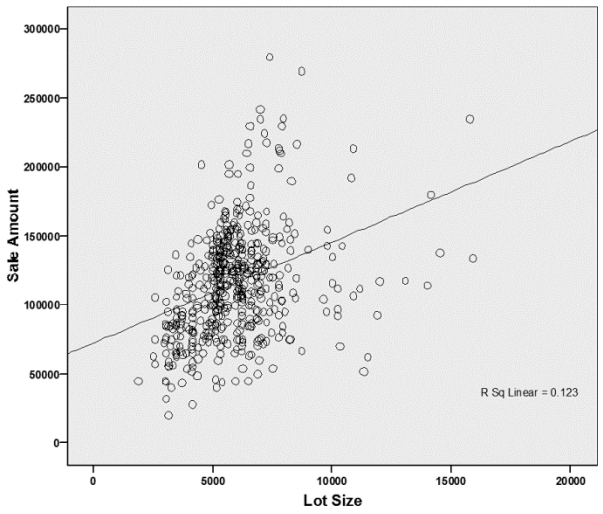
If you scroll to the top of the Data View window, you will now find the first eight records are four sets of duplicates. You may scroll to the right to confirm that the new PrimaryLast variable is a 1 for the most recent sale and a zero for the older sale. We will use the same steps as above for the propcode variable to eliminate these four sales based on the PrimaryLast variable.

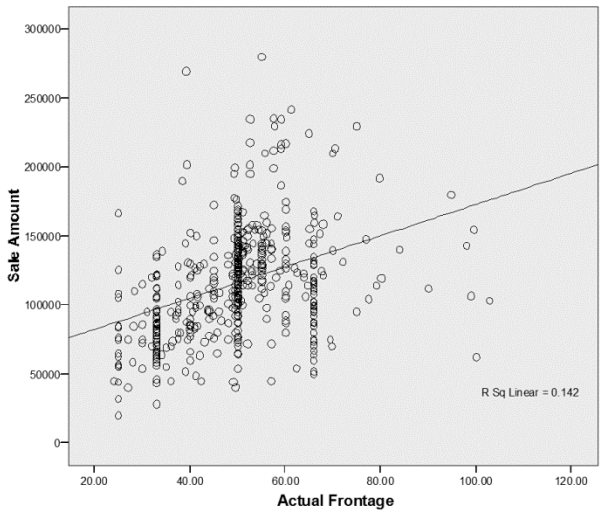
- Data → Select Cases → If condition is satisfied → If…
- In the condition box enter: PrimaryLast = 1 → Continue.
- Click Delete unselected cases → OK.
- This should leave you with 521 records – confirm this in the Data View window.
- Scroll to the right and right click on the PrimaryLast variable name → Edit → Clear (as we won't need this variable again).
- File → Save As (and provide a new name for the file with 521 unique SFD records; we will use "Ontario521").

Finally we should check for records that seem to have outliers in any given variable, that is, any values that are odd in relationship to the other variables. It is ideal to identify these records early and eliminate them before any model building takes place. To do this we will look at several scatterplots of variables against sale price. All we are looking for is odd values. In this case, for land we will look at site area, frontage, and depth, while for building we will look at total floor area, ground floor area, second floor area, and finished basement area. We will delete any records that seem to stand out.

You should note that eliminating records should be done with care. Ideally, you should confirm data problems with the source of the data (be it a document or a person). You want your data to be as "clean" as possible, but on the other hand you want to be careful not to eliminate important variation unnecessarily – over-managing the data is a problem too.

Use Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → Define. Set sale_amt as the Y-axis and a variety of variables for the X-axis: e.g., lot size, frontage, depth, total living area, ground floor area, second floor area, or total finished basement area.





No anomalies are readily apparent from these scatterplots.

We now have our data to begin the modelling process. We will follow the nine steps outlined in the previous lessons. However, we will abbreviate the explanation in some of the earlier steps and instead spend more time on the testing phase of MRA modelling. Once again, the nine steps are:

1. Describe an appropriate general model to use and state this model using standard mathematical symbols.

2. Review the variables in the database and identify those which are suitable to use as independent variables for the type of model defined in Step 1.

3. Examine the potential independent variables looking for relationships with each other and with the dependent variable using graphical analysis, cross tabs, and correlation analysis.

4. (a) Create any transformations necessary to make variables suitable for the chosen model structure.
   (b) Create any additional transformations required to remove problems of collinearity identified in Step 3.

5. Repeat Step 3 with new variables.

6. List a final group of potential variables for calibration.

7. Calibrate the model using an appropriate method.

8. Test and evaluate the model.

9. State your conclusions as to model quality.

STEP 1: Specifying the Model

We will develop and test an additive model to estimate the value of single family detached residential property for a county in Southern Ontario based on the variables given in the database. The additive general model that is often applied to residential property is:

$$MV = LV + BV$$

where

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

Land value is determined by a number of items including land size and location characteristics, such as view, proximity to schools, traffic noise, and neighbourhood. Building value is determined by items related to the physical dwelling on the property such as square footage of living area, number of bedrooms, number of bathrooms, and quality of construction.

Given the list of variables available in our database we can be somewhat more specific and produce a general model for the building value (and hence market value):

$$MV = b_0 + \sum(b_i \times STRUCTURE\_VARIABLE_i) + \sum(b_j \times LAND\_VARIABLE_j)$$

where

MV = estimated selling price (or market value) of the property;
b_0 = constant;
b_i, b_j = coefficients of the independent variables;
$STRUCTURE\_VARIABLE_i$ = any variable associated with the building or buildings on the land; and,
$LAND\_VARIABLE_j$ = any variable associated with the land or location of the property.

STEP 2: Reviewing the Variables

We will now list the variables and the category each falls under:

FACTOR	VARIABLES
Structure	strpcode, quality, condtion, yr_blt, renoyear, renotype, yrblteff, sty_full, sty_part, splilvl, area_tot, area1, area2, areau, bsmtarea, bsmttype, bsmtfin, bsmt_ht, fireplcs, bathfull, bathhalf, bedrooms, heattype, aircond, porchtyp, porchpts, shedarea, garatta, gardeta, garcpta
Land	hnbnd, lotsize, lotsz_um, frontage, depth, access, sercd_hy, sercd_sa, sercd_wa, ab_ind, ab_comm, ab_inst, ab_multi, ab_railw, ab_htraf, ab_mtraf, ab_ltraf, ab_playg, ab_walkw, ab_green, sitefair, sitegood, wfront_r, culdesac, corner

Information only	propnum, address, propcode (all are now 301), saledate (will use to determine if time adjusting the sale prices is necessary), sale_amt, salecode (note that all sales now have the same 005 sale code for single detached dwellings), random

In the next step, we will closely examine the variables to select which of those listed above will be included in the modelling process. If you have not already done so in the preliminary analysis of the database, you should familiarize yourself with all of its property characteristics now – before we begin to make changes to the variables.

STEP 3: Examining the Variables

In this section we will use the tools available to us to examine the variables in our database, determining their relationship to sale price and to each other. In this process, we will learn a great deal about our data and likely come up with a number of questions concerning it. For answers to many of these questions we will either have to rely on our local market knowledge or go back to the source of the data to get this local knowledge.

The statistical tools we will use for examining variables and the relationships have been shown in previous lessons. These include descriptive statistics (frequency distributions and crosstabulation tables), charts or graphs (scatterplots and boxplots), and correlation coefficients and simple linear regression.

Step 3(a): Time Adjustment

First, we examine the sale date range in our database. Since the sales dates are provided in a single Year/Month variable we will use a frequency distribution to get the range. Use Analyze → Descriptive Statistics → Frequencies, and select saledate as the Variable.

Frequency Table for Sale Date (YYYYMM) variable

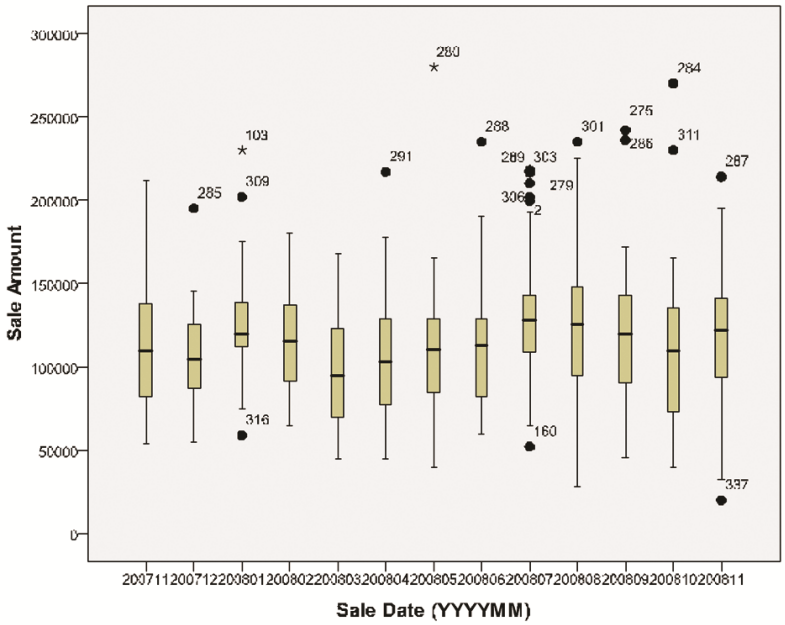
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	200711	38	7.3	7.3	7.3
	200712	33	6.3	6.3	13.6
	200801	21	4	4	17.7
	200802	15	2.9	2.9	20.5
	200803	26	5	5	25.5
	200804	26	5	5	30.5
	200805	48	9.2	9.2	39.7
	200806	52	10	10	49.7
	200807	50	9.6	9.6	59.3
	200808	65	12.5	12.5	71.8
	200809	68	13.1	13.1	84.8
	200810	36	6.9	6.9	91.7
	200811	43	8.3	8.3	100
	Total	521	100	100	

The results indicate that the sale dates range from November 2007 to November 2008, a span of 13 months. Our valuation date is November 30, 2008, which means we will be estimating our market values for this date.

Because of the potential for market movement leading up to our November 30, 2008 target date, we must determine if the sale prices need time adjusting. In a rapidly moving market it is important to ensure that all of your sales have the same baseline. For example, in a market that is moving upwards, you may need to factor the older sales up to align them with the more recent sales. Otherwise, your model will try to account for market

movement in the unrelated land or building variables and the end result will be less accurate variable coefficients. In a falling market, you would want to factor the older sales downward.

We will first examine the sale price and time of sale variables graphically to see the trend in prices. Because the sale date variable is discrete (separate entries for each month), it is best examined by a boxplot:



There does not seem to be much market movement in our sales sample. However, we will confirm this statistically using a Kruskal-Wallis test.

To use the Kruskal-Wallis test, we must create a new month number variable. Because there are only 13 months in total, we will recode the saledate variable into the range 1 to 13. Open a new syntax file and type in the following commands (don't forget the periods at the end of each line):

```
RECODE saledate (200711=1) (200712=2) (200801=3) (200802=4) (200803=5) (200804=6)
(200805=7) (200806=8) (200807=9) (200808=10) (200809=11) (200810=12) (200811=13) INTO
Sale_Month.
VARIABLE LABELS Sale_Month 'Sale_Month'.
EXECUTE .
```

Save the syntax file (e.g., "Lesson 8 syntax") – you will be using this syntax file throughout this lesson.[1] 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.

[1] As an alternative, all transformations can be done using Transform → Compute Variable. However, using syntax files allows you to save the transformations for future reference and also apply them to other databases, which will be an advantage in this lesson.

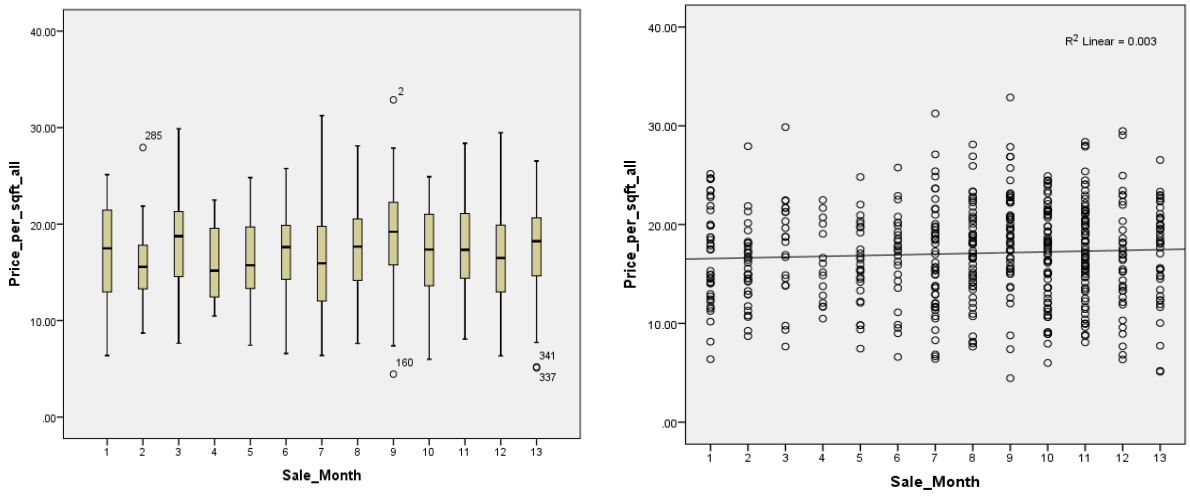
Confirm in the Data View window that a new variable called Sale_Month has been created, with values ranging from 1 to 13. You may also use Frequencies to verify your work.

One more data element is needed to carry out our time adjustment analysis and that is some unit of comparison for all sales in the database. Intuitively we know that total living area and lot size are two primary drivers for the sale price of a house. If we create a variable that is based on those three variables (total living area, lot size, and sale price), we should have a reasonable unit of comparison to test to see if prices have been moving within our study area.

In your syntax file, add the following transformation and run it:

```
COMPUTE Price_per_sqft_all = sale_amt / (area_tot + lotsize).
EXECUTE.
```

We will create a boxplot and scatterplot between Price_per_sqft_all and Sale_Month for a visual test on whether a time adjustment is necessary.



This is similar to our boxplot result earlier – there does not seem to be much market movement in our sales sample. However, a Kruskal-Wallis test will provide statistical evidence to confirm this. This test is non-parametric, meaning the validity of the test does not depend on the normal distribution of the data. The test ranks the variable in question from low to high (1 to 521) and then the rankings for all observations are grouped by month. The mean of the rankings in each month is then calculated. If it is assumed that all months have equal sale prices, there should be a similar distribution of high and low values in each month. This implies that the expected mean rank in each month should be approximately equal to the middle value in the database; in this case there are 521 observations so the expected value of the mean is approximately 260. If the observed mean rank for each month varies significantly from 260, this would indicate that the months are not valued equally and that there is some form of difference in sale price in various months.

- Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples…
- Enter Price_per_sqft_all as the Test Variable List and Sale_Month as the Grouping Variable.
- Click Define Range… and enter 1 for the Minimum and 13 for the Maximum.
- Continue → OK.

Kruskal-Wallis Test for Price per Square Foot of Total Living Area and Lot Size vs Sale Date

	Sale_Month	N	Mean Rank
Price_per_sqft_all	1	38	259.87
	2	33	210.94
	3	21	284.86
	4	15	222.47
	5	26	225.31
	6	26	256.00
	7	48	240.83
	8	52	267.04
	9	50	319.26
	10	65	260.14
	11	68	273.32
	12	36	239.61
	13	43	274.00
	Total	521	

Test Statistics(a,b)

	Price_per_sqft_all
Chi-Square	16.59
df	12
Asymp. Sig.	0.166

a Kruskal Wallis Test
b Grouping Variable: Sale_Month

The first part of this report shows the months, the number of sales in each month, and the mean rank. Scanning down the ranks for each month we see that the minimum is 210, the maximum is 319, and six of the months have mean ranks close to the expected value of 260.

The Kruskal-Wallis test also provides a chi-square test to determine if this apparent difference is actually statistically significant. The chi-square statistic is calculated using the expected value and the actual observed mean values for each month. The Asymp. Sig. column shows the probability of achieving a value of chi-square of this magnitude if the months had equal sale prices per square foot. Here, the calculated chi-square statistic is 16.59 and the probability of obtaining a test statistic of this amount if the months are equally valued is .166 or 16.6%. The threshold for rejecting the hypothesis that the months are equal is 5%. Our result at 16.6% is considerably greater than 5%, meaning we cannot be statistically certain the months are different. Therefore, we can confirm that no time adjustment is necessary. Remember that it is the Asymp. Sig. that determines if the test variable (Price_per_sqft_all) is equally distributed across the grouping variable (Sale_Month).

In markets that are increasing or decreasing, time adjustments are necessary and therefore this is a quite common requirement in modeling. Students interested in knowing more about making time adjustments may wish to review the Time Adjustment Illustration document found under Lesson 8's Online Readings on the course website. In Appendix 8.1 we have illustrated additional examples of Kruskal-Wallis tests. Students often have difficulties with the correct interpretation of these tests, and these examples are provided to add information on the correct use of these important tests.

Step 3(b): Descriptive Statistics

Next, use Analyze → Reports → Case Summaries (do NOT display cases) to produce a statistical summary of the following variables. For statistics, select mean, median, maximum, minimum, range, and standard deviation:

sale_amt	area_tot	bsmtarea
lotsize	area1	bsmtfin
frontage	area2	porchpts
depth	areau	

Case Summaries

	Mean	Median	Minimum	Maximum	Range	Std. Deviation
Sale Amount	115497.33	115000	20000	280000	260000	38754.008
lotsize	5888.74	5650	1854	15900	14046	1857.972
Actual Frontage	49.396	50	24	102.84	78.84	12.90131
Actual Depth	120.3095	112	51.5	318	266.5	26.58573
area_tot	1043.25	1020	370	2394	2024	299.656
Ground Floor Area	908.1	888	370	2000	1630	240.104
Second Floor Area	134.03	0	0	1400	1400	245.233
Third/Upper Floor Area	1.12	0	0	374	374	18.733
Total Basement Area	874.19	874	0	1878	1878	305.618
Total Finished Basement Area	333.8	312	0	1335	1335	310.346
Total Porch Points	9.34	6	0	57	57	9.61

The total floor area ranges from 370 to 2,394 square feet with a mean of 1,043; the sale prices range from $20,000 to $280,000 with a mean of $115,497. Total porch points range from zero to 57 with a median of 6–half of the properties have six or fewer porch points. In checking with the data source, we discovered that porch points are calculated using a combination of type of porch (uncovered, covered, enclosed), size, and quality. Generally one porch point is worth roughly $100 in value, thus the porch on a home with 50 porch points is worth twice that of a porch with only 25 porch points.

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

shedarea
garatta
gardeta
garcpta

We will not produce the output here, as the reports are long. They show how many properties do not have these features (value equal 0) and then the area for each property which has the feature. We find only 37 properties have sheds, ranging in size from 64 to 508 square feet. Only 34 properties have attached garages, many more (179) have detached garages, and there are only 18 with carports.

