Chapter 603

Multiple Regression for Appraisal

Introduction

This procedure is used to create a multiple regression model relating sale price to one or more property attributes based on a number of properties with known sale prices. The form of the model used in this procedure is

$$Sale\ Price = b_0 + b_1 \times Attr_1 + b_2 \times Attr_2 + \cdots + b_p \times Attr_p$$

Both numeric attributes (e.g., square feet, number of bathrooms, age) and categorical attributes (e.g., neighborhood) may be used in the model. The procedure creates binary (0 or 1) terms for categorical attributes. An example of an estimated model might look like

$$Sale\ Price = \$68{,}224 + \$77.51 \times SQFT + \$1.51 \times LOTSIZE - \$838.26 \times AGE + \$14{,}342 \times HERRICK - \$9{,}346 \times SKYGLADE + \$12{,}846 \times POOL$$

In this example, *SQFT*, *LOTSIZE*, and *AGE* are numeric terms, whereas *HERRICK*, *SKYGLADE*, and *POOL* are binary terms where the only possible values are 0 (No) and 1 (Yes).

The estimated model that is produced by running the procedure may be used to estimate the market values of properties for which no sale price is available.

The procedure also provides some options for evaluating whether the assumptions of using the model are met. Some of these include Normality tests and plots as well as multicollinearity diagnostics.

This procedure can be used to estimate the coefficients of a model with a given form, tailored to property value estimation. If you have a large number of attributes and you wish to have the software sift through and determine the best subset of terms, you may wish to instead use one of the subset selection regression procedures such as All Possible Regressions, Stepwise Regression, or Subset Selection in Multiple Regression. For more complex multiple regression models or diagnostics, you might consider the Multiple Regression procedure or the Hybrid Appraisal Models procedure.

Regression Models and Technical Details

If you wish to look into the estimation of regression models, residual diagnostics, regression assumptions, and other technical details in greater detail, you can examine the documentation chapter associated with the Multiple Regression procedure, or you may wish to consult a textbook on the subject of multiple regression. For example, there is a chapter on the subject in *Fundamentals of Mass Appraisal* (Gloudemans and Almy, 2011). A couple of technical aspects that commonly arise in multiple regression for appraisal purposes are mentioned below.

Categorical Attribute Terms

Multiple regression analysis, by its nature, requires that all terms of the model be numeric. Numeric attributes, such as square feet and age, fit nicely into the form of the multiple regression model. On the other hand,

categorical attributes, such as subdivision or property type, require a conversion to numeric terms in order to be used.

To create numeric columns from a categorical column, one of the categories must first be chosen as the reference or baseline category. Then a column is created for each of the other categories (but not the reference category). Each column is made up of ones when the value matches the column category and zeroes otherwise. Thus, the number of columns created is one fewer than the number of categories in the column. These columns are typically called binary columns or binary variables. When a categorical column is used in this procedure, the binary columns are not actually produced in the dataset, but instead are created and used internally. The following example illustrates the process of creating binary variables.

Suppose a property appraiser wishes to include an adjustment to property value based on the neighborhood of the property. In the dataset, the NBHD column appears as follows

NBHD
Cherry Farms
Cherry Farms
Cherry Farms
Cherry Farms
Homestead
Homestead
Homestead
Homestead
Spring Ridge
Spring Ridge
Spring Ridge
Spring Ridge
Spring Ridge

The investigator determines that the Spring Ridge subdivision is to be used as the reference category. Thus, a binary column will be created (internally) for both Cherry Farms and Homestead. The resulting columns are

NBHD	CF	HS
Cherry Farms	1	0
Cherry Farms	1	0
Cherry Farms	1	0
Cherry Farms	1	0
Homestead	0	1
Homestead	0	1
Homestead	0	1
Homestead	0	1
Spring Ridge	0	0
Spring Ridge	0	0
Spring Ridge	0	0
Spring Ridge	0	0
Spring Ridge	0	0

The CF column has a one whenever NBHD is Cherry Farms, and the HS column has a one whenever NBHD is Homestead.

If the Multiple Regression for Appraisal procedure in NCSS were to be run with NBHD as one of the categorical model terms, the software would create two numeric (binary) columns internally, and the regression analysis would use those two columns rather than the NBHD column. This way, all the terms in the model are numeric.

Multicollinearity

Multicollinearity arises when two terms in the model highly correlate with each other. This can cause a distortion in estimated coefficients for both terms. Multicollinearity is a common issue in property valuation data, since it is expected that may of the attributes will correlate with each other (e.g., square feet and number of bedrooms, or quality and age).

Multicollinearity can be detected by examining the scatter plots and correlations of each term of the model with each other term. It can also be detected by looking for large variance inflation factors (*VIF*). A common rule of thumb is that multicollinearity is likely an issue when the VIF is around or above 10.