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

propcode	salecode	ltsz_um
strpcode	quality	condtion
yr_blt	renoyear	renotype
yrblteff	sty_full	sty_part
splilvl	bsmttype	bsmt_ht
fireplcs	bathfull	bathhalf
bedrooms	heattype	aircond
porchtyp	access	sercd_hy
sercd_sa	sercd_wa	ab_ind
ab_comm	ab_inst	ab_multi
ab_railw	ab_htraf	ab_mtraf
ab_ltraf	ab_playg	ab_walkw
ab_green	sitefair	sitegood
wfront_r	culdesac	corner

Again, we will not provide all of these reports, but we will instead summarize a number of things that can be seen from these frequencies:

- all 521 sales are open market sales of single family detached dwellings;
- the land area for all properties is measured in square feet;
- the quality of the house ranges from 4 to 7, with 5 and 6 being the most common;
- the condition of the homes is either Poor, Fair, Average, or Good, and over 90% are average;
- year built ranges from 1903 to 2007;
- 293 properties have not had a renovation recorded;
- there are four type of renovations coded (it appears that for this variable a blank code and a zero mean the same thing);
- the effective year ranges from 1913 to 2007;
- 50 properties have two stories, 93 have part stories;
- there are 16 split level homes in three types and it appears that for this variable a code of N means not a split level;
- there are three types of basement finish, 173 properties have no basement finish;
- basement height, where it exists, ranges from 5.5 to 9.0 feet – the most common is 7.5 feet;
- 97 properties have one fireplace, nine have two;
- six houses have three full bathrooms;
- nine houses have more than four bedrooms;
- 92.5% of homes have forced air heating;
- 60 homes have air conditioning;
- 10% of homes have no porch recorded (blank or zero);
- all properties have road access;
- the data representing the availability of sanitary, water, and hydro services looks odd, as one would expect all properties in this southern Ontario county to have both – undefined likely means not coded;
- for abutting influences, one has industrial, 17 commercial, zero institutional, five multi-residential, three railway, and two for each of sports field/playground, public walkway, and green space;
- traffic influences 40 properties (at three different levels);
- seven properties have a "fair" site, five have a "good" one; and
- six properties are on waterfront, 12 are on a cul-de-sac, and 35 are on a corner.

From these frequency statistics we can easily eliminate a few variables, even at this early stage in our analysis. We know that we would like at least five occurrences of any given variable, and preferably 5%, meaning 26 sales for our database. Based on the fewer than five rule, we can immediately eliminate the following variables: abuts industrial, abuts institutional, abuts railway, abuts sports field/playground, abuts public walkway, and abuts green space. The access variable can be disregarded as all properties are identically coded. The three variables representing availability of sanitary services, water, and hydro are suspect and will also be eliminated (note: local knowledge will tell you that all properties in the sample should be coded "municipal" for these two variables).

Helpful Hint

When variables are "eliminated" from the analysis, this does NOT necessarily mean they get deleted from the database like is usually done with duplicate records. The majority of these "eliminated" variables are simply taken out of further consideration in the model building process. Often you will need these variables for testing purposes once the model has been calibrated. However, a constant variable or one that has questionable or highly suspect data can be physically removed from the database if desired (just be careful and remember to back-up your data regularly!)

Some variables are definitely on the edge for elimination: abuts multi-residential (5), waterfront (6), and the two site quality variables (seven occurrences of Fair, five of Good).

At this stage, we can recognize variables that will be difficult to use:

- basement height – may find a relationship between this variable and basement finish;
- number of storeys – the floor area variables will account for these;
- type of split level – may simply need a binary variable to indicate that the property is a split level, there are not enough of each type to warrant separate binary variables;
- the central heating variable has the same problem – perhaps a variable for unusual heating would be warranted;
- air conditioning will have to be translated from Y for "yes" and N for "no" to 1 and zero respectively (binary);
- the porch type variable may be already accounted for in the porch points variable (in fact, a crosstab between these two variables clearly shows the relationship); and,
- the effective year built variable should be partially based on the renovation variables.

As well, we can also recognize that a few variables cannot be used in their present forms:

- structure condition – we will have to translate Poor, Fair, Average and Good into another variable;
- structure quality – we will have to define the relationship between each of the numbers;
- effective year built will have to be transformed into age;
- finished basement type – we may wish to create a variable that represents the presence of a basement suite or apartment; and,
- the three traffic pattern variables will likely have to be rolled into one variable.

Many of these issues will be visually analyzed in the next section, and any transformations will be covered in Step 4.

Step 3(c): Graphical Analysis

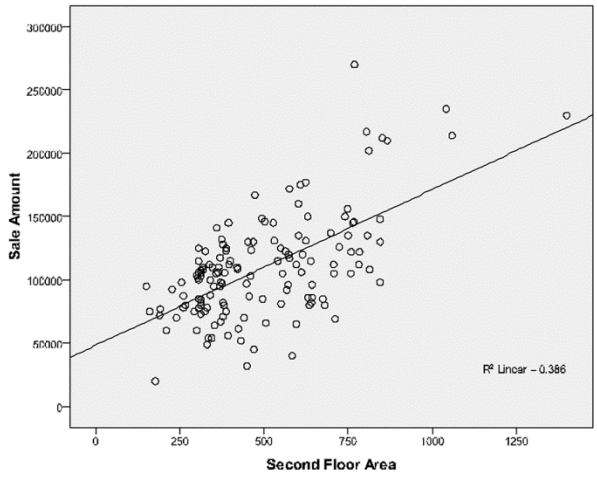
Scatterplots can be used to visually analyze the continuous variables we plan to use. We will plot these variables as independent variables and sale price (sale_amt) as the dependent variable. Earlier, while examining the data for outliers, we produced scatterplots of sale price against site area, frontage, depth, ground floor area, second floor area, and finished basement area. Looking back at those plots we can see a positive relationship between lot size and sale price and a slightly stronger one between frontage and sale price. Depth has very little relationship with sale price. Total living area has a strong relationship with sale price, but ground floor area is even stronger.

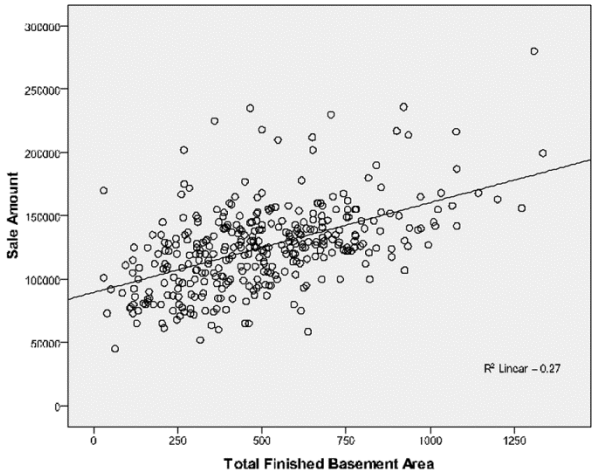
So that we can get a clearer picture of the relationship between sale price and the variables of second floor area and basement finish we will reproduce those scatterplots after filtering out the records with a zero value.

For second floor area, use:

- Data → Select Cases → If condition is satisfied → If…
- In the condition box enter: area2 > 0 → Continue.
- Make sure "Filter out unselected cases" rather than "Delete unselected cases" is selected.
- Click OK.
- Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter → click Define → X-axis is area2, Y-axis is sale_amt for the scatterplot → OK (a fit line has been added in our example).

Use the same procedure for bsmtfin.



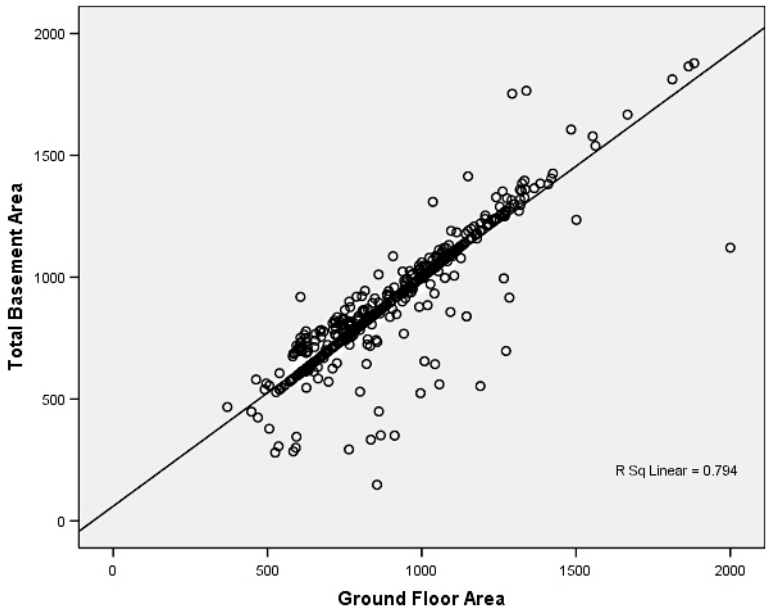


These plots show that both second floor area and basement finish have positive relationships with sale price.

Helpful Hint

After applying a filter it is very important to remember to turn the filter off – otherwise you will continue to use the filtered data in further steps; use Data → Select Cases → All Cases → OK

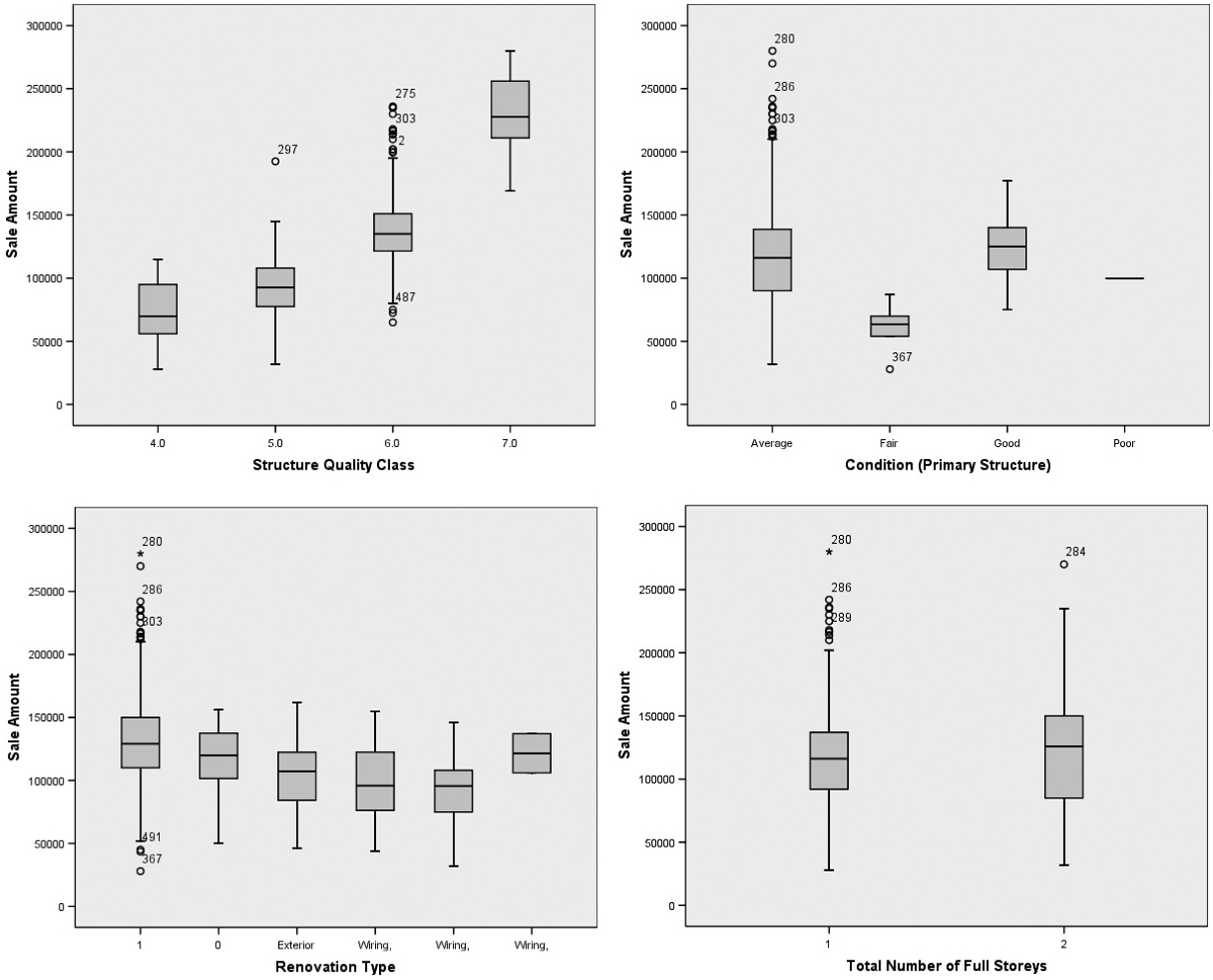
Up to this point we have ignored the basement area variable, focusing instead on the basement finish variable. The main reason for this is that the basement area variable is highly correlated to the first floor area – in many cases (one storey homes with a full basement for example) they will be equal. The following scatterplot illustrates this (Select Cases set to bsmtarea > 0). This highlights that basement finish area is a much better variable to consider in a model as it represents how much extra living space is in the basement of the home.



Now we will use boxplots to examine many of the variables discussed in the previous section. We have already established good relationships between sale price and the continuous variables that represent the land size and building size. Our boxplots will provide a visual representation of the relationship between our many discrete variables and sale price. Where there is little variation between the different values of a discrete variable in its relationship with sale price, it is likely that the variable would not be a good choice for inclusion in the regression model. Thus, we are looking for some separation in the boxes of the boxplots. We will only produce boxplots for those variables with five or more occurrences with the exception of the traffic influence variables.

Use Graphs → Legacy Dialogs → Boxplots → Simple → Define. Sale Amount is the Variable and the following variables should be placed in the Category Axis:

quality	condtion	renotype
sty_full	sty_part	splilvl
bsmttype	bsmt_ht	fireplcs
bathfull	bathhalf	bedrooms
heattype	aircond	porchtyp
ab_comm	ab_multi	ab_htraf
ab_mtraf	ab_ltraf	sitefair
sitegood	wfront_r	culdesac
corner		



We have displayed only a sample of the boxplots here, with the rest left to be run on your own. A number of things can be seen from these boxplots:

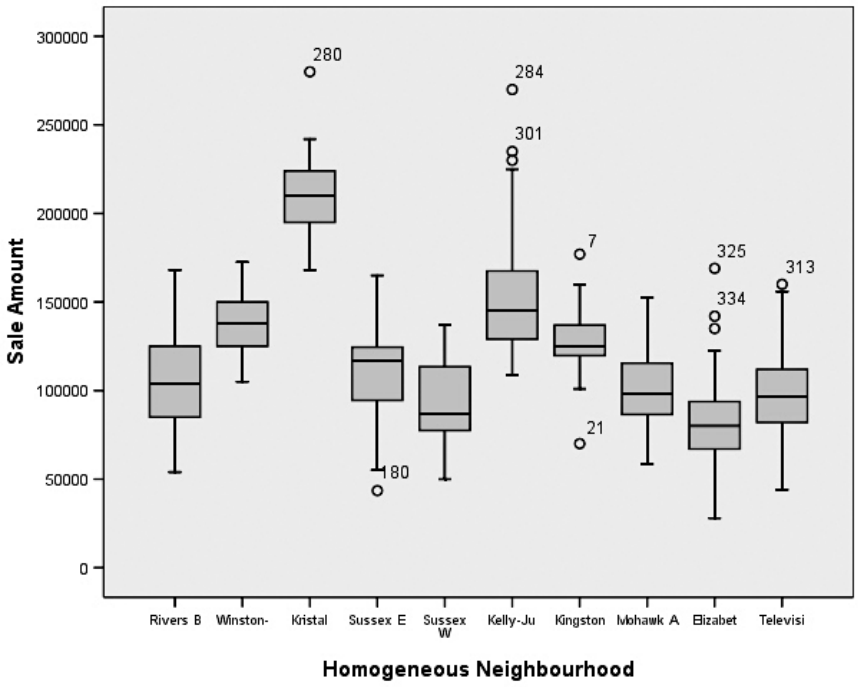
- quality has an impact on sale price;
- a condition of average or good fetches a higher price than fair or poor;
- the type of renovation, or even if the property has been renovated, does not seem to have much of an impact on sale prices in our data – an effective age variable will take any renovations into account;
- both the number of storeys and number of part storeys show very little impact on selling price;
- houses with a split-level design seem to have a higher price than those without;
- the fact that there is some basement finish positively affects the sale price, but there is no significant difference in the type of finish;
- basement height has an obvious impact on sale price;
- fireplaces and bathrooms have an impact on selling prices, as do bedrooms, although only to a point as three, four, and five are at similar levels (there is only one house with six bedrooms, so no conclusion can be drawn);
- the type of heating employed in the houses in our data sample seems to have little impact on selling price;
- air conditioning as well has little impact;
- the existence of a porch is important;
- abutting commercial seems to depress the price somewhat, abutting multi-family (with only five sales) may not really have an impact;

- being close to medium traffic seems not to affect the price at all, but close to heavy traffic (with two sales) or light traffic (with four sales) increases the selling price. It appears that traffic influence can be ignored in our sample;
- quality of site fair (seven sales) and quality of site good (with five sales) both have little or no influence on selling price in our sample; and
- the final three variables, waterfront, cul-de-sac, and corner lot, seem to have very little effect on selling price.

We can now cut out more variables: renotype, sty_full, sty_part, bsmttype, heattype, aircond, porchtyp (we will use porchpts instead), ab_multi, ab_htraf, ab_mtraf, ab_ltraf, sitefair, sitegood, wfront_r, culdesac, corner.

This leaves the following variables for possible inclusion: quality, condtion, splilvl, bsmt_ht, fireplcs, bathfull, bathhalf, bedrooms, and ab_comm. Most will need some transformation (Step 4).

One final boxplot to examine is selling price and neighbourhood. Select Sale Amount as the Variable and hnbhd (homogeneous neighbourhood) as the Category Axis (you may also try Price per Sqft vs Neighbourhood).



The boxplot shows significant differences between the selling prices across the neighbourhoods in our sample.

Step 3(d): Correlation Analysis

We have now determined a number of potential variables for our additive regression model. Many will have to be transformed so that they can be used in such a model. In this section we will use correlation analysis to determine if any variables of interest are correlated with each other. This is really a preliminary examination because another correlation analysis will have to be done once the transformations have been carried out.

Use Analyze → Correlate and run a bivariate correlation matrix with the following variables. Ensure that Pearson is the correlation coefficient type selected. Note that we removed the two extraneous rows from the correlation matrix (N and Sig), using the instructions from Lesson 3.

sale_amt	ab_comm	area_tot
area1	area2	areau
bathfull	bathhalf	bedrooms
bsmt_ht	bsmtarea	bsmtfin
depth	fireplcs	frontage
garatta	garcpta	gardeta
lotsize	porchpts	quality

Correlations

	sale_amt	ab_comm	area_tot	area1	area2	areau	bathfull	bathhalf	bedrooms	bsmt_ht	bsmtarea	bsmtfin	depth	fireplcs	frontage	garatta	garcpta	gardeta	lotsize	porchpts	quality
sale_amt	1	-.131**	.643**	.740**	.063	-.025	.441**	.175**	.408**	.443**	.695**	.504**	.058	.397**	.377**	.387**	.061	.069	.350**	-.180**	.711**
ab_comm	-.131**	1	-.077	-.061	-.033	-.011	.011	-.044	-.108*	-.081	-.074	-.120**	.021	-.042	.006	.050	-.033	-.007	.007	.000	-.066
area_tot	.643**	-.077	1	.596**	.629**	.132**	.355**	.284**	.577**	.222**	.484**	.243**	.054	.359**	.215**	.384**	.043	-.023	.209**	.043	.495**
area1	.740**	-.061	.596**	1	-.247**	-.053	.426**	.109*	.347**	.332**	.794**	.507**	-.006	.337**	.362**	.332**	.079	.006	.294**	-.291**	.681**
area2	.063	-.033	.629**	-.247**	1	.137**	.013	.236**	.359**	-.053	-.183**	-.198**	.070	.108*	-.086*	.146**	-.024	-.035	-.028	.324**	-.066
areau	-.025	-.011	.132**	-.053	.137**	1	.046	.060	.087*	-.018	-.031	-.031	.008	.018	-.076	-.015	-.011	.028	-.058	.169**	.051
bathfull	.441**	.011	.355**	.426**	.013	.046	1	-.231**	.315**	.247**	.401**	.465**	-.036	.274**	.177**	.206**	.072	-.002	.118**	-.112*	.366**
bathhalf	.175**	-.044	.284**	.109*	.236**	.060	-.231**	1	.150**	.097*	.076	.138**	.053	.218**	.023	.194**	.073	.024	.052	.007	.167**
bedrooms	.408**	-.108*	.577**	.347**	.359**	.087*	.315**	.150**	1	.177**	.294**	.321**	-.007	.188**	.157**	.124**	.052	-.011	.125**	.001	.370**
bsmt_ht	.443**	-.081	.222**	.332**	-.053	-.018	.247**	.097*	.177**	1	.708**	.363**	.018	.179**	.203**	.131**	.072	.028	.171**	-.204**	.414**
bsmtarea	.695**	-.074	.484**	.794**	-.183**	-.031	.401**	.076	.294**	.708**	1	.516**	.026	.290**	.333**	.304**	.040	.043	.295**	-.207**	.640**
bsmtfin	.504**	-.120**	.243**	.507**	-.198**	-.031	.465**	.138**	.321**	.363**	.516**	1	-.052	.302**	.228**	.230**	.098*	-.005	.156**	-.222**	.513**
depth	.058	.021	.054	-.006	.070	.008	-.036	.053	-.007	.018	.026	-.052	1	-.033	-.158**	-.008	-.057	.135**	.558**	.044	-.024
fireplcs	.397**	-.042	.359**	.337**	.108*	.018	.274**	.218**	.188**	.179**	.290**	.302**	-.033	1	.126**	.235**	.002	.013	.079	.005	.332**
frontage	.377**	.006	.215**	.362**	-.086*	-.076	.177**	.023	.157**	.203**	.333**	.228**	-.158**	.126**	1	.294**	.068	.088*	.711**	-.152**	.289**
garatta	.387**	.050	.384**	.332**	.146**	-.015	.206**	.194**	.124**	.131**	.304**	.230**	-.008	.235**	.294**	1	-.046	-.147**	.231**	-.005	.284**
garcpta	.061	-.033	.043	.079	-.024	-.011	.072	.073	.052	.072	.040	.098*	-.057	.002	.068	-.046	1	-.059	.011	-.095*	.125**
gardeta	.069	-.007	-.023	.006	-.035	.028	-.002	.024	-.011	.028	.043	-.005	.135**	.013	.088*	-.147**	-.059	1	.171**	-.031	.027
lotsize	.350**	.007	.209**	.294**	-.028	-.058	.118**	.052	.125**	.171**	.295**	.156**	.558**	.079	.711**	.231**	.011	.171**	1	-.101*	.211**
porchpts	-.180**	.000	.043	-.291**	.324**	.169**	-.112*	.007	.001	-.204**	-.207**	-.222**	.044	.005	-.152**	-.005	-.095*	-.031	-.101*	1	-.248**
quality	.711**	-.066	.495**	.681**	-.066	.051	.366**	.167**	.370**	.414**	.640**	.513**	-.024	.332**	.289**	.284**	.125**	.027	.211**	-.248**	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 of interest are those with sale_amt. Although we already have investigated the relationship between sale_amt and these variables, the correlations confirm our findings. The variables areatot, area1, bsmtarea, and quality all have high correlations with selling price. Note that quality has values of 4, 5, 6, and 7 and therefore is not really usable in its present form. If we used it as is, this incorrectly implies a quality 6 house is 1½ times better than a quality 4 house. In Step 4, we will transform the quality variable to create a linear relationship between the quality levels.

As expected areatot is correlated with area1 and area2; and area1 is correlated with bsmtarea and quality. Also, frontage is correlated with lotsize. Bedrooms is somewhat correlated to areatot, which is expected intuitively, since more bedrooms should be associated with larger total living area.

This preliminary correlation analysis leads us to a two immediate conclusions:

- areatot cannot be in a model with area1 and area2; and
- frontage cannot be in a model with lotsize.

Step 3(e): Summary of Variable Examination

The variable examination is an initial step towards selecting variables for the model. We have found variables that can be ignored because they have constant values for all sales (e.g., access), too few occurrences in the data (e.g., ab_ind), or have little impact on the selling price (e.g., corner). The correlation analysis showed variable combinations to avoid in model specification.

Our analysis also uncovered variables that need transformation to be of use in an additive model (e.g., quality and condition). In the next step, we will carry out the necessary transformations and then re-examine the data in the following step.

STEP 4: Transformations

Now that we have a good understanding of the data in our database, the next step is to transform any variables that are not in a usable format for an additive model. We will carry out these transformations, and then examine the new variables and their relationships with each other and to the un-transformed variables.

Transformations are required for the following variables:

- quality: it must be linearized, in order to avoid the model attaching numerical meaning to the 4, 5, 6, 7 coding (e.g., quality 6 is not necessarily 1½ times better than quality 4).
- condition: must be linearized as well; the current Poor, Fair, Average, and Good are unusable in our model.
- yrblteff: we need to create an age variable (this will account for any renovations).
- splilvl: create a binary variable that indicates if the house has a split level design.
- bsmt_ht: basement height seems to have an impact on selling price, but nine foot high ceilings are likely not 1½ times more valuable than six foot high ceilings. We will create three binary variables for under-height basements (5.5', 6', and 6.5'), low height basements (7' and 7.5'), and normal height basements (8' and over).
- bathrooms: we will incorporate full and half bathrooms into one numerical variable.
- neighbourhood: we will create a binary variable for nine neighbourhoods, keeping neighbourhood C10 as a "control" or "reference" neighbourhood (it has the second most sales and from the boxplots appears to be one of the "middle-of-the-road" neighbourhoods).

For the first two proposed transformations, we needed to go back to the source of the data for more information on quality and condition. For example, in appraisal terms, what is the percentage discount or addition for the various measures? After contacting the source of the data we were given the linear relationships for quality and condition:

- for quality, 4 = 0.55, 5 = 0.75, 6 = 1.00, and 7 = 1.10.
- for condition, Poor = 0.65, Fair = 0.75, Average = 1.00, and Good = 1.10.

Thus, a home with a 6 for quality is 1.82 times "better" than a home with a 4 for quality (1.00 ÷ 0.55); a home with a 7 quality is 1.10 times better than a home with a 6 (1.10 ÷ 1.00) – or put another way, a quality 7 home is a 10% improvement over a quality 6 home.

In checking with our data source on quality and condition, we also confirmed that the relationship for porch points is more-or-less linear. In other words, twenty porch points are worth twice as much as ten porch points and four times more than five porch points. As such, the variable porchpts can be used as is [this was discussed briefly in Step 3(b)].

To create the necessary transformations, we will return to our syntax file from earlier, and create and run the following transformations:

```
RECODE quality (4=.55)  (5=.75)  (6=1.00)  (7=1.10) INTO Lin_Qual.
RECODE condtion ('A'=1)  ('F'=.75)  ('G'=1.1)  ('P'=.65)  INTO  Lin_Cond.
COMPUTE Effage = 2009 – yrblteff.
RECODE splilvl ('N'=0)  (ELSE=1)  INTO  split_lvl.
RECODE bsmt_ht (5.5 thru 6.5=1)  (ELSE=0)  INTO  bsmt_under.
RECODE bsmt_ht (7 thru 7.5=1)  (ELSE=0)  INTO  bsmt_low.
RECODE bsmt_ht (8 thru Highest=1)  (ELSE=0)  INTO  bsmt_norm.
COMPUTE bathrooms = bathfull  +  (0.5 * bathhalf).
COMPUTE RiversBend = 0.
COMPUTE WinstonWell = 0.
COMPUTE KristalEstates = 0.
COMPUTE SussexEast = 0.
COMPUTE SussexWest = 0.
COMPUTE KellyJuno = 0.
COMPUTE Kingston = 0.
COMPUTE Mohawk = 0.
COMPUTE ElizabethJune = 0.
IF (hnbhd = 'C01') RiversBend = 1.
IF (hnbhd = 'C02') WinstonWell = 1.
IF (hnbhd = 'C03') KristalEstates = 1.
IF (hnbhd = 'C04') SussexEast = 1.
IF (hnbhd = 'C05') SussexWest = 1.
IF (hnbhd = 'C06') KellyJuno = 1.
IF (hnbhd = 'C07') Kingston = 1.
IF (hnbhd = 'C08') Mohawk = 1.
IF (hnbhd = 'C09') ElizabethJune = 1.
EXECUTE.
```

The transformations are explained as follows:

- the first transformation creates a new variable called Lin_Qual, the linearized quality variable;
- the second transformation creates the linearized condition variable Lin_Cond;
- effective age is created using 2009 as the base year;
- the split level variable is created as a binary variable;
- the three basement height variables are created as binaries;
- the bathroom count variable is created by summing the full and half bathroom variables; and
- nine neighbourhood binary variables are created using the neighbourhood names for the new variable names and leaving neighbourhood C10, Television Area, as the control neighbourhood.

Before continuing, you should check the results of these transformations. As they are fairly straightforward, you may review the data manually or use Analyze → Descriptive Statistics → Crosstabs to ensure that the correct source yields the correct binary. Two crosstab examples are given below, the new split level variable and the new basement height variables. Your output should look similar to the tables below:

		splilvl				Total
		Back or Front Split	Side Split	Yes, Unconventional Split	N	
split_lvl	0	0	0	0	505	505
	1	3	11	2	0	16
Total		3	11	2	505	521

		bsmt_low		
		0	1	Total
bsmt_ht	0	22	0	22
	5.5	3	0	3
	6	6	0	6
	6.5	40	0	40
	7	0	99	99
	7.5	0	201	201
	8	136	0	136
	8.5	9	0	9
	9	5	0	5
Total		221	300	521

		bsmt_norm		
		00	1.00	Total
bsmt_ht	0	22	0	22
	5.5	3	0	3
	6	6	0	6
	6.5	40	0	40
	7	99	0	99
	7.5	201	0	201
	8	0	136	136
	8.5	0	9	9
	9	0	5	5
Total		371	150	521

		bsmt_under		
		0	1	Total
bsmt_ht	0	22	0	22
	5.5	0	3	3
	6	0	6	6
	6.5	0	40	40
	7	99	0	99
	7.5	201	0	201
	8	136	0	136
	8.5	9	0	9
	9	5	0	5
Total		472	49	521

One last set of transformations needs to be done. The quality and condition variables are now linearized, which means they are now multiplicative variables. Because they are factors to be multiplied against price, this makes them impossible to use in an additive model (this is analogous to the difference between percentage adjustments and dollar adjustments in an appraisal using the direct comparison approach). However, we can account for these multiplicative factors by applying them to all of the building related area variables. The following transformations will carry out that task – add them to your syntax file and run them:

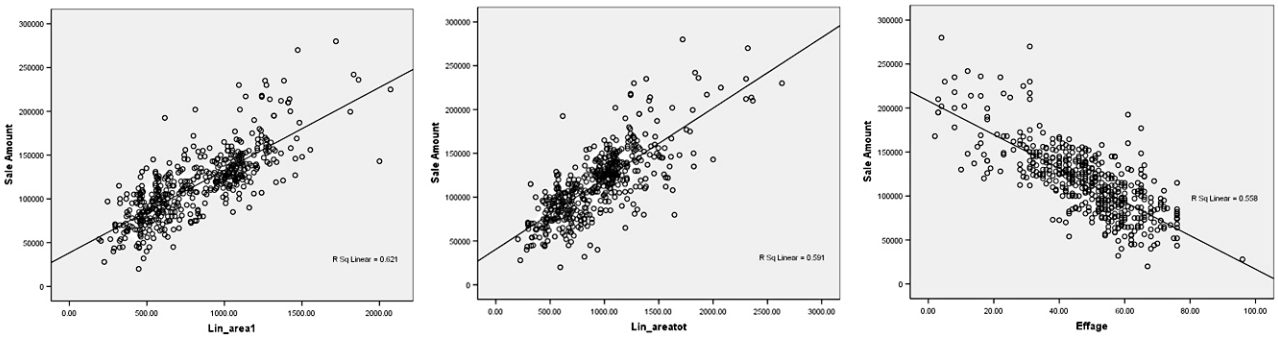
```
COMPUTE Qual_Cond = Lin_Qual * Lin_Cond.
COMPUTE Lin_areatot = Qual_Cond * area_tot.
COMPUTE Lin_area1 = Qual_Cond * area1.
COMPUTE Lin_area2 = Qual_Cond * area2.
COMPUTE Lin_areau = Qual_Cond * areau.
COMPUTE Lin_bsmtfin = Qual_Cond * bsmtfin.
COMPUTE Lin_attgar = Qual_Cond * garatta.
COMPUTE Lin_detgar = Qual_Cond * gardeta.
COMPUTE Lin_carport = Qual_Cond * garcpta.
COMPUTE Lin_shed = Qual_Cond * shedarea.
EXECUTE.
```

This completes all of the transformations we know are necessary at this point. We will now move on to examining these new variables.

STEP 5: Examining the Transformed Variables

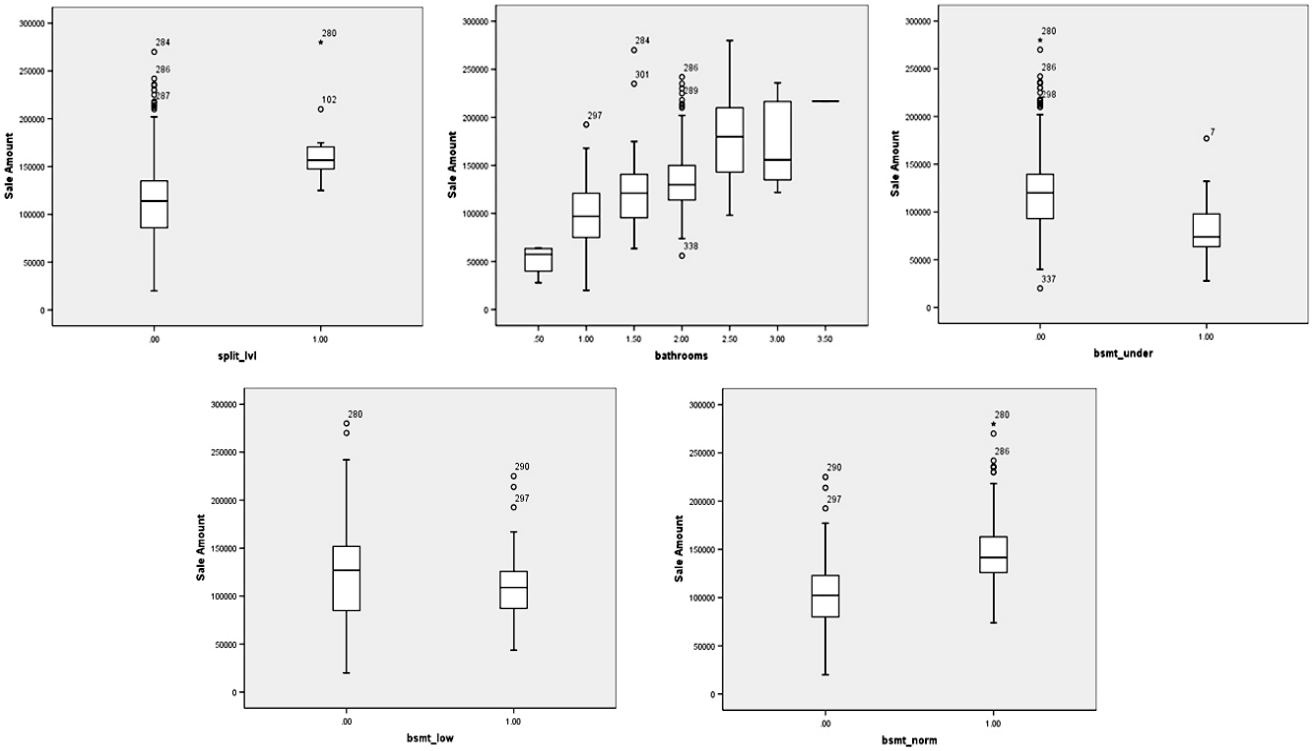
We will examine the newly transformed variables using the techniques applied in Step 3: scatterplots, boxplots, and a correlation matrix.

The new linearized variables and the effective age variable should be checked against sale price, to verify the strength of the relationship. Scatterplots of lin_areatot, lin_area1, and effage will be shown here.



These show strong relationships. The application of condition and quality has strengthened the correlation between the two main area variables and the sale price. As expected there is a negative relationship between age and selling price.

We have already examined many variables in Step 3 with boxplots, so here we will only look at the new variables: split_lvl, bathrooms, bsmt_under, bsmt_low, and bsmt_norm.



The new split_lvl variable shows its impact on sale price as does the bathroom count variable. For the basement variables, bmst_under and bsmt_norm show some separation, but bsmt_low does not. It can be eliminated from further analysis and will act as the control value for this set of binary variables. .

Next we examine a new correlation matrix, replacing the non-linearized building area values with their linearized counterparts and adding all the new variables except bsmt_low.

There are no correlations of great concern: none outside of ± 0.8, but a few approaching that threshold. The linearized first floor area and the linearized total area are at 0.794, meaning they should not be in a model together. We have already discussed frontage and lotsize. We need to watch carefully the linearized basement finish variable and the linearized first floor area, as well as the effective age variable with both the first floor area and the normal height basement variable. It appears that newer homes have larger first floor areas and normal height basements. This last relationship certainly makes sense.

These relationships indicate that the application of the quality and condition multipliers to the building area variables has taken the age of the building into account already. Therefore, the effective age variable will not be necessary in further analysis.

The database is ready for model specification.