Two common ways to correct for multicollinearity are

1. When two columns are highly correlated with each other, remove one of the two columns from the model.

2. Using scatterplots or other tools, look for one or two outlying properties and remove them from the analysis. An outlying property is one that is far away from the bulk of the properties and does not fit the general trend.

Data Structure

Each column of the spreadsheet (dataset) represents a property attribute and each row represents a property. A sale price column is required. At least one (but likely more) attribute column(s) is needed to run the Multiple Regression for Appraisal procedure. A column may contain a continuous range of values, such as square feet or number of bathrooms, or a set of discrete values, such as neighborhood or style.

The following dataset of residential property sales gives an example of what a multiple regression model dataset may look like.

Recent Sales dataset (subset)

Sale_ Price	Main_ SF	Walls_ Type	Baths	BS_SF_ Fin	Age	Pool	Garage	Lot_SF	Lake_ Front	NBHD
147900	2612	Brick	2.5	0	23	0	2	14778	Yes	Park Grove
184000	2478	Siding	2.5	0	26	0	2	8465	Yes	Park Grove
225000	2617	Wood	2.5	0	26	0	2	8277	Yes	Park Grove
108561	2354	Siding	2.5	0	26	0	2	8277	Yes	Park Grove
165500	2603	Siding	2.5	0	24	0	2	8277	Yes	Park Grove
191000	2549	Brick	2.5	510	22	0	2	10280	Yes	Park Grove
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	.	.	
261000	2177	Siding	2.5	0	8	0	3	8170	Yes	Wood Village
260000	2337	Siding	2.5	0	8	0	3	8312	Yes	Wood Village
268900	2413	Brick	2.5	0	8	0	3	14238	Yes	Wood Village
281000	2015	Brick	2.5	0	8	0	2	9800	Yes	Wood Village
300000	2453	Brick	3.5	936	9	0	3	9361	Yes	Wood Village
225000	2536	Siding	2.5	0	9	0	2	9367	Yes	Wood Village

Recent Sales dataset column definitions

Sale_Price: Purchase price
Main_SF: Non-basement square feet
Walls_Type: Material of exterior walls
Baths: Number of (finished) bathrooms
BS_SF_Fin: Finished basement square feet
Age: Age in years of the residence
Pool: 0 = No pool, 1 = Pool
Garage: Number of attached garage spaces
Lot_SF: Size of lot in square feet
Lake_Front: Lake front property
NBHD: Name of subdivision

Missing Values

Rows with missing values for any of the columns in the analysis are ignored. That is, the whole row is removed from the analysis when there is a missing value for any used column in that row.

When the value of the sale price is missing (i.e., it is left blank), but values for all other used columns are non-missing, the estimated sale price for that row is generated (see Estimated Property Values and Confidence Limits report).

Procedure Options

This section describes the options available in this procedure.

Columns, Model Tab

Columns

Sale Price

Specify one or more columns containing the sale price values. If more than one column is specified, a separate analysis is produced for each column. You may type the column name or number directly, or you may use the column selection tool by clicking the column selection button to the right. The sale price in dependent variable (Y) in the regression model. It is also known as the dependent, response, or predicted variable. The form of the regression model is

Y = B0 + B1 * X1 + B2 * X2 + ...

where B0 is a baseline value (the intercept) and B1, B2,... are the coefficients for each of the explanatory variables (property attribute columns) X1, X2,...

Numeric X's

Specify one or more numeric attribute variables (columns). Numeric attributes are attributes that are measured. Examples include square feet, number of bathrooms, and age. The columns listed here become the X's of the model (X1, X2,...). At least one column in either the Numeric X's or Categorical X's is required. You may type the column names or numbers directly, or you may use the column selection tool by clicking the column selection button to the right.

Categorical X's

Specify one or more categorical attribute variables (columns) here. Categorical attributes are attributes that are categorized, not measured. Examples include neighborhood, pool or no pool, shed or no shed, and property type. Multiple regression is based solely on numeric variables. Thus, the software creates binary (0, 1) variables representing the categories of the column. The number of binary variables created is one less than the number of categories. First, a reference category is selected, either by default or by direct specification. Next, a variable is created for each category that is not the reference category. In each created variable, the value is 1 for properties that are of the category of the variable, and the value is 0 if the property is any of the other categories. For example, suppose Neighborhood is one of the columns of the dataset, where three neighborhood are represented: TM, RB, GV. Suppose TM is designated the reference category. Then, two variables will be created, one for GV, and one for RB.

Prop	NBHD	GV	RB
1	RB	0	1
2	RB	0	1
1	GV	1	0
2	GV	1	0
1	GV	1	0
2	TM	0	0
1	TM	0	0
2	TM	0	0
.	.	.	.
.	.	.	.
.	.	.	.

It can be seen here that neighborhood TM is represent by 0's for both GV and RB.