Correlations

	sale_amt	ab_comm	bathrooms	bedrooms	bsmt_under	bsmt_norm	depth	Effage	fireplcs	frontage	Lin_area1	Lin_area2	Lin_areatot	Lin_areau	Lin_attgar	Lin_bsmtfin	Lin_carport	Lin_detgar	Lin_shed	lotsize	RiversBend	WinstonWell	KristalEstates	SussexEast	SussexWest	KellyJuno	Kingston	Mohawk	ElizabethJune	porchpts	split_lvl
sale_amt	1	-.131**	.519**	.408**	-.312**	.544**	.058	-.747**	.397**	.377**	.788**	.135**	.769**	-.025	.397**	.541**	.068	.129**	.053	.350**	-.124**	.246**	.478**	-.050	-.155**	.346**	.093*	-.179**	-.310**	-.180**	.226**
ab_comm	-.131**	1	-.007	-.108*	-.022	-.069	.021	.082	-.042	.006	-.079	-.031	-.088*	-.011	.054	-.123**	-.033	-.019	.048	.007	-.004	-.082	-.036	-.011	.045	-.062	-.053	-.067	.371**	.000	.030
bathrooms	.519**	-.007	1	.381**	-.179**	.344**	-.014	-.454**	.369**	.188**	.501**	.143**	.528**	.072	.293**	.535**	.104*	.046	.063	.141**	.059	.107*	.271**	.024	.048	-.117**	.174**	.053	-.040	-.123**	-.050	-.110*	.102*
bedrooms	.408**	-.108*	.381**	1	-.135**	.205**	-.007	-.274**	.188**	.157**	.380**	.363**	.555**	.087*	.126**	.336**	.058	.021	.011	.125**	-.106*	.217**	.042	-.042	.114**	-.088*	.053	-.040	-.056	.222**	.218**	.001	.080
bsmt_under	-.312**	-.022	-.179**	-.135**	1	-.205**	.027	.332**	-.113**	-.130**	-.298**	.151**	-.164**	.112*	-.081	-.254**	-.058	-.032	-.052	-.088*	.057	-.127**	-.063	-.042	.114**	-.088*	-.044	-.056	.222**	.218**	-.057
bsmt_norm	.544**	-.069	.344**	.205**	-.205**	1	-.026	-.639**	.251**	.185**	.528**	-.060	.423**	-.038	.263**	.416**	.143**	.056	.018	.126**	-.073	.193**	.306**	-.036	-.088*	.271**	-.136**	-.073	-.155**	-.195**	.133**
depth	.058	.021	-.014	-.007	.027	-.026	1	.009	-.033	-.158**	-.016	1	.080	.035	.008	-.003	-.058	-.059	.030	.558**	-.087*	-.097*	.064	-.013	-.121**	.095*	-.158**	-.016	.011	.044	.013
Effage	-.747**	.082	-.454**	-.274**	.332**	-.639**	.009	1	-.296**	-.270**	-.701**	.073	-.564**	.095*	-.290**	-.501**	-.105*	-.065	-.037	-.205**	.034	-.247**	-.460**	.036	.111*	-.250**	.094*	.184**	.253**	.276**	-.193**
fireplcs	.397**	-.042	.369**	.188**	-.113**	.251**	-.033	-.296**	1	.126**	.365**	.154**	.413**	.018	.240**	.317**	.000	.041	.026	.079	-.094*	.043	.176**	-.058	-.086*	.154**	.066	-.072	-.021	.005	.085
frontage	.377**	.006	.188**	.157**	-.130**	.185**	-.158**	-.270**	.126**	1	.360**	-.061	.274**	-.076	.291**	.239**	.070	.103*	.055	.711**	.123**	.127**	.073	.073	.031	.237**	.015	-.177**	-.271**	-.152**	.058
Lin_area1	.788**	-.079	.501**	.380**	-.298**	.528**	-.016	-.701**	.365**	.360**	1	-.131**	.794**	-.022	.339**	.620**	.108*	.103*	-.006	.281**	-.137**	.342**	.347**	-.041	-.151**	.323**	-.058	-.216**	-.180**	-.292**	.184**
Lin_area2	.135**	-.031	.143**	.363**	.151**	-.060	.080	.073	.154**	-.061	-.131**	1	.497**	.156**	.193**	-.150**	-.007	-.049	.096*	-.003	-.048	-.094*	-.019	-.078	-.029	.020	.135**	.063	.103*	.294**	.102*
Lin_areatot	.769**	-.088*	.528**	.555**	-.164**	.423**	.035	-.564**	.413**	.274**	.794**	.497**	1	.125**	.412**	.450**	.090*	.063	.055	.241**	-.150**	.241**	.291**	-.084	-.150**	.294**	.030	.152**	-.087*	-.070	.222**
Lin_areau	-.025	-.011	.072	.087*	.112*	-.038	.008	-.064**	-.076	-.022	.156**	.125**	1	-.015	-.028	-.150**	-.011	.037	.034	-.058	-.023	-.027	-.012	-.017	-.015	-.020	-.017	-.022	.179**	.169**	-.011
Lin_attgar	.397**	.054	.293**	.126**	-.081	.263**	-.003	-.290**	.240**	.239**	.339**	.193**	.412**	-.015	1	.252**	-.045	-.142**	-.062	.232**	-.097*	.190**	.036	-.043	.278**	-.045	-.059	-.050	-.002	.161**
Lin_bsmtfin	.541**	-.123**	.535**	.336**	-.254**	.416**	-.058	-.501**	.317**	.239**	.620**	-.150**	.450**	-.028	.252**	1	.104*	.057	-.049	.161**	-.102*	.333**	.088*	-.044	-.100*	.291**	-.042	-.161**	-.195**	-.233**	.098*
Lin_carport	.068	-.033	.104*	.058	-.058	.143**	-.059	-.105*	.000	.070	.108*	-.007	.090*	-.011	.104*	1	-.053	-.006	.012	.032	.094*	-.035	.099*	-.045	.022	-.052	-.033	-.060	-.100*	.190**
Lin_detgar	.129**	-.019	.046	.063	-.032	.056	.132**	-.065	.041	-.032	.103*	-.049	.063	.037	-.142**	.057	-.049	1	-.045	.182**	-.035	.043	.038	-.056	.010	-.031	-.070	-.026	-.022	-.064	.017
Lin_shed	.053	.048	.063	.011	-.052	-.018	.030	-.037	.026	.055	-.006	.096*	.055	.034	-.062	-.049	-.006	-.045	1	.069	.004	.010	.051	-.056	.010	.013	.009	-.017	.028	.000	.018
lotsize	.350**	.007	.141**	.125**	-.088*	.126**	.558**	-.205**	.079	.711**	.281**	-.003	.241**	-.058	.232**	.161**	.012	.182**	.069	1	.027	.027	.108*	.054	-.049	.275**	-.092*	-.139**	-.218**	-.101*	.038
RiversBend	-.124**	-.004	-.059	-.106*	.057	-.073	-.087*	.034	-.094*	.123**	-.137**	-.048	-.150**	-.023	-.023	-.102*	.032	-.035	.004	.027	1	-.169**	-.073	-.107*	-.095*	-.128**	-.109*	-.137**	-.126**	.019	-.034
WinstonWell	.246**	-.082	.107*	.217**	-.127**	.193**	-.097*	-.247**	.043	.127**	.342**	-.094*	.241**	-.027	-.097*	.333**	.094*	.043	.010	.027	-.169**	1	-.087*	-.127**	-.113*	-.152**	-.129**	-.163**	-.149**	-.200**	.099*
KristalEstates	.478**	-.036	.271**	.042	-.063	.306**	.064	-.460**	.176**	.073	.347**	-.019	.291**	-.012	.190**	.088*	-.035	.038	.051	.108*	-.073	-.087*	1	-.055	-.049	-.066	-.056	-.071	-.065	.029	.025
SussexEast	-.050	-.011	.024	-.048	-.042	-.036	-.013	.036	-.058	.073	-.041	-.078	-.084	-.017	.036	-.044	.099*	.032	-.056	.054	-.107*	-.127**	-.055	1	-.072	-.097*	-.082	-.104*	-.095*	-.072	-.051
SussexWest	-.155**	.045	-.103*	-.117**	.114**	-.088*	-.121**	.111*	-.086*	.031	-.151**	-.029	-.150**	-.015	-.043	-.100*	-.045	-.072	.010	-.049	-.095*	-.113*	-.049	-.072	1	-.086	-.073	-.092*	-.084	.010	-.045
KellyJuno	.346**	-.062	.172**	.174**	-.088*	.271**	.095*	-.250**	.154**	.237**	.323**	.020	.294**	-.020	.278**	.291**	.022	-.031	.013	.275**	-.128**	-.152**	-.066	-.097*	-.086	1	-.098*	-.124**	-.113**	-.080	.085
Kingston	.093*	-.053	-.047	.053	-.044	-.136**	-.158**	.094*	.066	.015	-.058	.135**	.030	-.017	-.045	-.042	-.052	-.070	.009	-.092*	-.109*	-.129**	-.056	-.082	-.073	-.098*	1	-.105*	-.096*	.010	-.010
Mohawk	-.179**	-.067	-.123**	-.040	-.056	-.073	-.016	.184**	-.072	-.177**	-.216**	.063	-.152**	-.022	-.059	-.161**	-.033	-.026	-.017	-.139**	-.137**	-.163**	-.071	-.104*	-.092*	-.124**	-.105*	1	-.121**	.024	-.030
ElizabethJune	-.310**	.371**	-.050	-.072	.222**	-.155**	.011	.253**	-.021	-.271**	-.180**	.103*	-.087*	.179**	-.050	-.195**	-.060	-.022	.028	-.218**	-.126**	-.149**	-.065	-.095*	-.084	-.113**	-.096*	-.121**	1	.217**	-.022
porchpts	-.180**	.000	-.110*	.001	.218**	-.195**	.044	.276**	.005	-.152**	-.292**	.294**	-.070	.169**	-.002	-.233**	-.100*	-.064	.000	-.200**	.029	.010	.010	.024	.217**	1	-.076				
split_lvl	.226**	.030	.102*	.080	-.057	.133**	.013	-.193**	.085	.058	.184**	.102*	.222**	-.011	.161**	.098*	.190**	.017	.018	.038	-.034	.099*	.025	-.051	-.045	.085	-.010	-.030	-.022	-.076	1

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

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

STEP 6: List the Variables for Calibration

The next step in the variable selection process is to use regression to evaluate the candidate variables. From the preceding analysis we have determined a final list of variables for the variable selection process. We will try four different combinations. The following 22 variables will be common to each of the combinations:

ab_comm	lin_carport	SussexEast
bathrooms	lin_detgar	SussexWest
bedrooms	lin_shed	KellyJuno
bsmt_under	porchpts	Kingston
bsmt_norm	split_lvl	Mohawk
fireplcs	RiversBend	ElizabethJune
lin_attgar	WinstonWell	
lin_bsmtfin	KristalEstates	

Below are the variables that will differentiate our four combinations. They represent different arrangements of the variables that will account for building size and land size in the regression:

- Model 1: lin_areatot and lotsize
- Model 2: lin_areatot and frontage and depth
- Model 3: lin_area1, lin_area2, and lin_areau and lotsize
- Model 4: lin_area1, lin_area2, and lin_areau and frontage and depth

We will run the regressions as follows:

- Select Analyze → Regression → Linear and enter the above variables as Independent variables and sale_amt as the Dependent variable.
- Ensure the Method is Enter.
- Click the Statistics button, and select R squared change, Estimates, Model Fit, Descriptives, and Collinearity Diagnostics → Continue.
- We will be running four sets of regressions, so you may wish to save the syntax file for easy runs of the three other models. You can do this by clicking Paste, and the commands will be pasted into the open syntax file (or a new syntax file). You can then run this by selecting it and clicking Run. You can change the variables and run it again. This is very helpful if you want to re-run models later.[2]

Helpful Hint

When running multiple iterations of a model, it is important to remember to substitute the correct variables for each other. A common error is to leave one or more variables in the regression, such as lin_areatot with lin_area1 (and perhaps lin_area2 and lin_areau). This will bring multicollinearity in your regressions and give you incorrect results!

[2] 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 (however, you must click paste before running the regression or other procedure). 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 use your mouse to select the regression section of the file, from the REGRESSION command to the period at the end of the "/METHOD" line, and then 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.

The Model Summary, ANOVA and Coefficients Tables are shown for the first regression, but to save space only the Model Summaries are shown for the other three models:

Model 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.911[a]	.830	.822	16357.594

a. Predictors: (Constant), Lot Size, Abuts Commerical, RiversBend, Lin_carport, Lin_shed, SussexWest, Total Porch Points, Total Number of Fireplaces, Kingston, split_lvl, Lin_detgar, KristalEstates, SussexEast, bsmt_under, Total Number of Bedrooms, Mohawk, Lin_attgar, KellyJuno, bsmt_norm, bathrooms, ElizabethJune, Lin_bsmtfin, WinstonWell, Lin_areatot

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	648258858401	24	27010785767	100.948	.000[b]
	Residual	132715156991	496	267570881		
	Total	780974015392	520			

a. Dependent Variable: Sale Amount

b. Predictors: (Constant), Lot Size, Abuts Commerical, RiversBend, Lin_carport, Lin_shed, SussexWest, Total Porch Points, Total Number of Fireplaces, Kingston, split_lvl, Lin_detgar, KristalEstates, SussexEast, bsmt_under, Total Number of Bedrooms, Mohawk, Lin_attgar, KellyJuno, bsmt_norm, bathrooms, ElizabethJune, Lin_bsmtfin, WinstonWell, Lin_areatot

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	44511.843	4326.545		10.288	.000		
Abuts Commerical	-91.052	4510.427	.000	-.020	.984	.800	1.250
bathrooms	2462.689	1944.595	.032	1.266	.206	.539	1.856
Total Number of Bedrooms	-299.065	1128.794	-.006	-.265	.791	.620	1.612
bsmt_under	-11889.058	2682.782	-.090	-4.432	.000	.837	1.194
bsmt_norm	8772.757	1981.915	.103	4.426	.000	.638	1.568
Total Number of Fireplaces	2982.636	1808.266	.035	1.649	.100	.760	1.315
Lin_attgar	15.972	8.556	.043	1.867	.063	.654	1.529
Lin_bsmtfin	12.530	3.224	.101	3.886	.000	.506	1.977
Lin_detgar	14.756	4.014	.073	3.676	.000	.872	1.147
Lin_carport	-15.106	14.740	-.020	-1.025	.306	.893	1.119
Lin_shed	8.571	16.766	.010	.511	.609	.950	1.052
split_lvl	13445.017	4391.275	.060	3.062	.002	.895	1.118
Total Porch Points	-45.663	81.756	-.011	-.559	.577	.834	1.200
RiversBend	6607.620	2822.071	.056	2.341	.020	.591	1.693
WinstonWell	16142.686	2925.449	.156	5.518	.000	.431	2.318
KristalEstates	64875.222	4804.475	.314	13.503	.000	.633	1.579
SussexEast	7887.460	3313.845	.054	2.380	.018	.675	1.481
SussexWest	2666.920	3567.286	.016	.748	.455	.721	1.387
KellyJuno	22659.090	3355.786	.178	6.752	.000	.491	2.037
Kingston	24569.625	3325.804	.169	7.388	.000	.655	1.527
Mohawk	1942.114	2893.846	.016	.671	.502	.593	1.686
ElizabethJune	-12521.548	3300.572	-.097	-3.794	.000	.525	1.906
Lin_areatot	42.341	3.157	.404	13.412	.000	.377	2.649
Lot Size	1.847	.441	.089	4.183	.000	.765	1.307

a. Dependent Variable: Sale Amount

Model 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
2	.911[a]	.830	.822	16368.972

a. Predictors: (Constant), Actual Depth, Lin_attgar, SussexEast, Abuts Commerical, Total Porch Points, Lin_shed, Mohawk, Lin_carport, SussexWest, Total Number of Bedrooms, KristalEstates, Lin_detgar, Kingston, split_lvl, bsmt_under, RiversBend, Total Number of Fireplaces, KellyJuno, Actual Frontage, bsmt_norm, bathrooms, ElizabethJune, Lin_bsmtfin, WinstonWell, Lin_areatot

Model 3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
3	.913[a]	.834	.825	16196.953

a. Predictors: (Constant), Lot Size, Lin_area2, Lin_carport, Abuts Commerical, RiversBend, SussexWest, Lin_shed, KristalEstates, Lin_areau, SussexEast, Lin_detgar, split_lvl, Kingston, bsmt_under, Total Number of Fireplaces, Mohawk, Total Porch Points, KellyJuno, Total Number of Bedrooms, Lin_attgar, bsmt_norm, bathrooms, ElizabethJune, Lin_bsmtfin, WinstonWell, Lin_area1

Model 4

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
4	.914[a]	.835	.825	16190.986

a. Predictors: (Constant), Actual Depth, Lin_attgar, Lin_areau, SussexEast, Abuts Commerical, Lin_shed, Lin_carport, Mohawk, bsmt_under, RiversBend, Lin_detgar, KristalEstates, SussexWest, split_lvl, Total Porch Points, Total Number of Bedrooms, Kingston, Total Number of Fireplaces, KellyJuno, Actual Frontage, bsmt_norm, Lin_area2, bathrooms, ElizabethJune, Lin_bsmtfin, WinstonWell, Lin_area1

These model summaries show very little difference between the four models. However, Model #4 has the highest value of R^2 and lowest value for the SEE. This appears to be the optimal group of variables, so we will continue with them.

One issue that must be examined at this stage is the VIF (or tolerance) values. VIF values greater than 3.33 (or tolerance values less than 0.3) indicate multicollinearity to a degree that MUST be addressed. High VIF values can cause artificial results in the adjusted R2 and SEE statistics. It does not occur in this case, but if there were a variable or variables with that condition, the best approach is to remove one problem variable at a time until the issue disappears. If one of the variables with a VIF $>$ 3.33 is a very important variable (such as square footage of living area), remove the variable with the next greatest VIF. Eventually all VIF values will be less than 3.33 and you can choose your "best" variable combination from the remaining variables.

Creating the Model and Test Database

Step 8 states that mass appraisal models should be fully tested before they are used in real application (e.g., to estimate the market value of a city's properties for assessment purposes). The performance of a model can be tested by comparing the estimates produced by the model to actual sales observations. However, if the sales used to calibrate the model are also used in testing the model, then the results can appear better than they really are. This is because a few properties with specific characteristics may skew the results, which was referred to in Lesson 7 as "chasing the sales". A better practice in mass appraisal model building is to withhold a portion of the sales database when calibrating the model and then test the model against this group of sales. Because the model will eventually be applied to properties outside of the database, this testing method ensures that generalizing the results outside of the sample database will produce accurate results.

The model should only be tested in this manner if there are sufficient sales remaining to calibrate the model. As few as 30 sales may be sufficient to calibrate a simple model. But, as stated in Lesson 7, this restricts the number of variables that can be included in the model. Most model builders prefer to have a database with at

least five times as many sales as there are variables in the original database. The Ontario database contained 62 original variables, so the desired minimum number of sales for calibration is 310.

For testing a model, the minimum number of sales is also 30, but it is desirable to have at least 100 to allow some stratification in the testing process (stratification refers to separation into subgroups, i.e. by neighbourhood). Since the Ontario database contains 521 sales, the desirable levels can be attained for both the model and test databases. Given 521 sales in the entire database, two-thirds or 347 sales will be used to create the model and 174 sales will be reserved for testing the model.

Just to make sure we are all on the same page, we have provided a fresh copy of the "Ontario521" database, with all the necessary transformations carried out. The new file is called "Ontario521complete.sav". You may continue with the file you have or you may use this new file.

We will now look at the variable named Random. This is a random number between 0 and 100 and you will find the "Ontario521complete" database is sorted based on this variable.[3] We will use the Random variable to create two subsets of the "Ontario521complete" database. The 347 sales with the lowest values for RANDOM will become "OntarioModel", used to calibrate the model; the 174 sales with the highest values for RANDOM will become "OntarioTest", the database we will use for testing the model.

CAUTION

At this point, you should make a backup of the "Ontario521complete" database in case you mistakenly overwrite it. One further note on saving databases: when you save the "OntarioModel" and "OntarioTest" databases below, ensure you use Save As... and not Save ("Save" would overwrite the Ontario521Complete database).

Go to the Data View window, locate the variable Random, and scroll down to rows 347 and 348.

Helpful Hint

A quick method to locate the Random variable is to click Utilities → Variables and then Random → Go To.

The value of the random number at row #347 in the spreadsheet display is 61.41 and #348 is 61.44. We will use these numbers to split the database randomly. If you click on the #347 Random cell the value displayed in the variable information box at the top of the spreadsheet is 61.4128568675369. Because the spreadsheet display is rounded, we will use a value between those displayed for these two rows to split the file (i.e., 61.42).

Now, set a filter to select value of Random $<$ 61.42 and specify that unselected cases should be deleted.

- Data → Select Cases → If Condition is Satisfied → If…
- Random $<$ 61.42
- Continue → Delete unselected cases → OK (deletes cases not meeting the selection criterion).

Check the database to ensure that there are exactly 347 cases. Values of the variable Random should range from .08 to 61.41 (this can be checked in Data View). When satisfied, click File → Save As and name the resulting file OntarioModel.sav.

[3] To sort the database, you can use Data → Sort Cases and select Random as the Sort By variable. Ensure Ascending is selected and press OK.

To create the test database, proceed as follows:

- File → Open → Data → Ontario521Complete.sav (re-opens the Ontario data base with all 521 cases).
- Data → Select Cases → If Condition is Satisfied → If…
- Random > 61.42
- Continue → Delete unselected cases → OK (deletes cases not meeting the selection criterion).

Check to ensure there are 174 remaining cases. Values of Random should range from 61.44 to 99.83. Then click File → Save As and name the file OntarioTest.sav. This saves the test file for future use.