Reference Value

A specific reference value (category) may be designated by entering the value in parentheses. If no reference value is designated, the Default Reference Value is used.

Examples of Valid Categorical X's

POOL

POOL(Yes)

POOL(No)

NBHD

NBHD(Island Grove)

NBHD(West Park)

Default Reference Value

This option specifies the default reference value used when generating internal numeric (binary) variables from categorical variables. This default is used when the reference value is not designated between parentheses after the name of the categorical variable.

Possible Choices

- First Value after Sorting

  Use the first value after sorting as the reference value.

- Last Value after Sorting

  Use the last value after sorting as the reference value.

Regression Model Preview

Regression Model Preview

This box shows the terms in the current regression model based on the Numeric X's and Categorical X's entered.

Orientation

Press the Orientation (List / Across) button to toggle between a horizontal and vertical view of the model.

List / Across Preview Orientation

Press this button to toggle between a horizontal (Across) and vertical (List) preview of the model. The orientation has no impact on the actual model that is used.

Reports Tab

The following options control which reports are displayed.

Select Reports

The following options control which reports are displayed.

Select Reports – Summaries

Run Summary

Check to display this report. This report summarizes the multiple regression results. It presents the number of variables and rows used, basic statistical results such as R^2 and mean square error, and whether the routine completed normally.

Descriptive Statistics

Check to display this report. This report provides the count, arithmetic mean, standard deviation, minimum, and maximum of each variable. It is particularly useful for checking that the correct variables were used.

Correlations

Check to display this report. This report provides the Pearson correlations for all variables with all other variables. These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with another may cause collinearity problems. Note that these correlations may differ from pair-wise correlations generated by the Correlation Matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

Select Reports – Regression Coefficients

Coefficient T-Tests

Check to display this report. This reports the regression coefficients, standard errors, and significance tests. The significance tests are used to determine whether the coefficients (slopes) are significantly different from 0.

Coefficient Confidence Intervals

Check to display this report. This report provides the regression coefficients and their standard errors and confidence intervals.

Estimated Model (Reading Form)

Check to display this report. This report provides the estimated least-squares regression line in reading form. The number of decimals for each coefficient is specified on the Report Options.

Estimated Model (Transformation Form)

Check to display this report. This report displays the estimated model with the complete number of decimals available. This model can be copied and pasted as a transformation to the Column Info portion of the spreadsheet to give property value estimates.

Select Reports – ANOVA

ANOVA Summary

Check to display this report. This report provides the common ANOVA table summary. Essentially it tests whether the model as a whole has any estimation value.

ANOVA Detail

Check to display this report. This report provides the analysis of variance test and the R^2 for each term in the model.

Select Reports – Assumptions

Normality Tests

Check to display this report. This report provides the results of several normality tests of the residuals including the Shapiro-Wilk test and the Anderson-Darling test. Normality of residuals is one of the assumptions upon which multiple regression analysis is based.

Multicollinearity

Check to display this report. Multicollinearity arises when two terms in the model highly correlate with each other. This can cause a distortion in estimated coefficients for both terms. Multicollinearity is a common issue in property valuation data, since it is expected that may of the attributes will correlate with each other (e.g., square feet and number of bedrooms). This report provides information useful in assessing the amount of multicollinearity in the data. The two diagnostics in this report are the VIF (variance inflation factor) and the R^2 with other variables. High VIFs (say, greater than 10) and high R^2 with other variables (say, greater than 0.6 or 0.7) indicate a potential multicollinearity problem.

Select Reports – Row-by-Row Lists

Row-by-Row Lists to Show

This option makes it possible to limit the number of rows shown in the lists. This may be useful when you have a large number of rows of data.

Only Rows with no Sale Price

Only those rows for which the sale price value is blank are displayed. This would be used to display only those properties for which the market value is to be estimated.

All Rows

All rows are displayed.

Estimated Property Values and Confidence Limits

Check to display this list report. This report gives the estimated property values with confidence limits for each (specified) row of the dataset.

Residuals (Actual - Estimated) and Percent Error

Check to display this report. This report displays the residuals (actual value - estimated value) and the associated absolute percent error (100 * |residual| / actual value) for each (specified) row.

Alphas and Confidence Levels

Tests Alpha

Alpha is the significance level used in conducting the hypothesis tests.

Recommended

The value of 0.05 is usually used. This corresponds to a chance of 1 out of 20.

Range

Typical values range from 0.01 to 0.20.

Assumptions Alpha

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. This is used in the Normality Tests report. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests use a larger alpha such as 0.10, 0.15, or 0.20. A reasonable choice is 0.20.

Confidence Level

Enter the confidence level (or confidence coefficient) for the confidence intervals reported in this procedure. Note that, unlike the values of alpha, the confidence level is entered as a percentage. The recommended value is 95. The range is 80 to 99.99.