At this point there are three data files:

1. Ontario521complete.sav with 521 cases;
2. OntarioModel.sav with 347 cases; and
3. OntarioTest.sav with 174 cases.

Open OntarioModel.sav, the file to be used in model development. Again, ensure that the file has 347 cases.

List a Final Group of Potential Variables for Calibration in the Model Database

Previously we established the variables to be used in an additive model calibration. Those selected here are the variables from Model #4 earlier:

lin_area1	lin_detgar
lin_area2	lin_shed
lin_areau	RiversBend
frontage	WinstonWell
depth	KristalEstates
ab_comm	SussexEast
bathrooms	SussexWest
bedrooms	KellyJuno
bsmt_under	Kingston
bsmt_norm	Mohawk
fireplcs	ElizabethJune
lin_attgar	porchpts
lin_bsmtfin	split_lvl
lin_carport	

There are a few binary variables in this group, so the next procedure is important to ensure that there are enough sales of each of the binary variables in the reduced database to allow them to be included.

Use Analyze → Descriptive Statistics → Frequencies to determine the number of cases equal to 1 for ab_comm, bsmt_under, bsmt_norm, split_lvl, and the nine neighbourhood variables.

Binary Frequencies (summarized in a table)

Variable	= 0 (no)	= 1 (yes)
ab_comm	339	8
bsmt_under	319	28
bsmt_norm	243	104
split_lvl	339	8
RiversBend	307	40
WinstonWell	287	60
KristalEstates	332	15
SussexEast	318	29
SussexWest	326	21
KellyJuno	313	34
Kingston	318	29
Mohawk	305	42
ElizabethJune	315	32

The variables split_lvl and ab_comm could be eliminated from further consideration because they each have only eight sales – above the 5 minimum, but well below the 5%, or 17, preferred (we want 17 now that our model database has 347 cases). However, we will make the decision to keep them in the mix. Generally, it is better to have more variables rather than fewer going into a regression. You can let the results of regression help you decide which variables are significant. Kristal Estates, at 15 sales, is slightly below the 17 threshold, but is well above the 5 and as such it also will be retained.

We will now use the stepwise regression technique to build our model. Our model is based on additive multiple regression. As mentioned earlier, the procedures described in this lesson will follow a middle course which will allow for both a good prediction of market value and variable coefficients which are reasonable and rational from an appraisal perspective. In order to achieve this dual goal, 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 0.1%).

However, keep in mind that these criteria are not clear-cut in all cases and occasionally must be tempered by appraisal judgment. Occasionally there will be conflicts in achieving all three of these standards for some variables.

We will now proceed with a stepwise regression as the final step in selecting the variables for our model:

- Analyze → Regression → Linear.
- Select sale_amt as the Dependent variable.
- Select the following as Independent variables:

Lin_area1	fireplcs	SussexEast
Lin_area2	Lin_attgar	SussexWest
Lin_areau	Lin_bsmtfin	KellyJuno
frontage	Lin_carport	Kingston
depth	Lin_detgar	Mohawk
ab_comm	Lin_shed	ElizabethJune
bathrooms	split_lvl	porchpts
bedrooms	RiversBend	
bsmt_under	WinstonWell	
bsmt_norm	KristalEstates	

Ensure the Method is Stepwise.

- Click the Statistics button and select Estimates, Model Fit, Descriptives, R Squared Change, and Collinearity Diagnostics → Continue.
- Click the Options button. Under "Use probability of F", put in 0.15 for Entry and 0.20 for Removal → Continue. (A variable is entered into the model if the significance level of its F value is less than the Entry value and is removed if the significance level is greater than the Removal value; if you wanted fewer variables, you could reduce the PIN and POUT criteria; to include more variables you could increase PIN and POUT).
- You may wish to click Paste here, to save the commands to a syntax file.
- Click OK to run, or if you used Paste, you may go to the syntax file, block select the regression commands, and run it from there.

Your SPSS output should include the following:

Stepwise Regression: Model Summary #4

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782[a]	.611	.610	24164.787
2	.817[b]	.667	.665	22384.929
3	.839[c]	.705	.702	21118.328
4	.865[d]	.749	.746	19509.290
5	.873[e]	.763	.759	18974.624
6	.883[f]	.780	.776	18312.113
7	.888[g]	.789	.785	17946.551
8	.893[h]	.798	.793	17594.729
9	.897[i]	.804	.799	17343.400
10	.900[j]	.810	.804	17113.969
11	.903[k]	.815	.809	16894.555
12	.905[l]	.819	.812	16765.679
13	.906[m]	.821	.814	16696.633
14	.907[n]	.822	.815	16643.914
15	.908[o]	.824	.816	16579.084

a. Predictors: (Constant), Lin_area1
b. Predictors: (Constant), Lin_area1, KristalEstates
c. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2
d. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune
e. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston
f. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno
g. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar
h. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin
i. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under
j. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell
k. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm
l. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm, Actual Frontage
m. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm, Actual Frontage, bathrooms
n. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm, Actual Frontage, bathrooms, split_lvl
o. Predictors: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm, Actual Frontage, bathrooms, split_lvl, Actual Depth

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
15	(Constant)	32185.002	6816.616		4.722	.000		
	Lin_area1	51.307	4.516	.442	11.362	.000	.351	2.850
	KristalEstates	57907.413	5420.181	.305	10.684	.000	.652	1.534
	Lin_area2	32.487	4.589	.182	7.080	.000	.806	1.240
	ElizabethJune	-17911.015	3348.905	-.134	-5.348	.000	.844	1.185
	Kingston	23049.902	3466.296	.165	6.650	.000	.861	1.162
	KellyJuno	19859.178	3620.589	.153	5.485	.000	.684	1.463
	Lin_detgar	16.685	4.798	.082	3.477	.001	.950	1.053
	Lin_bsmtfin	8.542	4.062	.070	2.103	.036	.482	2.075
	bsmt_under	-10910.622	3540.461	-.077	-3.082	.002	.852	1.174
	WinstonWell	10294.309	2928.273	.101	3.515	.001	.646	1.548
	bsmt_norm	7704.657	2391.898	.091	3.221	.001	.660	1.516
	Actual Frontage	201.065	75.416	.069	2.666	.008	.786	1.272
	bathrooms	4348.916	2308.513	.058	1.884	.060	.568	1.759
	split_lvl	11699.757	6135.930	.045	1.907	.057	.934	1.071
	Actual Depth	70.874	37.346	.046	1.898	.059	.903	1.107

a. Dependent Variable: Sale Amount

Excluded Variables[a]

Model		Beta in	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
15	Abuts Commerical	-.003[p]	-.115	.909	-.006	.798	1.253	.351
	Total Number of Bedrooms	.012[p]	.418	.676	.023	.640	1.563	.337
	Total Number of Fireplaces	.037[p]	1.394	.164	.077	.746	1.340	.341
	Lin_attgar	-.008[p]	-.295	.768	-.016	.708	1.412	.344
	Lin_carport	-.006[p]	-.240	.811	-.013	.921	1.086	.351
	Lin_shed	-.010[p]	-.436	.663	-.024	.930	1.075	.350
	Total Porch Points	.016[p]	.609	.543	.034	.793	1.261	.343
	RiversBend	.024[p]	.927	.355	.051	.811	1.233	.350
	SussexEast	.026[p]	1.040	.299	.057	.871	1.149	.351
	SussexWest	.004[p]	.142	.887	.008	.867	1.153	.351
	Mohawk	.002[p]	.094	.925	.005	.804	1.244	.351
	Lin_areau	.009[p]	.370	.711	.020	.896	1.115	.350

a. Dependent Variable: Sale Amount
p. Predictors in the Model: (Constant), Lin_area1, KristalEstates, Lin_area2, ElizabethJune, Kingston, KellyJuno, Lin_detgar, Lin_bsmtfin, bsmt_under, WinstonWell, bsmt_norm, Actual Frontage, bathrooms, split_lvl, Actual Depth

The output shows the fifteen steps taken by the Stepwise procedure and the variables introduced at each step. The statistics we are most interested in are only those in step number 15 in all of the tables.

The following list summarizes the critical values at each of the fifteen steps in the step-wise process:

Step	Variable Added	Adjusted R^2	Standard Error
1	Lin_area1	.610	24,165
2	KristalEstates	.665	22,385
3	Lin_area2	.702	21,118
4	ElizabethJune	.746	19,509
5	Kingston	.759	18,975
6	KellyJuno	.776	18,312
7	Lin_detgar	.785	17,947
8	Lin_bsmtfin	.793	17,595
9	bsmt_under	.799	17,343
10	WinstonWell	.804	17,114
11	bsmt_norm	.809	16,895
12	frontage	.812	16,766
13	bathrooms	.814	16,697
14	split_lvl	.815	16,644
15	depth	.816	16,579

The Excluded Variables table shows which variables were not in the model at each step. The most important step is 15, as this shows the 12 variables excluded at this point in this process. All of these had t-values below the stated preference of 1.6, and thus also have probabilities below the acceptable 90% confidence level. This can be seen by subtracting Sig. (the probability that the coefficient is equal to zero) from one. For example, the Mohawk neighbourhood has a significance of .925 indicating a 92.5% probability that this coefficient is equal to zero, or, in other words, a confidence level of only 7.5%. This is far too low to justify attempting to force this variable into the model.

The PIN and POUT thresholds that were chosen for this stepwise regression ("probability in" of 0.15 and "probability out" 0.20 respectively) were set so that the variables included in the model would meet our three criteria. With a t-statistic of 1.394 the variable fireplcs is just on the edge of acceptability, it could be forced into the model or remain excluded from the model without too much impact either way. We will leave it out of the model.

Ignoring fireplaces we find no other excluded variables that are close to the t-statistic interval of $\pm$ 1.6. The two neighbourhoods SussexEast and RiversBend are the closest at 1.040 and 0.927 respectively, far enough from the 1.6 critical value that we would not consider forcing them into the model. As such, we will make no adjustments to the PIN and POUT settings.

For our 15 included variables, the model has produced an adjusted R^2 of 0.816 and SEE of 16,579. The mean value for sale_amt in this data base is 115,784, which means that the COV is 16,579 $\div$ 115,784 = 14.32%. A COV of between 10% and 20% is indicative of a good model, and a COV under 10% would be an excellent result. The F statistic is 103.56 with a Significance of 0.000. It is important that this Significance (or probability of the F statistic being this large by "chance") is less than 0.05 or 5%.

All of the VIF (and tolerance) statistics look fine, with VIF $<$ 3.3 and tolerance $>$ 0.3, indicating no evidence of multicollinearity in the model.

It appears that we have found our "best" model. It uses fifteen variables to estimate the sale price (although in strict model building terms, the five neighbourhood variables can be considered one variable):

Lin_area1	bsmt_norm	KristalEstates
Lin_area2	frontage	ElizabethJune
Lin_detgar	bathrooms	Kingston
Lin_bsmtfin	split_lvl	KellyJuno
bsmt_under	depth	WinstonWell

This completes the variable selection process.

STEP 7: Model Calibration

Now that we have a final list of variables to include in the model, our model is ready to be calibrated. In other words, we can now calculate the coefficients for each of the variables in the regression model. These coefficients were displayed in the final step of the stepwise process, but the Enter process produces a more compact report 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:

- Analyze → Regression → Linear.
- Select sale_amt as the Dependent variable.
- Select the following as Independent variables: Lin_area1, Lin_area2, frontage, depth, bathrooms, bsmt_under, bsmt_norm, split_lvl, Lin_bsmtfin, Lin_detgar, KristalEstates, ElizabethJune, Kingston, KellyJuno, and WinstonWell.
- Ensure the Method is Enter.
- Click the Statistics button, and select Estimates, Model Fit, Descriptives, Collinearity Diagnostics, Casewise Diagnostics (outside 3 standard deviations) → Continue.
- Click Plots and check Histogram and Normal Probability Plot → Continue.
- Click Save and check Unstandardized for Predicted Values and Unstandardized for Residuals → Continue. (This will create two new variables, PRE_1 which are the model's estimated sale prices, and RES_1 which are the residual values between the predicted and actual sale prices. We will use these values in a scatterplot.)
- Paste, to paste the regression commands into a syntax file.
- Click OK to run, or if you used Paste, you may go to the syntax file, block select the regression commands and run it from there.

Your SPSS output should include the following:

Enter Regression: 15 Variables

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Actual Depth, bsmt_under, KristalEstates, Lin_detgar, split_lvl, Kingston, KellyJuno, ElizabethJune, Lin_area2, WinstonWell, Actual Frontage, bathrooms, bsmt_norm, Lin_bsmtfin, Lin_area1[b]	.	Enter

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

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.908[a]	.824	.816	16579.084

a. Predictors: (Constant), Actual Depth, bsmt_under, KristalEstates, Lin_detgar, split_lvl, Kingston, KellyJuno, ElizabethJune, Lin_area2, WinstonWell, Actual Frontage, bathrooms, bsmt_norm, Lin_bsmtfin, Lin_area1
b. Dependent Variable: Sale Amount

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	426962538311.375	15	28464169220.758	103.557	.000[b]
	Residual	90980658840.210	331	274866038.792		
	Total	517943197151.585	346			

a. Dependent Variable: Sale Amount
b. Predictors: (Constant), Actual Depth, bsmt_under, KristalEstates, Lin_detgar, split_lvl, Kingston, KellyJuno, ElizabethJune, Lin_area2, WinstonWell, Actual Frontage, bathrooms, bsmt_norm, Lin_bsmtfin, Lin_area1

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	32185.002	6816.616		4.722	.000		
bathrooms	4348.916	2308.513	.058	1.884	.060	.568	1.759
bsmt_under	-10910.622	3540.461	-.077	-3.082	.002	.852	1.174
bsmt_norm	7704.657	2391.898	.091	3.221	.001	.660	1.516
Lin_bsmtfin	8.542	4.062	.070	2.103	.036	.482	2.075
Lin_detgar	16.685	4.798	.082	3.477	.001	.950	1.053
split_lvl	11699.757	6135.930	.045	1.907	.057	.934	1.071
WinstonWell	10294.309	2928.273	.101	3.515	.001	.646	1.548
KristalEstates	57907.413	5420.181	.305	10.684	.000	.652	1.534
KellyJuno	19859.178	3620.589	.153	5.485	.000	.684	1.463
Kingston	23049.902	3466.296	.165	6.650	.000	.861	1.162
ElizabethJune	-17911.015	3348.905	-.134	-5.348	.000	.844	1.185
Lin_area1	51.307	4.516	.442	11.362	.000	.351	2.850
Lin_area2	32.487	4.589	.182	7.080	.000	.806	1.240
Actual Frontage	201.065	75.416	.069	2.666	.008	.786	1.272
Actual Depth	70.874	37.346	.046	1.898	.059	.903	1.107

a. Dependent Variable: Sale Amount

Casewise Diagnostics[a]

Case Number	Std. Residual	Sale Amount	Predicted Value	Residual
172	4.757	192500	113631.15	78868.853
301	3.218	202000	148648.51	53351.490

a. Dependent Variable: Sale Amount

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	32587.95	235761.36	115784.03	35128.276	347
Residual	-45081.730	78868.852	.000	16215.729	347
Std. Predicted Value	-2.368	3.415	.000	1.000	347
Std. Residual	-2.719	4.757	.000	.978	347

a. Dependent Variable: Sale Amount

We want to confirm our model coefficients meet our expectations for size and sign (positive or negative). In general, the coefficients look reasonable:

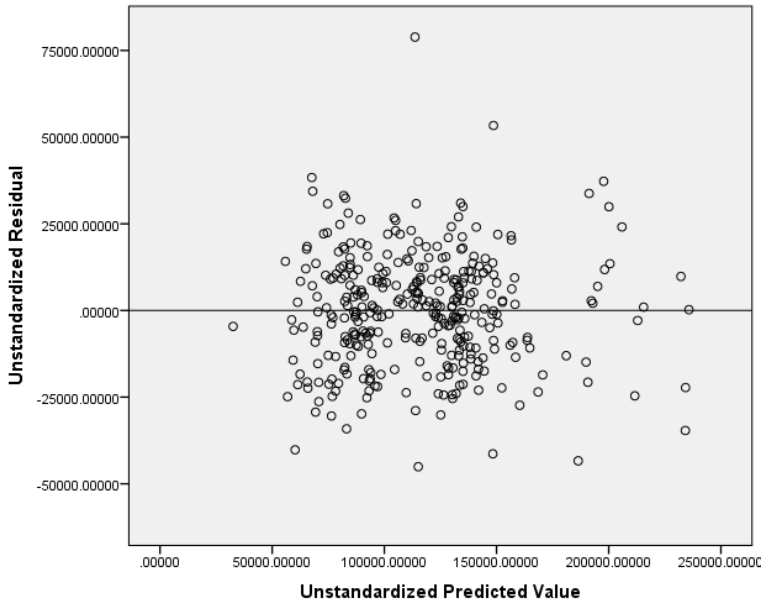
- Each square foot of first floor living area adds $51.31, while each square footage of second floor area adds $32.49.
- Basement finish is worth just over $8.50 per square foot, while a detached garage adds almost $16.70 per square foot.
- Each bathroom is worth $4,348.92 and an under-height basement is a negative influence of almost $11,000. In contrast to that, a normal height basement adds $7,705.
- A split level adds $11,700
- Each foot of frontage adds $201.07, while each foot of depth adds $70.87.
- Neighbourhoods Kristal Estates, Winston-Wellington, Kelly-Juno and Kingston are preferable to our control neighbourhood (Television Area) by amounts varying from $10,294.31 to $57,907.41; while Elizabeth-June appears to be less desirable. These neighbourhood adjustments can be validated somewhat by using local knowledge and appraisal judgment.

The Casewise Diagnostics report shows two sales with a residual (error) of more than three standard errors (standard deviations from the mean): sales 172 and sale 301. The histogram confirms the existence of two outliers. The residuals appear approximately normally distributed, which is one of the assumptions of regression analysis. The normal P-P plot examines the same distribution 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.

You should plot the residuals against several key property characteristics to ensure that the plots show a horizontal distribution randomly centred on 0. However, one crucial plot at this point is the residuals against the predicted values. Regression analysis assumes 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. On the Reference Line window, set Y Axis Position to 0). The plot should look as follows:



The residuals appear randomly distributed about 0, as desired.

At this stage we should examine the two outliers, sale 172 and sale 301. Both properties are in neighbourhood 6, but otherwise they appear to have little in common and have no obvious data problems. It seems the sale prices are simply quite high compared to the other sales in the neighbourhood, for no reason that is apparent in the data. They could be an anomaly or there could be some other influence at work on the sale prices. A deeper investigation could be undertaken at this point, to confirm why these two sales are outside the norm. However, for our purposes here we will simply filter out these two properties, setting a filter to eliminate properties with residuals outside the range of $\pm$50,000 (roughly, three times the standard deviation of the residuals: 16,215.73 $\times$ 3). We will then re-run the regression without these two sales.

- Data → Select Cases → If condition is satisfied → If…
- In the condition box enter: RES_1 > -50000 AND RES_1 < 50000 → Continue → OK.
- Make sure "Filter out unselected cases" rather than "Delete unselected cases" is selected.
- Re-run the regression, either using Dialog Recall or the commands in the syntax file.

Enter Regression: 15 Variables (Outliers Removed)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.916[a]	.839	.832	15702.235

a. Predictors: (Constant), Actual Depth, bsmt_under, KristalEstates, Lin_detgar, split_lvl, Kingston, KellyJuno, ElizabethJune, Lin_area2, WinstonWell, Actual Frontage, bathrooms, bsmt_norm, Lin_bsmtfin, Lin_area1

b. Dependent Variable: Sale Amount

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	423429418649.659	15	28228627909.977	114.490	.000[b]
	Residual	81118299171.790	329	246560179.853		
	Total	504547717821.449	344			

a. Dependent Variable: Sale Amount

b. Predictors: (Constant), Actual Depth, bsmt_under, KristalEstates, Lin_detgar, split_lvl, Kingston, KellyJuno, ElizabethJune, Lin_area2, WinstonWell, Actual Frontage, bathrooms, bsmt_norm, Lin_bsmtfin, Lin_area1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	33058.024	6465.357		5.113	.000		
	Lin_area1	53.432	4.291	.466	12.452	.000	.350	2.860
	Lin_area2	33.870	4.352	.192	7.783	.000	.805	1.241
	frontage	158.659	72.078	.055	2.201	.028	.784	1.276
	depth	66.844	35.429	.044	1.887	.060	.902	1.109
	bathrooms	4051.974	2189.557	.054	1.851	.065	.570	1.754
	bsmt_under	-10496.015	3354.763	-.075	-3.129	.002	.852	1.174
	bsmt_norm	7913.233	2270.394	.095	3.485	.001	.662	1.510
	Lin_bsmtfin	9.743	3.852	.080	2.529	.012	.484	2.068
	split_lvl	12760.942	5815.033	.050	2.194	.029	.933	1.072
	KristalEstates	56593.768	5142.426	.302	11.005	.000	.650	1.539
	KellyJuno	14681.780	3525.709	.111	4.164	.000	.683	1.464
	Kingston	22857.709	3283.110	.166	6.962	.000	.861	1.161
	ElizabethJune	-18303.101	3172.829	-.139	-5.769	.000	.844	1.185
	WinstonWell	9289.153	2779.435	.092	3.342	.001	.644	1.553
	Lin_detgar	17.730	4.547	.088	3.899	.000	.951	1.052

a. Dependent Variable: sale_amt

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	31910.39	236708.20	115311.77	35084.177	345
Residual	-44262.957	38998.887	.000	15356.074	345
Std. Predicted Value	-2.377	3.460	.000	1.000	345
Std. Residual	-2.819	2.484	.000	.978	345

a. Dependent Variable: Sale Amount

The adjusted R^2 has improved to 0.832 and the SEE has dropped slightly to 15,702. The COV has dropped slightly too, $15{,}702 \div 115{,}312 = 13.62\%$. As well, there are no longer any residual values outside of three standard deviations away from zero (the range is -2.819 to 2.484).

Helpful Hint

If some records had appeared on the Casewise Diagnostics table after this second run of the model, those records would need to be investigated and most likely removed and the model re-run. This process continues until no further outliers remain in the model. The regression statistics R^2, SEE, COV should improve slightly with each iteration of this outlier removal process.

One other important thing to note is that none of the t-statistics has dropped inside of our cutoff of ± 1.6. If this had happened a decision would have to be made whether or not to remove the affected variable(s) from the model. If the variable is deemed important from an appraisal perspective an argument can be made to retain the variable in the model. If the decision is made to remove the variable the regression is rerun without the variable and the Casewise Diagnostics and t-statistics are checked again. If low t-statistics happen to multiple variables, the variable with the t-statistic closest to zero should be removed first.

The calibration step ends when there are no further Casewise Diagnostics are reported and the t-statistics are all outside of ± 1.6 (or support has been given to keep any variables with low t-statistics).

The output from this regression satisfies our two conditions. As such, we have our final model.

Helpful Hint

This model has worked out exceptionally well. The R^2 is high, the SEE is low, and the F-statistic is large, indicating good overall results. Furthermore, all coefficients have the expected sign and a scale that makes sense, plus acceptably large t-statistics and low significance. The VIF and Tolerance are acceptable too, indicating no significant multicollinearity. However, in practice and on your project, you may find that models do not always turn out so nicely. There is no "perfect" model – you can only do the best with what you have. You may find the optimal result has some negative features, such as a lower than preferred R^2 or coefficients that do not make intuitive sense, and you may have to accept some negative aspects in return for an overall model that is the best in other measures. All of these statistics are intertwined, so improving one often means worsening another – in other words, trade-offs must be made. In choosing what is "best", sometimes modelers must choose the "least bad" option.

STEP 8: Test and Evaluate the Model

We will first examine the model in the database used to create it (OntarioModel) and then in a separate database of sales not used in the model's creation (OntarioTest). Finally, we will test the model using the entire database (Ontario521complete).

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. We could easily transform a PAR variable by dividing the model's predicted value by the sale_amt. However, there are two shortcomings to this approach. First, there is no predicted value for the two outliers excluded from the final model, so we would not have the option to include them in further testing. Second, we will need to apply the model against the test file (OntarioTest) and in the full database (Ontario521complete). To accomplish this, we will instead write a transformation that uses the model coefficients and variables to calculate the predicted values.

First, remove any filters still present. Then open a syntax file and enter the following transformations:

```
COMPUTE Predval =  33058.024 + 53.432*Lin_area1 + 33.870*Lin_area2 + 158.659*frontage
                   + 66.844*depth + 4051.974*bathrooms - 10496.015*bsmt_under
                   + 7913.233*bsmt_norm + 9.743*Lin_bsmtfin + 17.730*Lin_detgar
                   + 9289.153*WinstonWell + 56593.768*KristalEstates+14681.780*KellyJuno
                   + 22857.709*Kingston - 18303.101 * ElizabethJune.

COMPUTE PAR=       Predval/sale_amt.
```

Execute the transformations. You can confirm the Predval is correct by comparing means for PRE_1 and Predval – the mean value should be close to 115,312. If the mean of Predval is not close to the mean of PRE_1, it is likely you have made an error in your Predval transformation.

Before testing the Predval and PAR variables, we will create one further variable that can be used in stratifying our testing. We will recode our existing neighbourhood variable into a new numeric variable (the existing neighbourhood code variable in this database is named "hnbhd" and has values of C01, C02, C03…to C10. These are in a text format and as such cannot be used in many of the analysis tools within SPSS. We will simply transform the C## values to the corresponding numbers 1 through 10 and call the new variable Neigh_num).

Add the following transformation to the syntax file and execute it.

```
RECODE hnbhd ('C01'=1) ('C02'=2) ('C03'=3) ('C04'=4) ('C05'=5) ('C06'=6) ('C07'=7)
('C08'=8) ('C09'=9) ('C10'=10) INTO Neigh_num.
VARIABLE LABELS Neigh_num 'Neigh'.
EXECUTE.
```

We will now proceed to test the Predval and PAR variables using a variety of tools including the Kruskal-Wallis test and Ratio Statistics.

Model Testing: Model Database

Ratio statistics will provide an in-depth look at our results. If we have done a good job of predicting the selling price of the houses in our Model database, the mean and median PAR should be close to 1.000 – in other words, generally predicted values are equal to sale prices.

We will also examine the PARs in each neighbourhood, to confirm the PARs are equally distributed. If we find that some neighbourhoods are over- or under-predicted, then we can consider neighbourhood-specific adjustments.
Ratio statistics are calculated as follows:

- Analyze → Descriptive Statistics → Ratio…
- Select Predval as the Numerator, sale_amt as the Denominator, and Neigh_num 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.

Ratio Statistics for Predval / sale_amt

Group	Mean	95% Confidence Interval for Mean		Median	95% Confidence Interval for Median			Min	Max	Range	COD
		Lower Bound	Upper Bound		Lower Bound	Upper Bound	Actual Coverage				
Rivers Bend	.994	.937	1.051	.957	.935	1.030	96.2%	.632	1.563	.931	.131
Winston-Wellington	1.011	.979	1.044	1.002	.936	1.035	97.3%	.812	1.396	.584	.098
Kristal Estates	1.006	.952	1.061	.986	.941	1.111	96.5%	.838	1.181	.343	.076
Sussex East	1.007	.941	1.073	.978	.918	1.011	97.6%	.700	1.500	.800	.121
Sussex West	1.034	.932	1.136	1.004	.890	1.094	97.3%	.719	1.701	.982	.155
Kelly-Juno Area	.981	.937	1.025	.998	.943	1.050	97.6%	.553	1.177	.625	.093
Kingston Area	1.012	.955	1.069	.989	.962	1.021	97.6%	.789	1.632	.844	.090
Mohawk Area	1.033	.967	1.099	1.005	.944	1.071	95.6%	.710	1.691	.981	.153
Elizabeth-June Area	1.095	.947	1.244	1.028	.942	1.103	98.0%	.660	2.998	2.339	.208
Television Area	1.080	1.020	1.139	1.039	.963	1.142	96.4%	.771	1.650	.880	.152
Overall	1.027	1.005	1.048	.989	.978	1.011	95.9%	.553	2.998	2.446	.132

The overall statistics for the model database (bottom of table) show a median PAR of 0.989 and a mean PAR of 1.027 (note: Ratio Statistics displays the ratios as decimal numbers instead of percentages, e.g., 0.989 rather than 98.9%). This shows our prediction is quite good, with predicted results very close to actual sales prices (within 1.1% for the median and 2.7% for the mean). As the mean is greater than the median we suspect there may be some high outliers. The coefficient of dispersion (COD) is 13.2%, which is higher than the desired 10%, but in the acceptable 10-20% range.

The report also shows acceptable levels for all neighbourhoods. The lowest median PAR is 0.957 and the highest is 1.039. The means range from 0.994 to 1.095. All of the 95% confidence intervals around the medians contain 1.000, meaning there is less than 5% probability the median PAR is not equal to 100%, or we can be 95% confident that predicted values are statistically equal to sale prices across the neighbourhoods in this database. We can conclude no adjustments are needed for any of the neighbourhoods. Although the COD for Elizabeth-June exceeds 20%. This may be caused, in part by the extreme PAR value of 2.998.

Boxplots could also be used as a visual test, highlighting differences in PAR by Neigh_num. We can also confirm our conclusion with a Kruskal-Wallis test:

- Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples…
- Enter PAR as the Test Variable and Neigh_num as the Grouping Variable.
- Click Define Range, enter 1 for Minimum and 10 for Maximum.
- Continue → OK.

The output should appear as below.

Kruskal-Wallis: OntarioModel

Ranks

	Neigh	N	Mean Rank
PAR	1.00	40	155.80
	2.00	60	173.40
	3.00	15	177.53
	4.00	29	159.79
	5.00	21	168.43
	6.00	34	166.74
	7.00	29	172.45
	8.00	42	173.74
	9.00	32	182.34
	10.00	45	202.36
	Total	347	

Test Statistics[a,b]

	PAR
Chi-Square	5.986
df	9
Asymp. Sig.	.741

a. Kruskal Wallis Test
b. Grouping Variable: Neigh

The mean ranks here should be approximately 174 (half of 347). This report shows excellent results, with seven of the neighbourhoods close to the target value. Neighbourhoods 1 and 4 are slightly low, while neighbourhood 10 is high. The chi-square test, with the Asymp. Sig. at 74.1%, tells us that we cannot reject the hypothesis that all neighbourhoods have had their predicted selling prices equally modeled. This is a strong result.

Neighbourhood Adjustment

This model did not require a neighbourhood adjustment, but if it did, the following explains how to do so. The median PAR for River's Bend (neighbourhood 1) is 95.7%, which means on average it is over 4% undervalued. To adjust for this we would apply a factor of 1 ÷ 0.957, or 1.045, to the predicted selling prices of all properties in neighbourhood 1. This would increase their value and raise the median PAR for River's Bend to 100%. The syntax to do this would be as follows:

```
COMPUTE Nbhd1Fac=Neigh_num.
RECODE Nbhd1Fac(1=1.045)( ELSE=1).
COMPUTE NewPredVal=Predval * nbhd1fac.
```

Ratio statistics would then be re-run with the new predicted selling price, NewPredVal, to confirm its results were improved. If more than one neighbourhood needed to be adjusted, the set of transformations above would be run for each such neighbourhood.

Model Testing: Test Database

Now that we are satisfied with the model, we will also test it in the OntarioTest database, seeing how it performs using data not involved in the model building process. This is a better test of the model, since the OntarioModel database could potentially have problems with "chasing the sales".

Before we move on to the test database, first ensure you save your OntarioModel data file!

Open OntarioTest.sav. Go to your syntax file and block and run the last three transformations, beginning with the calculation of Predval and ending with the recode of hnbhd. The variables in OntarioTest should now match the variables in OntarioModel. Run the Ratio Statistics for the entire database. Because there are very few observations in some of the neighbourhoods, no neighbourhood analysis will be done for this database.

Ratio Statistics: OntarioTest

Mean		.985
95% Confidence Interval for Mean	Lower Bound	.963
	Upper Bound	1.008
Median		.962
95% Confidence Interval for Median	Lower Bound	.936
	Upper Bound	1.003
	Actual Coverage	96.0%
Minimum		.704
Maximum		1.581
Range		.877
Coefficient of Dispersion		.118

The results look good. The median PAR of 0.962 is slightly below 1.00, but the confidence interval includes 1.000. The mean PAR is 0.985 (98.5%) with a coefficient of dispersion equal to 0.118 (11.8%).

To more precisely quantify the distribution of the ratios, we will use Frequencies in Descriptive Statistics to calculate quartiles and cutpoints 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 Quartiles and Cut Points (10).
- Continue → OK

Frequency Statistics: PAR variable in OntarioTest

N	Valid	174
	Missing	0
Percentiles	10	.826
	20	.856
	25	.876
	30	.897
	40	.931
	50	.962
	60	1.017
	70	1.048
	75	1.071
	80	1.087
	90	1.145

Based on the first quartile (25th percentile) and third quartile (75th percentile), 50% of the observations have a PAR between 0.876 and 1.071. In addition, 80% of the observations fall between 0.826 and 1.145.

These results indicate that the model created in the OntarioModel database provides accurate estimates for selling prices in another database of sale properties not used to create the model. This implies the model would create good selling price estimates when applied to all of the properties in this region of Ontario.

The next phase of the testing requires the maximum number of properties available or many of the tests can produce invalid results. Therefore, we will evaluate the model's performance using all 521 sales in the Ontario521complete.sav database.

Model Testing: Complete Database

Once again, ensure you save the OntarioTest.sav database before proceeding!

Open the Ontario521complete.sav database. Run the syntax commands to create the Predval, PAR, and Neigh_num variables (as was done for the OntarioTest database). We will now run a selection of statistical tests to evaluate the performance of the model at predicting selling prices for single family dwelling in this region of Southern Ontario.

First off, we will run Ratio Statistics for the overall database.

Ratio Statistics for Predval / Sale Amount

Mean		1.013
95% Confidence Interval for Mean	Lower Bound	.996
	Upper Bound	1.029
Median		.986
95% Confidence Interval for Median	Lower Bound	.965
	Upper Bound	1.000
	Actual Coverage	95.6%
Minimum		.553
Maximum		2.998
Range		2.446
Coefficient of Dispersion		.128

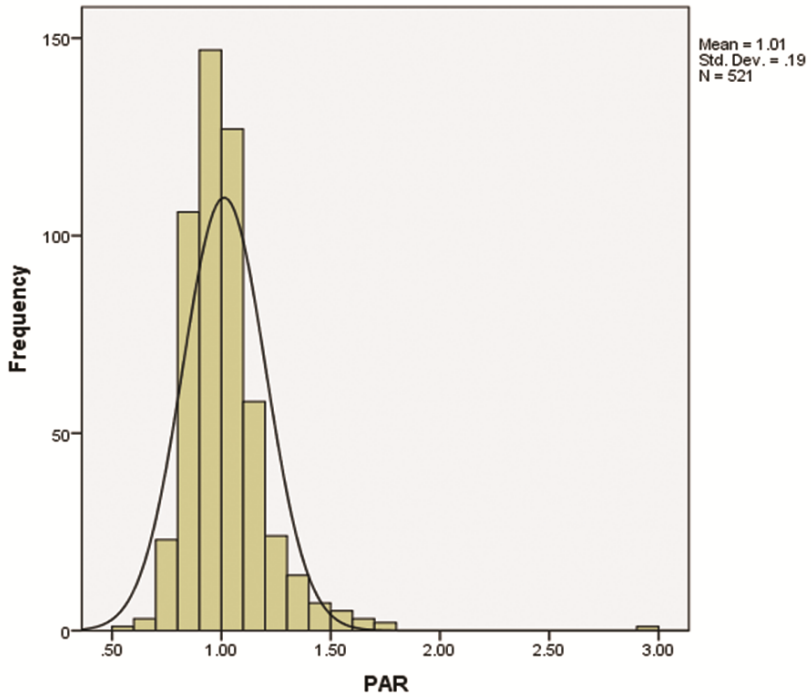
The model is performing reasonably well. The median PAR is 0.986 and the mean is 1.013.

Our previous results indicated there may be some high outliers affecting results. We can use Frequencies to investigate. As before, we will calculate quartiles and cut points for 10 equal groups for the PAR variable.

N	Valid	521
	Missing	0
Mean		1.013
Median		.986
Std. Deviation		.190
Minimum		.553
Maximum		2.998
Percentiles	10	.832
	20	.880
	25	.897
	30	.918
	40	.949
	50	.986
	60	1.020
	70	1.066
	75	1.085
	80	1.109
	90	1.220

Based on the first quartile (25th percentile) and third quartile (75th percentile), 50% of the observations have a PAR between 0.897 and 1.085, or roughly within 10% of the median value. In addition, 80% of the observations fall between 0.832 and 1.220, or within 22% of the median. These measures indicate that the PARs are quite tightly bunched in the centre, but with a significant number of high outliers. This distribution appears to be slightly skewed on the high side (to the right). This can be visually confirmed using a histogram.

Re-open the Frequencies window and click Charts. Under Chart Type, select Histogram and Normal Curve.



The chart shows that indeed there are more values of PAR near the 1.50 mark than there are near the 0.50 mark. Also, there is one extreme outlier near 3.00. The one outlier near 3.00 has a selling price of only $20,000 with a predicted value of almost $60,000. This sale should be examined more closely to verify the price and the conditions surrounding the sale. One course of action at this point would be to eliminate the sale from the database and rebuild the model.

Alternatively, we could perform an action called "trimming" for our tests. We would do this by setting a filter to identify PAR > 1.58 OR PAR < 0.44 or, if using the more conservative two standard deviations, PAR > 1.39 OR PAR < 0.63. If you use Analyze → Descriptive Statistics → Frequencies and select Display frequency tables, you will find 19 observations of PAR which are greater than 1.39 and one that is less than 0.63; there are six greater than 1.58 and none less than 0.44.

If the distribution were normal, we would expect to have 2.5% of the observations (or 13 sales) fall more than two standard deviations from the mean in either direction. Since there are two low outliers and 19 high outliers, this also indicates that PAR is not normally distributed. Because the distribution is somewhat more concentrated in the centre, and the truncation on the left is compensated by a longer tail on the right, the ratios will be somewhat more concentrated than the above normal probability rules would suggest.

The coefficient of dispersion (COD) is 12.8%. For this kind of study a COD of less than 15% is a good result; less than 10% would be an excellent result.

Neighbourhood Analysis

To test if neighbourhoods are evenly valued in the complete database, first we will run the Kruskal-Wallis test on the PAR variable across the neighbourhood numbers.

- Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples...
- Enter PAR as the Test Variable and Neigh_num as the Grouping Variable
- Click Define Range, enter 1 for Minimum and 10 for Maximum
- Continue → OK.

Ranks

	Neigh	N	Mean Rank
PAR	1.00	65	231.86
	2.00	87	264.68
	3.00	19	269.26
	4.00	39	235.41
	5.00	31	274.13
	6.00	54	260.80
	7.00	40	272.13
	8.00	61	268.26
	9.00	52	225.52
	10.00	73	301.77
	Total	521	

Test Statistics[a,b]

	PAR
Chi-Square	12.509
df	9
Asymp. Sig.	.186

a. Kruskal Wallis Test
b. Grouping Variable: Neigh

The mean ranks here should be approximately 261 (half of 521). Six neighbourhoods are very close to the target value. Neighbourhoods 1, 4, and 9 are low, while neighbourhood 10 is high. The chi-square test (at 18.6%) shows we cannot reject the hypothesis that all neighbourhoods are equally modeled, but we will continue with a ratio statistics test to ensure that the neighbourhoods are equally modeled. We will use Predval/Sale_Amt (or PAR) as the ratio and set hnbhd as the Group Variable.