Definition

The interpretation of confidence level is that if confidence intervals are constructed across many analyses at the same confidence level, the percentage of such intervals that surround the true value of the parameter is equal to the confidence level.

Report Options Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. The number of decimal places shown in plots is controlled by the Tick Label Settings buttons on the Axes tabs.

Column Labels

Column Names

Specify whether to use column names, column labels, or both to label output reports.

Names

Column names are the column headings that appear on the data table. They may be modified by clicking the Column Info button on the Data window or by clicking the right mouse button while the mouse is pointing to the column heading.

Labels

This refers to the optional labels that may be specified for each column. Clicking the Column Info button on the Data window allows you to enter them.

Both

Both the column names and labels are displayed, one after the other.

Comments

1. Most reports are formatted to receive 12 to 18 characters for column names.

2. Column Names cannot contain spaces or math symbols (like + - * / . ,), but column labels can.

Stagger Label and Output

When writing a row of information to a report, some column names/labels may be too long to fit in the space allocated. If the name (or label) contains more characters than entered here, the rest of the output for that line is moved down to the next line.

Recommended

Most reports are designed to hold a label of up to around 20 characters.

Hint

Enter 1 when you always want each row's output to be printed on two lines. Enter 100 when you want each row printed on only one line. Note that this may cause some columns to be misaligned.

Decimal Places

Precision

This option is used when the number of decimal places is set to ***All***. It specifies whether numbers are displayed as single (7-digit) or double (13-digit) precision numbers in the output. All calculations are performed in double precision regardless of the Precision selected here.

Decimal Places

Specify the number of digits after the decimal point to display on the output of values of this type.

All keyword

Select All to display all digits available. The number of digits displayed by this option is controlled by whether the Precision option is Single or Double. The choice for this option in no way influences the accuracy with which the calculations are made.

Plots Tab

Select Plots

These options control the inclusion and the settings of each of the plots.

Select Plots – Sale Price Scatter Plots

Sale Price vs X Plot

Check this box to display a scatter plot for each model term with sale price on the vertical axis and the term value on the horizontal axis. These plots show the individual relationship between the sale price and each term. They can be used to determine whether a relationship exists, whether it is a straight line or curved, and/or whether there are outliers.

Select Plots – Distribution of Residuals

Histogram of Residuals

Check this box to display a histogram and/or density trace of the residuals. This histogram can be used to evaluate the assumption that the residuals are normally distributed (bell-shaped). Normality of residuals is one of the assumptions upon which multiple regression analysis is based.

Normal Probability Plot of Residuals

Check this box to display a normal probability plot of the residuals. If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this line reflect departures from normality. Deviating points at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Use of this graphic tool with very small sample sizes is not recommended. If the residuals are not normally distributed, then the t-tests on regression

coefficients, the F-tests, and any interval estimates have questionable validity. Normality of residuals is one of the assumptions upon which multiple regression analysis is based.

Storage Tab

Data Storage Options

Storage Option

This option controls whether the values for the selected check boxes below are stored to the dataset when the procedure is run.

- **Do not store data**
  No data are stored to the dataset.

- **Store in empty columns only**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**
  Beginning at the *Store First Item In*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Item In

The first saved value is stored in this column. Each additional value that is specified is stored in the columns immediately to the right of this column. Leave this box blank if you want the data storage to begin in the first blank column dataset. Any existing data in these columns is automatically replaced.

Data Storage Options – Select Items to Store

Store Estimated Sale Price

Indicated whether to store each estimated sale price to a column on the dataset.

Store Residuals (Actual - Estimated)

Indicated whether to store each residual to a column on the dataset.

Store Estimated Sale Price Lower Confidence Limit

Indicated whether to store each estimated sale price lower confidence limit to a column on the dataset.

Store Estimated Sale Price Upper Confidence Limit

Indicated whether to store each estimated sale price upper confidence limit to a column on the dataset.

Example 1 – Multiple Regression for Appraisal – All Reports

This section presents an example of estimating the coefficients of a multiple regression model based on the Recent Sales dataset. The Recent Sales dataset contains the sale price and attribute information about 125 properties. The property values of 3 properties without sale price information are to be estimated. The attribute values for these 3 properties are given in the last three rows (126, 127, and 128) of the dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Multiple Regression for Appraisal window.

1 Open the Recent Sales dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Recent Sales.NCSS**.
- Click **Open**.

2 Open the Multiple Regression for Appraisal window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Multiple Regression for Appraisal** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the columns.

- On the Multiple Regression for Appraisal window, select the **Columns, Model tab**.
- Set the **Y** box to **Sale_Price**.
- Set the **Numeric X's** box to **Main_SF, Baths-Age, Garage, Lot_SF**.
- Set the **Categorical X's** box to **Walls_Type(Siding), Pool(0), Lake_Front(No), NBHD(Park Grove)**. A set of binary columns will be made internally for each Categorical X. The values in parentheses are the reference or baseline categories.