Ratio Statistics for Predval/Sale Amount

		95% Confidence Interval for Mean			95% Confidence Interval for Median						
Group	Mean	Lower Bound	Upper Bound	Median	Lower Bound	Upper Bound	Actual Coverage	Min	Max	Range	COD
Rivers Bend	.986	.943	1.030	.949	.906	.998	95.4%	.632	1.569	.937	.132
Winston-Wellington	1.002	.975	1.028	1.002	.946	1.028	96.9%	.768	1.396	.628	.099
Kristal Estates	.997	.950	1.045	.986	.941	1.076	98.1%	.826	1.181	.355	.077
Sussex East	.990	.939	1.041	.949	.917	.989	97.6%	.700	1.500	.800	.111
Sussex West	1.035	.959	1.111	1.011	.891	1.094	97.1%	.719	1.701	.982	.149
Kelly-Juno Area	.982	.950	1.014	.990	.959	1.032	96.0%	.553	1.177	.625	.088
Kingston Area	1.010	.967	1.053	.997	.973	1.035	96.2%	.789	1.632	.844	.087
Mohawk Area	1.025	.976	1.074	.996	.944	1.036	96.0%	.710	1.691	.981	.141
Elizabeth-June Area	1.024	.926	1.122	.942	.877	1.022	96.4%	.660	2.998	2.339	.196
Television Area	1.062	1.017	1.107	1.039	.972	1.084	96.6%	.704	1.650	.947	.144
Overall	1.013	.996	1.029	.986	.965	1.000	95.6%	.553	2.998	2.446	.128

The CODs show low dispersion in the ratio of predicted to actual sale amounts (PAR) in neighbourhoods Winston-Wellington, Kristal Estates, Kelly-Juno, and Kingston (CODs range from 7.7% to 9.9%) and higher in the other six neighbourhoods (CODs ranging from 11.1% to 19.6%). The very high COD in Elizabeth-June is caused by one outlier of 2.998. If this sale is filtered out, the COD drops to 15.7%.

We see that three neighbourhoods are somewhat under-valued, Rivers Bend, Sussex East, and Elizabeth-June all have medians that are less than 0.950. As well, two neighbourhoods (Rivers Bend and Sussex East) do not have 1.000 contained within in their 95% confidence interval around the medians.

At this point, because of the Ratio Statistics result we could consider adjusting the model to fix the problem of a slight under-valuation in Rivers Bend, Sussex East, and Elizabeth-June. However, we will leave these adjustment for now.

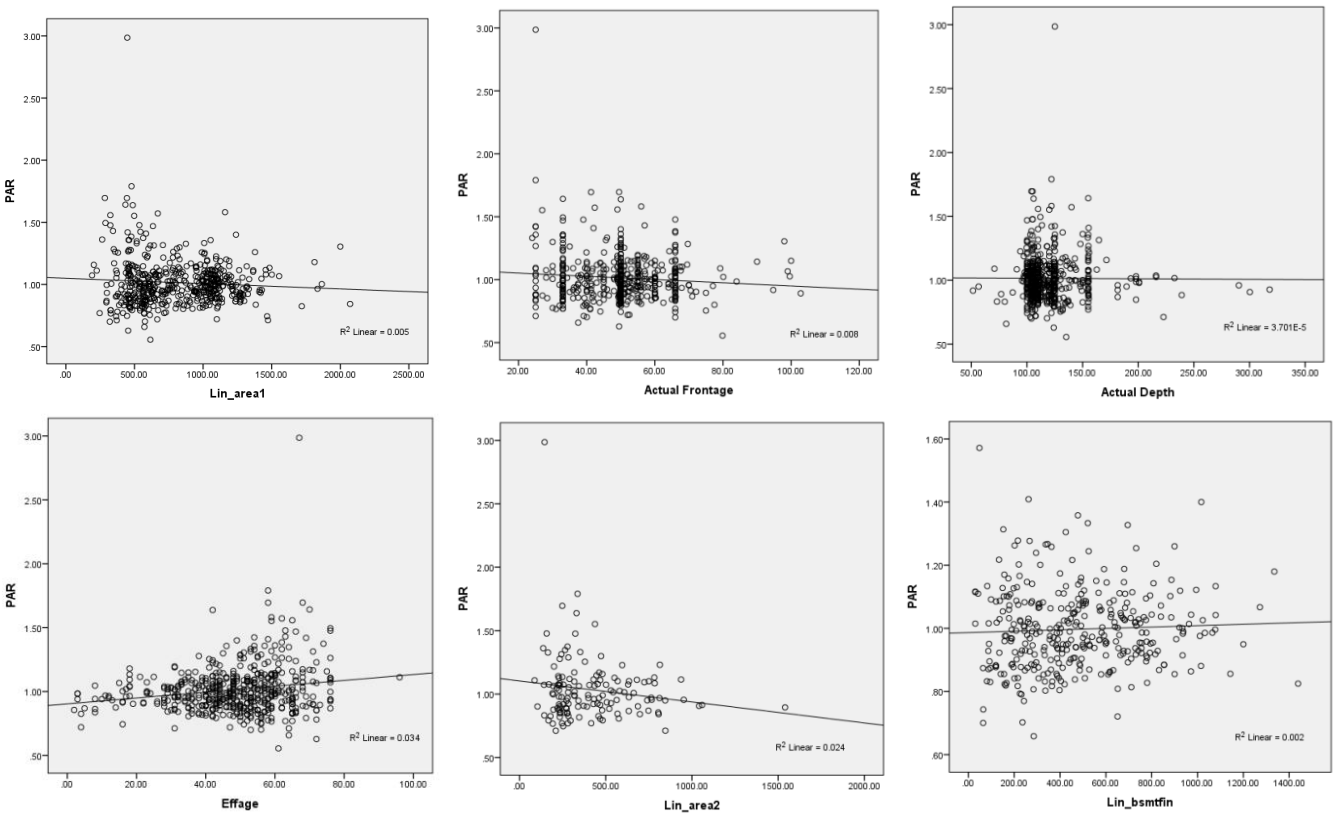
Examination of Property Characteristics

Ideally we want to find no relationship between any particular characteristic and the PARs. If we found a relationship, this implies that the PARs vary according to that characteristic. For example, a positive relationship between PAR and frontage would indicate that as frontage increases, the predicted selling price to actual selling price ratio increases. Therefore, the predicted selling prices of properties with wide frontage would be much higher than selling prices for properties with narrower frontage. Put another way, the properties with large frontage are over-valued by more than those with smaller frontages.

Any significant relationship between the PAR and a property characteristic indicates that selling prices developed by the model are biased.

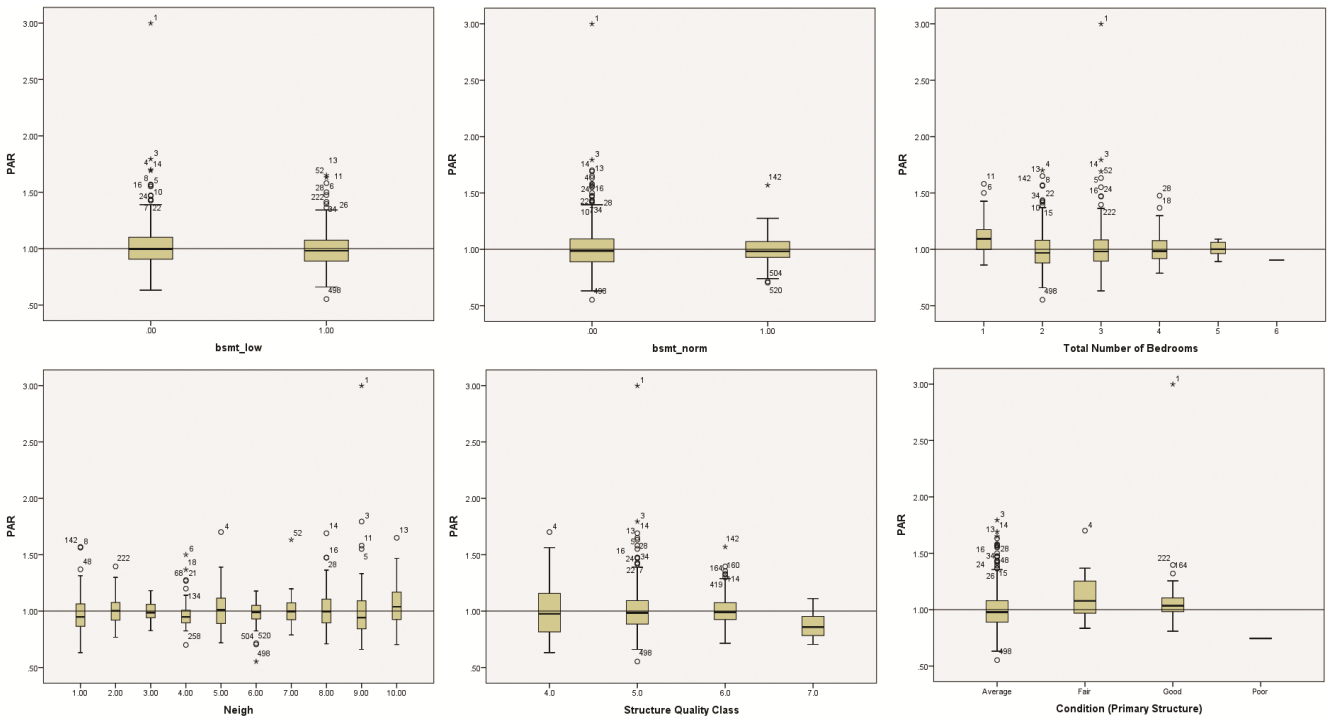
Data displays are very useful for determining whether or not relationships exist between PARs and property characteristics. We will use scatter diagrams and boxplots to look for relationships. Scatter diagrams are useful for comparing PARs with continuous data, while boxplots are useful for comparing PARs with discrete data.

Consider scatter diagrams for Lin_area1, frontage, depth, Effage, Lin_area2, and Lin_bsmtfin. For the latter two, we eliminated sales with a zero value.



These six charts show there is little to no relationship between these variables and PAR.

For boxplots, we will examine bsmt_low, bsmt_norm, bedrooms, Neigh_num, quality, and condition. We added a reference line at 1.000.



The boxplots show our model is producing fairly consistent results. The model does a good job of predicting the selling price whether the house has a basement with low or normal ceiling height. The one bedroom houses appear to produce a slightly higher predicted selling price; nothing can be said about six bedroom homes as there is only one in our database. The neighbourhoods appear equitable. Higher quality homes (quality 7) produce lower predicted values than their selling prices. Finally, condition appears to have no effect on the predicted selling price (and there is only one occurrence of Poor).

Correlation and multiple regression can also be used to check for relationships between PAR and various property characteristics – essentially we want no relationships to appear. Go to Analyze → Correlate → Bivariate and produce a correlation matrix of the following variables: PAR, Lin_areatot, lotsize, fireplcs, Effage, porchpts and sty_full.

Correlations

	PAR	Lin_areatot	lotsize	fireplcs	Effage	porchpts	sty_full
PAR	1	-.082	-.078	-.078	.192**	.018	-.008
Lin_areatot	-.082	1	.241**	.413**	-.564**	-.070	.387**
lotsize	-.078	.241**	1	.079	-.205**	-.101*	-.081
fireplcs	-.078	.413**	.079	1	-.296**	.005	.143**
Effage	.192**	-.564**	-.205**	-.296**	1	.276**	.026
porchpts	.018	-.070	-.101*	.005	.276**	1	.252**
sty_full	-.008	.387**	-.081	.143**	.026	.252**	1

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

The variable most strongly correlated with PAR is Effage but the correlation is quite low. This is a good result.

Go to Analyze → Regression → Linear and run a regression with the dependent variable PAR and the independent variables as above: Lin_areatot, lotsize, fireplcs, Effage, porchpts, and sty_full. Use the Enter method so that the impact of all of these variables can be observed (ensure you do not use the Stepwise method, as this will exclude some variables from the analysis). The following output results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.209[a]	.044	.033	.18644

a. Predictors: (Constant), Total Number of Full Storeys, Effage, Lot Size, Total Number of Fireplaces, Total Porch Points, Lin_areatot

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.816	6	.136	3.913	.001[b]
	Residual	17.866	514	.035		
	Total	18.682	520			

a. Dependent Variable: PAR
b. Predictors: (Constant), Total Number of Full Storeys, Effage, Lot Size, Total Number of Fireplaces, Total Porch Points, Lin_areatot

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.911	.058		15.799	.000
	Lin_areatot	4.548E-005	.000	.089	1.391	.165
	lotsize	-5.715E-006	.000	-.056	-1.235	.217
	fireplcs	-.015	.020	-.037	-.770	.442
	Effage	.003	.001	.230	4.056	.000
	porchpts	-.001	.001	-.035	-.750	.454
	sty_full	-.025	.033	-.039	-.757	.449

a. Dependent Variable: PAR

Effage is the only variable with a high Beta value, showing it has the most impact in explaining variations in PAR. Beta values indicate the relative importance of the variables in explaining variations in the dependent variable. The variables with the highest beta values are most important in the model.

However, the R^2 shows these variables in total only explain 4.4% of the variation in PAR. The summary of a stepwise regression using the same variables is included below so that you can see the effect of adding each variable in the R^2 statistics (use a PIN of 0.40 and a POUT of 0.45). These results confirm that Effage has the only noticeable relationship with PAR (R^2 is 0.037 with Effage and increases to just 0.042 with the three additional variables). This relationship is negligible and can be ignored. If there was a significant relationship between PAR and one of the variables included in this type of testing, we would have to look more closely at the descriptive and ratio statistics for that variable and determine if an adjustment was necessary. As well, if that variable had been excluded from the model, we would have to re-evaluate why we took that particular variable out of the model. Fortunately we do not have to do that here.

Stepwise Regression Analysis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.192[a]	.037	.035	.18621
2	.196[b]	.038	.035	.18624
3	.199[c]	.040	.034	.18628
4	.204[d]	.042	.034	.18628

a. Predictors: (Constant), Effage
b. Predictors: (Constant), Effage, Lot Size
c. Predictors: (Constant), Effage, Lot Size, Total Porch Points
d. Predictors: (Constant), Effage, Lot Size, Total Porch Points, Lin_areatot

The final tool that we will use to test model performance is the Mann-Whitney or M-W test. This test, like the Kruskal-Wallis test, is found in the Nonparametric Tests menu of SPSS. The Mann-Whitney test is a specialized form of the K-W test. While the K-W test provides an objective test of the valuation among all values of a discrete variable, the M-W tests the level of valuation between the two aspects of a binary variable. Prior to the completion of any modeling project, the M-W test should be conducted on all suitable (e.g., binary) variables to ensure there are no obvious modeling errors.

The SPSS steps for the Mann-Whitney test are as follows:

- Analyze → Nonparametric Tests → Legacy Dialogs → 2 Independent Samples
- Enter PAR as the Test Variable and ab_comm as the Grouping Variable
- Click Define Groups and enter 0 and 1.
- Continue → OK to display the following report:

Mann-Whitney: PAR by ab_comm

	Abuts Commerical	N	Mean Rank	Sum of Ranks
PAR	No	504	261.40	131748.00
	Yes	17	249.00	4233.00
	Total	521		

Test Statistics[a]

	PAR
Mann-Whitney U	4080.000
Wilcoxon W	4233.000
Z	-.334
Asymp. Sig. (2-tailed)	.738

a. Grouping Variable: Abuts Commerical

The Asymp. Sig. shows the probability that homes that abut commercial uses are valued in our model at the same level as those that do not. The probability of 0.738 is greater than our critical value of 0.05. This means at a 95% confidence level, we can conclude properties abutting and not abutting commercial uses are appraised equally in our model.[4] You can also confirm this result visually by viewing a boxplot of PAR by ab_comm.

Let's use the Mann-Whitney test to examine properties on corners. Click Dialog Recall to return to the Mann-Whitney input panel and change the Grouping Variable to corner (don't forget to define groups, which will be 0 and 1, as in the last test). We are testing whether properties on corners and those not on corners are valued at the same percentages of the actual selling prices in our model. The results are as follows:

Mann-Whitney: PAR by corner

	Corner Lot	N	Mean Rank	Sum of Ranks
PAR	No	486	256.81	124810.00
	Yes	35	319.17	11171.00
	Total	521		

Test Statistics[a]

	PAR
Mann-Whitney U	6469.000
Wilcoxon W	124810.000
Z	-2.367
Asymp. Sig. (2-tailed)	.018

a. Grouping Variable: Corner Lot

Here the probability (Asymp. Sig.) is 0.018 which is less than .05. Thus, at 95% confidence, we can conclude that properties on corners and those not on corners are not valued at the same level.[5] You can confirm this

[4] The Mann-Whitney tests a null hypothesis, e.g., properties with an attribute are valued at the same level as those without the attribute. If the test fails, then the null hypothesis is rejected and the alternative hypothesis accepted, e.g., properties with an attribute are valued at a different level than those without. The M-W test uses Z-values to calculate a probability estimate. For a two-tailed test with a 95% confidence interval, the critical value of Z is 1.96 and the critical probability limit is .05. A Z value in the interval ± 1.96, or a probability of greater than .05, indicates that the null hypothesis cannot be rejected (i.e., conclude they are valued the same). A Z value outside the interval ± 1.96, or a probability of less than .05, indicates the null hypothesis should be rejected (i.e., conclude they are valued differently). For ab_comm, the Z value is -0.334 which is within the range of ± 1.96. The associated probability is 0.738, which is greater than .05. We conclude the null hypothesis cannot be rejected; the properties with and without commercial influence are valued equally. For more background on Mann-Whitney tests, see *Advanced Computer-Assisted Mass Appraisal*, page 13.5.

[5] The null hypothesis is that properties on corners and those not on corners are valued in our model at equal percentages of the selling prices. The alternative hypothesis is that properties on corners and those not on corners are not valued at the same level. The calculated Z-value is -2.367, which is outside the range of ± 1.96. The prob. is 0.018 which is less than .05. Thus, at 95% confidence, we can reject the null hypothesis and instead accept the alternative hypothesis: these properties are not valued equally. Note here that there is no test for the alternative hypothesis. The test is always of the null hypothesis, which is rejected or not. If the null is rejected, the alternative is automatically accepted.

visually by viewing a boxplot of PAR by corner. This characteristic may justify an adjustment being made to the model. However, we will not carry out this adjustment here.

Adjusting a Model for a Specific Property Characteristic

In our model, we did not carry out an adjustment for corner, even though the results indicate one may be needed. Running a Ratio Statistics for PAR with corner as Group Variable shows that the median for corner = No is 0.982 and corner = Yes is 1.040. The 95% confidence interval (C.I.) for the median when corner equals 0 does not include 1, meaning that an adjustment is justified (for corner = 1, the C.I. does include 1.00).

Ratio Statistics for Predval/Sale Amount

		95% Confidence Interval for Mean			95% Confidence Interval for Median						
Group	Mean	Lower Bound	Upper Bound	Median	Lower Bound	Upper Bound	Actual Coverage	Min	Max	Range	COD
No	1.009	.992	1.026	.982	.962	.998	95.9%	.553	2.998	2.446	.127
Yes	1.070	1.011	1.128	1.040	.981	1.090	95.9%	.819	1.473	.653	.118
Overall	1.013	.996	1.029	.986	.965	1.000	95.6%	.553	2.998	2.446	.128

To carry out the adjustment for this model you would do the following:

1. calculate the corner adjustment factor: 1 ÷ 0.982 = 1.018
2. enter and run the following transformations:

```
COMPUTE Corner0Fac=corner.
RECODE Corner0Fac(0=1.018)( ELSE=1).          [creates a factor for properties with corner = 0]
COMPUTE NewPredVal=Predval * Corner0Fac.      [applies the adjustment factor to the PredVal to
                                               create NewPredVal variable]
COMPUTE NewPAR = NewPredVal / sale_amt.       [re-calculates a corner-adjusted PAR]
```

3. Re-run the Ratio Statistics to confirm the results are improved.

Ratio Statistics for NewPredval/Sale Amount

		95% Confidence Interval for Mean			95% Confidence Interval for Median						
Group	Mean	Lower Bound	Upper Bound	Median	Lower Bound	Upper Bound	Actual Coverage	Min	Max	Range	COD
No	1.027	1.010	1.044	1.000	.980	1.016	95.9%	.563	3.052	2.490	.127
Yes	1.070	1.011	1.128	1.040	.981	1.090	95.9%	.819	1.473	.653	.118
Overall	1.030	1.013	1.046	1.003	.981	1.018	95.6%	.563	3.052	2.490	.127

The median for corner = No (or 0) is now 1.000 and the 95% Confidence Interval around the median brackets 1.000.

If other property characteristics needed adjusting, then a set of similar transformations to those above would be run for each such characteristic.

When testing an actual model, you should examine all binary variables with the Mann-Whitney test and make any adjustments necessary. This will ensure that your model does not have any systematic under- or over-valuations for property attributes. This is an important requirement of appraisal uniformity. Consider if this model was for property tax assessments – the valuation must be seen as "accurate and equitable" by taxpayers. It is obviously important that people who live on corners, have waterfront, sloped lots, or any other attribute, can see they are assessed on an equivalent basis to everyone else and thus pay only their "fair share" of property taxes and no more! As another example, assume this model was instead forming the basis for a commercial automated valuation model (AVM), used by lenders underwriting mortgage loans. The lenders' risk

minimization demands valuations be as accurate as possible. If a lender lost considerable money on a foreclosed corner property and subsequently discovered corner lots were systematically under-valued by the AVM, then the modeller would probably face some difficult questions.

STEP 9: State Conclusions on Model Quality

Based on the statistical results, we seem to have a reasonably good model for predicting sale prices of single family dwellings in this market area in southern Ontario. The model could possibly be improved by making adjustments for Rivers Bend, Sussex East, and Elizabeth-June and for corner properties. As well, to be comprehensive, we must test for all characteristics in order to confirm the model is valuing all properties uniformly. Once we are convinced the model is as accurate and uniform as possible, then the model can be applied to all of the properties in the market area, estimating the market value of properties that did not sell in 2008.

Summary: Comprehensive Model Building

This lesson advanced on the model building foundations introduced in Lessons 6 and 7. We added data screening and model testing to the basic steps shown earlier. In this lesson, we:

- described a general additive model for a database of single family dwellings;
- screened the data to ensure it is optimal for modelling;
- 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;
- tested the performance of the model using a number of statistical tools; and
- stated conclusions about the quality of the model.

Using our Ontario example, we illustrated all nine of the steps in model building, showing a comprehensive model building example. This completes our coverage of model building techniques.

Keep in mind that model building is an intense statistical exercise and we have only brushed the surface of what is possible. If you want more depth on comprehensive modelling and the possibilities beyond our relatively simple examples, you may be interested in the BUSI 444 course, *Advanced Computer-Assisted Mass Appraisal*. As well, Lesson 9 will further examine common applications of modelling in valuation work in looking at automated valuation models (AVMs).

Appendix 8.1: Additional Kruskal-Wallis Tests

Two additional Kruskal-Wallis tests are illustrated here. Note that these are not relevant to the model presented in this lesson. They are provided here only as additional examples to better illustrate how to use Kruskal-Wallis tests. The first additional Kruskal-Wallis test will examine SAR (sale to assessment ratio) by a neighbourhood variable.

Kruskal-Wallis Test

	NBHD	N	Mean Rank
SAR	1	78	274.67
	2	53	259.51
	3	39	235.67
	4	23	203.91
	5	83	194.91
	6	107	204.17
	7	70	224.97
	Total	453	

Test Statistics [a, b]

	SAR
Chi-Square	22.752
df	6
Asymp. sig.	.001

a Kruskal Wallis Test
b Grouping Variable: NBHD

The significance (Asymp. Sig.) shows that there is almost zero probability that the NBHDs are equally valued (0.1% to be exact). The expected value of mean ranks should be in the centre of the distribution, or 453 divided by 2 = 226. The mean ranks above show only NBHD 7 having a value close to the expected value of 226. NBHDs 1 to 3 have values higher than expected, indicating possible undervaluation (recall that this is a sale to assessment ratio). NBHDs 4 to 6 have lower values indicating possible overvaluation. Because of the large number of sales in NBHDs 1, 5, and 6, the mean ranks are likely a true indicator of the problem. This is because with a larger number of sales, the confidence interval will be very narrow. This means that if the expected value does not fall within the confidence interval, you can be at least 95% confident that the test result is statistically not equal to the expected value. Contrast this to NBHDs 3 and 4 which have few sales – the results above will not be as reliable because the confidence interval is likely to be wide and may include the expected value. There is enough variation in the mean ranks shown above to indicate that the confidence interval of the median will need to be checked and the ratios possibly adjusted. This review requirement is also shown by the value of the significance below the acceptable level of .05.

The second test will examine a land adjustment factor across neighbourhoods.

Kruskal-Wallis Test

	NBHD	N	Mean Rank
LFACTOR	1	78	201.74
	2	53	161.36
	3	39	256.13
	4	23	207.09
	5	83	238.57
	6	107	245.37
	7	70	253.36
	Total	453	

Test Statistics [a, b]

	LFACTOR
Chi-Square	24.284
df	6
Asymp. sig.	.000

a Kruskal Wallis Test
b Grouping Variable: NBHD

Again, here the significance shows that at least some of the NBHDs are not equally valued. The mean ranks range from a very low value for NBHD 2 to a high value for NBHD 3. This difference is likely great enough to cause the need to review the values in more detail using Ratio Statistics. This would be recommended even if the significance had shown a pass value over .05. Large differences in the mean ranks between two of the groups may indicate that there is a problem that needs correction, even when the chi-square test shows a passing value.

What is emphasized here is the use of the significance (Asymp. Sig.) as the main determining factor. The use of the mean ranks is only an indication of potential problems and cannot be used as the main analysis tool. However, large variations in the mean ranks should indicate the need for further examination of the data with more accurate tests that include the use of confidence intervals, such as Ratio Statistics.

Review and Discussion Questions

1. You are beginning development of a regression model to predict street level retail rents for several prestige retail districts in the City of Vancouver (e.g., South Granville, Robson Street). You have purchased data from BC Assessment and have begun migrating the data from an Excel format into SPSS. What is your next step?

2. Eliminating outliers is an initial data screening activity. What type of problems can you encounter with elimination of these unexplainable data occurrences?

3. In the time adjustment discussion earlier in this lesson, we transformed sale price into sale price per total square foot (finished area square feet and lot size square feet combined). Try experimenting with other units of comparison to see if there is any correlation between sale date and these new variables. How might you decide which units of measure to attempt since there are so many different possibilities?

4. Prior to building a model to predict property value based on a group of property sales, time adjusting those sales may be important. What time adjustment is likely required for sales in a market that is (a) fairly constant, (b) rising, or (c) declining? What tools are most effective to determine the need for a time adjustment?

5. In a modelling exercise, you are attempting to determine if a "sea-glimpse" view is significant in terms of necessitating an adjustment in the model. You have completed a Kruskal-Wallis test and found the following results:

    1. N = 200
    2. Mean Rank = 168 for view = 1; Mean Rank = 85 for view = 0
    3. Chi-Square = 6.2
    4. Asymp. Sig. = 0.04

    What is the expected mean rank? What can you conclude from these results?

6. One of the problems in building a regression model for real estate valuation is dealing with variables that have a multiplicative relationship with the dependent variable, sales price. For example, market research may indicate that the Exterior Maintenance of a house will impact the building value according to a percentage scale, for example, -15% for deferred maintenance. How would you account for this factor in an additive regression model?

7. You are building a regression model to predict the market value of single family residential building lots in a neighbourhood experiencing rapid growth. You need to determine which data characteristic to use for describing lot size. Some possibilities are a standard lot with an adjustment factor for size, lot area in square feet, lot width and depth, and lot area in acres. What should you consider in making this choice?

8. You are building a model to predict market rents for retail properties. If you have 20 variables in your regression model, what is the minimum number of rental transactions you need in your dataset?

9. What information does the SPSS Casewise Diagnostics report provide? What action should you take if data occurrences are noted in the report?

10. An analyst has completed a valuation model for real estate prices and is now recommending applying the model to assess all houses in the city for property tax purposes. However, the Assessor first wants to test the ability of the model to reliably predict outcomes for real estate prices. What do you recommend?

11. Why is it necessary to examine the relationship between PARs and various data variables during the model testing process? If we use regression to analyze PAR in relation to the data variables, do we want the R^2 to be high or low?

12. Assume you have just completed a regression analysis to predict the improvement value of housing in a Canadian city. Your client is an insurance company, who will rely on the outcome for loan underwriting purposes. How would you document and explain the quality of the model to your client?

13. What is the easiest way to obtain the predicted values of a regression model for all sales in your database? What is the best method?

14. You have completed a mass appraisal model building exercise and are now in testing mode. You have found the following results. What factor should be applied to the predicted selling prices created by the model in order to bring the estimated sale prices in line with the target ratio of 1.000?

    (a) The median PAR for a neighbourhood is 1.087 with a 95% confidence interval of 1.035 to 1.111.
    (b) The median PAR for a neighbourhood is 1.017 with a 95% confidence interval of 0.975 to 1.111.

ASSIGNMENT 8

LESSON 8: Comprehensive Model Building – Data Screening and Testing

Marks: 1 mark per question.

1. Which of the following are statistical tools that can be used to examine variables in a database?

   A. Crosstabulation tables
   B. Scatterplots
   C. Chi-square test
   D. Simple linear regression

   (1) A and B only
   (2) A and D only
   (3) B, C, and D only
   (4) All of the above

2. Repeat Sales:

   (1) are useful for analyzing time trends.
   (2) are useful in multiple regression analysis.
   (3) are not useful for doing any analyses, and should be removed from the dataset.
   (4) None of the above.

3. Mass appraisal models should be fully tested before they are used in a real application. All of the following statements about testing mass appraisal models are true EXCEPT:

   (1) Testing with sale observations used in the model can give a representation that the results are better than they appear to be.
   (2) Testing with sale observations not used in the model can ensure that generalizing of the results outside of the sample database will produce accurate results.
   (3) "Chasing the sales" refers to withholding a portion of the sales database when calibrating the model and then testing the model against this group of sales.
   (4) "Stratification" refers to separating data into subgroups.

4. In a STEPWISE regression, a useful prediction of market value and variable coefficients can be found by meeting a few criteria. Which of the following criteria for the regression analysis is FALSE?

   (1) Values of t-statistics should be within an interval of ± 1.0.
   (2) Significance levels should be approximately .10 or lower.
   (3) Reduction in the standard error should be low, around 0.1%.
   (4) Confidence intervals should be at least 90%.

Assignment 8 continues on next page

5. Suppose you are in the middle of developing a comprehensive model for the purpose of data screening and testing. Assuming you have just calibrated the model using an appropriate method, which of the following requirements of the modeling process still needs to be completed?

(1) Complete graphical analysis to examine the relationships between variables.
(2) Test and evaluate the model.
(3) Describe an appropriate model.
(4) Create necessary transformations to make variables suitable for the model.

6. Refer back to the boxplots you drew for the relationships between sale amounts and categorical variables. All of the following are true EXCEPT:

(1) "Number of bathrooms" has an impact on sales price.
(2) "Homogeneous neighbourhood" has an impact on sales price.
(3) "Basement height" has little impact on sales price.
(4) "Number of storeys" has little impact on sales price.

7. Which of the following is/are TRUE regarding Beta values shown with the regression using PAR as the dependent variable?

A. A variable with a Beta value of zero explains much of the variation of the dependent variable.
B. A variable with a high Beta value explains a high proportion of the dependent variable's variance.
C. The "total number of full storeys" variable has a negative Beta value, given PAR being the dependent variable.
D. The "effage" variable has a higher Beta value than the "total porch points" variable, given PAR being the dependent variable.

(1) A only
(2) B and C only
(3) B, C, and D only
(4) All of the above

8. When regression modelling, Preliminary Data Screening is important for all but:

(1) finding duplicate sales and retaining only the most recent sale.
(2) finding and removing outliers.
(3) investigating the relationship of continuous variables with the dependent variable.
(4) ensuring the data is appropriate for the task.

9. In regression analysis it is important to state a General Model so that:

(1) it is easier to build the specific model.
(2) the variables in the database will be easier to use.
(3) the size of the constant and coefficients is determined before you start modelling.
(4) the end goal is well defined.

Assignment 8 continues on next page

10. The second step in regression analysis is Reviewing the Variables. The most important part of this step is to ensure:

    (1) that all variables in the database either fit into the General Model or are Information Only.
    (2) all variables in the database are identified for inclusion in the model.
    (3) variables needing transformation are identified.
    (4) outliers in the data are identified and their records removed from the analysis.

11. Lexi is building a regression model on commercial property in a large city in Canada. The market has been reasonably stable up to the last six months when things began to pick up. Her colleague tells her that there is no need for a time adjustment to the sale prices. Lexi runs a K-W test on the Sale Price to Assessed Value Ratio (SAR) by Sale Month and the result is an Asymp. Sig. of 0.050. Lexi should:

    (1) perform the time adjustment using the result of the K-W test as proof that a time adjustment is needed.
    (2) run a Mann-Whitney test on the SAR by month to get another result to analyze.
    (3) run ratio statistics on the SAR by month to get a clearer picture of the time trend.
    (4) use a cross-tabulation test to see if there is a relationship between the Sale Price and the Assessed Value.

12. During the Examine the Variables step it is very important to ensure that:

    (1) only continuous variables with a high correlation to the dependent variable are considered for Step 6 – List Variables for Calibration.
    (2) only discrete variables with significant separation between the boxes in a box plot against the dependent variable are considered for Step 6 – List Variables for Calibration.
    (3) any collinearity among possible independent variables is noted so that certain combinations of variables can be avoided during 6 – List Variables for Calibration.
    (4) All of the above.

13. In the database that Jean is using to create an additive regression model is a variable called "Style". It represents the architectural style of a home. It is numeric with values of 1 through 6. Jean has done some analysis and has found that the lower numbered styles always sell for less than the higher numbered styles. If she uses Style = 4 as the base, the ratios of the median selling prices for each Style relative to the base are 0.76, 0.85, 0.95, 1.00, 1.10, and 1.17. This means that Jean should:

    (1) use the variable Style as it is.
    (2) convert the Style variable to a set of five binary variables with Style = 4 as the control value.
    (3) convert the Style variable from 1 to 6 to the series of ratios and use the Style ratios in the regression.
    (4) Any of the three choices above would be fine.

Assignment 8 continues on next page

14. Lily has transformed a number of variables, some of which are simple binary variable transformations… from a YES/NO to a 1/0. The best method to ensure that such a transformation was successful is to run a:

(1) cross-tabulation table of the old variable and the new variable.
(2) scatter plot between the old variable and the new variable looking for an R^2 of close to or equal to 1.00.
(3) correlation between the old variable and the new variable looking for a correlation of greater than +0.900.
(4) Any of the above will work.

15. Consider the following transformations:

A. A View Direction with values of N, NE, E, SE, S, SW, W, NW into values of 1, 2, 3, 4, 5, 6, 7, 8 respectively.
B. A Quality variable with values of Poor, Fair, Average, Above Average and Excellent into values of 0.75, 0.85, 1.00, 1.05 and 1.15.
C. The sum of the count of bedrooms and the count of bathrooms to yield a new variable called BED-BATH.
D. A PARKING variable which adds up 2 for each multi-car garage, 1 for each single car garage and 1 for each carport.
E. A Garages variable that adds 1 for a multi-car garage, 0.5 for a single car garage, and 0.1 for a carport.

Which transformations would be effective in an additive regression model?

(1) Transformation C only
(2) Transformations A, B, and D
(3) Transformations C and E
(4) All of them

16. Megan is working on an additive regression model with Sale Price as the dependent variable and in her database of 500 records she has, among many other variables, two binary variables called HOT_TUB and MOUNTIAN_VIEW. She has done a box plot for both variables against the dependent variable and finds that there is a slight separation between the two boxes for the HOT_TUB variable, and a moderate separation in the boxes for the MOUNTAIN_VIEW variable. The occurrence count for HOT_TUB is 24 and for MOUNTAIN_VIEW is 16. She is listing the common variables for her set of initial models, she should:

(1) discard both variables as their occurrence counts are less than 25.
(2) keep only HOT_TUB since its occurrence count is so close to 25.
(3) keep only MOUNTAIN_VIEW since it is the only one of the two variables that appears to have an impact on the Sale Price.
(4) keep both variables.

Assignment 8 continues on next page

17. Steven has just completed a multiple regression model for Sale Price. The equation for the final model is:

Predicted Value = 33,255
+ 51.17 × first floor area
+ 31.55 × second floor area
+ 10.11 × basement finished area
+ 9.56 × lot size
+ 3,517.04 × bathrooms
+ 3,500.09 × pool
+ 850.01 × air conditioning
- 3,287.12 × carport

He is quite happy with the model, but notices that the t-statistic for the carport variable is -1.611. He is not quite sure what to do. He should:

(1) remove the carport variable from the list of variables and re-run the model since the t-statistic is negative.
(2) since the carport variable is not significantly adding to the model, remove it from the expression for the Predicted Value, recalculate the Predicted Value and proceed to the Test and Evaluate the Model step.
(3) look back to the List Variables for Calibration step to determine if one or more variables should have been included in the model in place of the carport variable.
(4) do nothing, proceed to the Test and Evaluate the Model step.

18. When testing a discrete variable against the Predicted Value to Actual Sale Price ratio to determine if the ratio is equitably distributed across the values of the discrete variable, the best combination of tests is:

(1) a Kruskal-Wallis test and a Cross-tabulation test.
(2) a Kruskal-Wallis test and a Ratio Statistics test.
(3) a Mann-Whitney test and a Compare Means test.
(4) a Kruskal-Wallis test and a Mann-Whitney test.

19. When testing a continuous variable against the Predicted Value to Actual Sale Price ratio to determine if the ratio is equitably distributed across the values of the continuous variable, the best combination of tests is:

(1) a scatterplot and a correlation matrix.
(2) a boxplot and a correlation matrix.
(3) a Kruskal-Wallis test and a scatter plot.
(4) a Kruskal-Wallis test and a Ratio Statistics test.

Assignment 8 continues on next page

20. When testing for equitable distribution of the Predicted Value to Actual Sale Price ratio across a set of neighbourhood codes a good result for a Ratio Statistics test is:

(1) the Lower Bounds of the 95% Confidence Intervals around the median for all neighbourhoods are greater than 1.00.
(2) the Upper Bounds of the 95% Confidence Intervals around the median for all neighbourhoods are less than 1.00.
(3) all neighbourhoods have median ratios between 0.95 and 1.05.
(4) all of the above.

20 Total Marks

Planning Ahead

Project 2 is based on the analysis illustrated in Lessons 6, 7, and 8. In particular, the project will closely follow the steps in Lesson 8. You are advised to begin work on Project 2 now, while the concepts are fresh in your mind. It is a challenging project and you will be much more satisfied with your results if you give yourself sufficient time to complete it.

End of Assignment 8