4 Specify the reports.

- Select the **Reports tab**.
- Check the boxes for all reports.

5 Specify the plots.

- Select the **Plots tab**.
- Check the boxes for all plots.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Report

Item	Value	Rows	Value
Sale Price Column	Sale_Price	Rows Processed	128
Number Ind. Variables	12	Rows Filtered Out	0
R^2	0.8914	Rows with X's Missing	0
Adj R^2	0.8798	Rows with Sale Price Missing	3
Coefficient of Variation	0.1393	Rows Used in Model Estimation	125
Mean Square Error	1.012713E+09		
Square Root of MSE	31823.14		
Average |Percent Error|	11.382		
Completion Status	Normal Completion		

This report summarizes the multiple regression results. Several model statistics are shown, as well as a summary of the rows in the analysis. Notice the Average absolute percent error (11.382) is similar to that found in the Hybrid Appraisal Models example (11.72). Details of the items listed may be viewed in the first example of the Multiple Regression procedure.

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Main_SF	125	2435.488	357.4667	1617	3663
Baths	125	2.702	0.4608635	1	3.5
BS_SF_Fin	125	369.016	531.7466	0	2289
Age	125	17.288	6.653663	5	26
Garage	125	2.376	0.4863292	2	3
Lot_SF	125	10265.49	2449.599	7869	20162
(Walls_Type="Brick")	125	0.296	0.4583279	0	1
(Walls_Type="Wood")	125	0.08	0.2723849	0	1
(Pool=1)	125	0.024	0.1536649	0	1
(Lake_Front="Yes")	125	0.888	0.3166355	0	1
(NBHD="Glen Lake")	125	0.064	0.2457379	0	1
(NBHD="Wood Village")	125	0.424	0.496179	0	1
Sale_Price	125	228406.9	91778.09	99900	610000

This report presents a brief numeric summary for each of the model terms, including the binary terms created from the categorical X's. Sometimes this report is useful for determining whether the proper columns were used.

Correlation Matrix

	Main_SF	Baths	BS_SF_Fin	Age
Main_SF	1.0000	0.4748	0.5861	-0.0296
Baths	0.4748	1.0000	0.6524	-0.2847
BS_SF_Fin	0.5861	0.6524	1.0000	-0.2002
Age	-0.0296	-0.2847	-0.2002	1.0000
Garage	0.2345	0.2521	0.2211	-0.4998
Lot_SF	0.4518	0.1386	0.2937	-0.1569
(Walls_Type="Brick")	0.4176	0.2396	0.3224	-0.0467
(Walls_Type="Wood")	0.0243	-0.2583	0.0176	0.3254
(Pool=1)	-0.0660	-0.0690	-0.0698	0.0090
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

This report gives the Pearson correlation of each term with each other term. Terms that are highly correlated (greater than, say, 0.4 or 0.5) may indicate a multicollinearity problem.

Regression Coefficients T-Tests

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Statistic to Test H0: β(i)=0	P-Value	Reject H0 at 5%?
Intercept	120082.9	57704.18	2.081	0.0397	Yes
Main_SF	48.04036	12.16535	3.949	0.0001	Yes
Baths	-5691.82	9414.582	-0.605	0.5467	No
BS_SF_Fin	25.94215	8.952463	2.898	0.0045	Yes
Age	-3839.612	1743.181	-2.203	0.0297	Yes
Garage	17188.51	7923.555	2.169	0.0322	Yes
Lot_SF	-0.3096872	1.478794	-0.209	0.8345	No
(Walls_Type="Brick")	29858.78	8122.794	3.676	0.0004	Yes
(Walls_Type="Wood")	24299.47	12460.23	1.950	0.0537	No
(Pool=1)	52406.63	19229.7	2.725	0.0075	Yes
(Lake_Front="Yes")	1630.331	12414.08	0.131	0.8958	No
(NBHD="Glen Lake")	191800.7	28562.95	6.715	0.0000	Yes
(NBHD="Wood Village")	198.096	23112.5	0.009	0.9932	No

This section reports the values and significance tests of the regression coefficients. The significance test is whether the coefficient estimate is statistically different from 0. Terms with larger P-values (closer to 1) indicate those terms may not be contributing well to the model. In this example, it appears that Baths, Lot_SF, and Lake_Front are all candidates for removal from the model. One of the subset selection procedures in NCSS should be used if you wish to have the software sort through and remove the terms of the model automatically.

Regression Coefficients Confidence Intervals

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	120082.9	57704.18	5749.489	234416.3
Main_SF	48.04036	12.16535	23.93627	72.14445
Baths	-5691.82	9414.582	-24345.61	12961.97
BS_SF_Fin	25.94215	8.952463	8.203996	43.68031
Age	-3839.612	1743.181	-7293.502	-385.7214
Garage	17188.51	7923.555	1489.004	32888.02
Lot_SF	-0.3096872	1.478794	-3.239727	2.620353
(Walls_Type="Brick")	29858.78	8122.794	13764.51	45953.06
(Walls_Type="Wood")	24299.47	12460.23	-388.872	48987.82
(Pool=1)	52406.63	19229.7	14305.45	90507.8
(Lake_Front="Yes")	1630.331	12414.08	-22966.58	26227.24
(NBHD="Glen Lake")	191800.7	28562.95	135206.9	248394.5
(NBHD="Wood Village")	198.096	23112.5	-45596.36	45992.55

Note: The T-Value used to calculate these confidence limits was 1.981.

The confidence limits for each coefficient give a feel for the variation in possible true coefficient values. Typically, larger numbers of properties analyzed result in narrower intervals.

Estimated Model (Reading Form)

Estimated Model: Estimated Market Value (Sale_Price) =
120082.92 + 48.04 * Main_SF - 5691.82 * Baths + 25.94 * BS_SF_Fin - 3839.61 * Age + 17188.51 * Garage - 0.31 * Lot_SF + 29858.78 * (Walls_Type="Brick") + 24299.47 * (Walls_Type="Wood") + 52406.63 * (Pool=1) + 1630.33 * (Lake_Front="Yes") + 191800.73 * (NBHD="Glen Lake") + 198.10 * (NBHD="Wood Village")

This report shows the model in reading form. The number of decimal places for the parameter estimates is set by the user.

Estimated Model (Transformation Form) Report

Estimated Model (Transformation Form)
This model can be copied and pasted as a transformation to the Column Info portion of the spreadsheet to give property value estimates.

Estimated Model: Estimated Market Value (Sale_Price) =
120082.916540956+48.0403578675254*Main_SF-5691.81998850623*Baths+25.9421538058507*BS_SF_Fin-3839.61158170051*Age+17188.5122535601*Garage-0.309687166595788*Lot_SF+29858.7826034505*(Walls_Type="Brick")+24299.4727693489*(Walls_Type="Wood")+52406.6271601821*(Pool=1)+1630.3309054377*(Lake_Front="Yes")+191800.732024224*(NBHD="Glen Lake")+198.096042298154*(NBHD="Wood Village")

This is the model with full precision coefficient estimates. This expression may be copied onto the Clipboard and pasted into a transformation cell of the dataset to estimate other properties. This expression is always provided in double precision.

Analysis of Variance

Source	DF	R^2 Lost If Term(s) Removed	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1		6.521214E+12	6.521214E+12		
Model	12	0.8914	9.310553E+11	7.758795E+10	76.614	0.0000
Error	112	0.1086	1.134238E+11	1.012713E+09		
Total(Adjusted)	124		1.044479E+12	8.423219E+09		

A P-value near 0 indicates that the model as a whole has predictive value for estimating sale prices. See the Multiple Regression documentation chapter for more details.

Analysis of Variance Detail

Source	DF	R^2 Lost If Term(s) Removed	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1		6.521214E+12	6.521214E+12		
Model	12	0.8914	9.310553E+11	7.758795E+10	76.614	0.0000
Main_SF	1	0.0151	1.579244E+10	1.579244E+10	15.594	0.0001
Baths	1	0.0004	3.701574E+08	3.701574E+08	0.366	0.5467
BS_SF_Fin	1	0.0081	8.503803E+09	8.503803E+09	8.397	0.0045
Age	1	0.0047	4.913327E+09	4.913327E+09	4.852	0.0297
Garage	1	0.0046	4.765655E+09	4.765655E+09	4.706	0.0322
Lot_SF	1	0.0000	4.441374E+07	4.441374E+07	0.044	0.8345
Walls_Type	2	0.0140	1.462025E+10	7.310125E+09	7.218	0.0011
Pool	1	0.0072	7.521661E+09	7.521661E+09	7.427	0.0075
Lake_Front	1	0.0000	1.74666E+07	1.74666E+07	0.017	0.8958
NBHD	2	0.1040	1.086739E+11	5.433695E+10	53.655	0.0000
Error	112	0.1086	1.134238E+11	1.012713E+09		
Total(Adjusted)	124		1.044479E+12	8.423219E+09		

These tests are essentially the same tests as the regression coefficients tests, except for the case where there is a categorical X with more than 2 categories. These tests provide a good measure of whether the term should be included in the model. Columns for which the P-value is close to 0 should be kept. More details about the meaning of each column in this table are given in the Multiple Regression chapter.

Normality Tests

Test Name	Test Statistic	P-Value	Reject Normality at 20%?
Shapiro Wilk	0.991	0.5959	No
Anderson Darling	0.246	0.7570	No
D'Agostino Skewness	0.042	0.9662	No
D'Agostino Kurtosis	1.544	0.1227	Yes
D'Agostino Omnibus	2.384	0.3036	No

This report gives the results of applying several normality tests to the residuals. The Shapiro-Wilk test is probably the most commonly used. When the residuals cannot be assumed to be Normally distributed, the reliability of the coefficient estimates may be in question. These tests are discussed in detail in the Normality Test section of the Descriptive Statistics procedure.

Multicollinearity Report

Independent Variable	Variance Inflation Factor	R^2 Versus Other I.V.'s
Main_SF	2.3156	0.5681
Baths	2.3051	0.5662
BS_SF_Fin	2.7748	0.6396
Age	16.4718	0.9393
Garage	1.8182	0.4500
Lot_SF	1.6067	0.3776
(Walls_Type="Brick")	1.6971	0.4107
(Walls_Type="Wood")	1.4104	0.2910
(Pool=1)	1.0691	0.0647
(Lake_Front="Yes")	1.8918	0.4714
(NBHD="Glen Lake")	6.0323	0.8342
(NBHD="Wood Village")	16.1030	0.9379

Both the Variance Inflation Factor (VIF) and R-squared Versus Other Independent Variables are useful measures of multicollinearity for that term. VIFs near to or greater than 10 typically indicate significant multicollinearity. R-squared values greater than 0.7 or 0.8 also indicate multicollinearity. There is a brief discussion of multicollinearity earlier in the chapter. The Multiple Regression chapter or a regression analysis text may be useful for learning more about dealing with multicollinearity. This report indicates that Age or a term that is highly correlated with Age should probably be removed from the model.

Estimated Property Values and Confidence Limits

Row	Actual Sale_Price	Estimated Sale_Price	Standard Error of Estimated	95% Lower Limit	95% Upper Limit
1	147900	204313.3	33107.13	138715.8	269910.8
2	184000	158453.3	32532.1	93995.14	222911.5
3	225000	189488.6	34125.84	121872.7	257104.6
4	108561	152554.5	32509.92	88140.3	216968.8
5	165500	172195.8	32522.43	107756.8	236634.8
6	191000	219749.8	32619.71	155118.1	284381.6
7	276000	239710.5	33746.93	172845.3	306575.8
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
122	268900	269901.3	33429.27	203665.5	336137.1
123	281000	234967.1	33459.74	168670.9	301263.3
124	300000	288083.7	33387.28	221931	354236.3
125	225000	226431.8	33038.35	160970.6	291893.1
126		227327.7	33284.24	161379.3	293276.2
127		223182.1	34296.47	155228	291136.2
128		195862.8	33108.41	130262.8	261462.9

This report gives the estimated property value and confidence limits for each property. Property values are also estimated for rows where the actual sale price is blank.

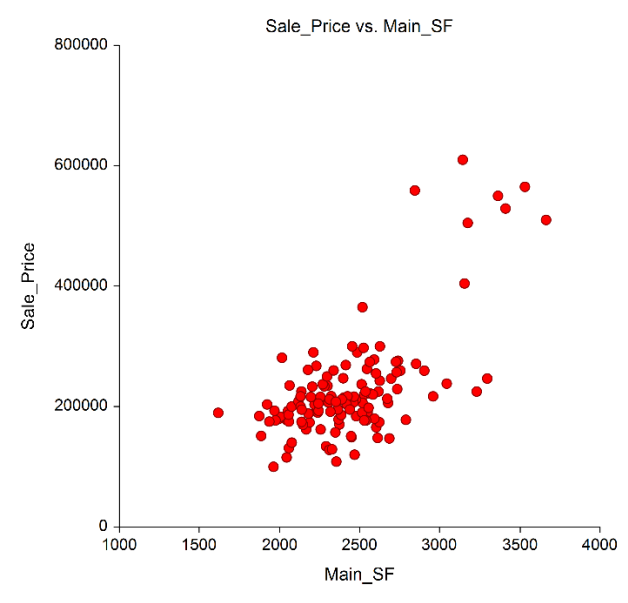
Residual Report Residuals (Actual - Estimated) and Percent Error Report

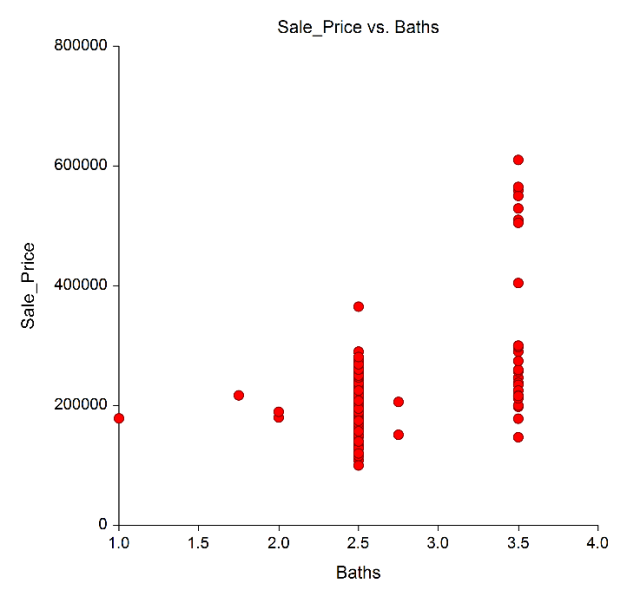
Row	Actual Sale_Price	Estimated Sale_Price	Residual	Absolute Percent Error	
1	147900	204313.3	-56413.3	38.14	
2	184000	158453.3	25546.67	13.88	
3	225000	189488.6	35511.37	15.78	
4	108561	152554.5	-43993.54	40.52	
5	165500	172195.8	-6695.815	4.05	
6	191000	219749.8	-28749.84	15.05	
7	276000	239710.5	36289.45	13.15	
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
122	268900	269901.3	-1001.278	0.37	
123	281000	234967.1	46032.91	16.38	
124	300000	288083.7	11916.34	3.97	
125	225000	226431.8	-1431.822	0.64	
126		227327.7			
127		223182.1			
128		195862.8			

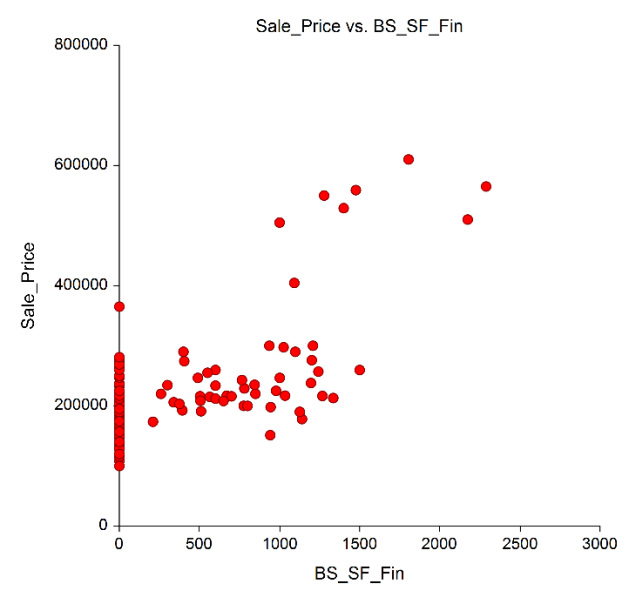
This section shows the distance between the actual sale prices and the estimated sale prices. This difference is also shown as a percent difference. No residuals or errors are given for the properties without a known sale price.

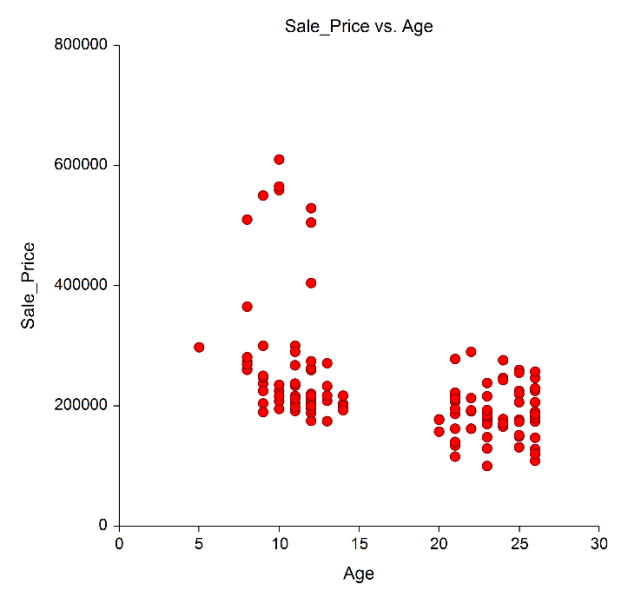
Plots of Sale Price versus Each X

It is often useful to examine sale price scatter plots as a preliminary step to forming the model. These plots can be used to identify trends, outliers, or other anomalies in the data.









(additional plots follow)

Histogram

The general purpose of the histogram is to determine the shape of the distribution of residuals, and, in particular, to determine if the residuals are Normally distributed (bell-shaped). The histogram below seems to show a very clean, symmetric distribution.

Probability Plot of Residuals

If the residuals are normally distributed, the data points of the Normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this line indicate departures from normality. If the residuals are not normally distributed, the validity of the tests and confidence limits of the report may be in question. The probability plot seems to indicate a Normal distribution, with only a few points at the ends away from the line.