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).

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.

Setup

To run this example, complete the following steps:

1 Open the Recent Sales example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Recent Sales** and click **OK**.

2 Specify the Multiple Regression for Appraisal procedure options

- Find and open the **Multiple Regression for Appraisal** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Columns, Model Tab	
Sale Price	**Sale_Price**
Numeric X's	**Main_SF, Baths-Age, Garage, Lot_SF**
Categorical X's	**Walls_Type(Siding), Pool(0), Lake_Front(No), NBHD(Park Grove)**
Reports Tab	
All Available Reports	**Checked** (click the *Check All* button)
Plots Tab	
All Available Plots	**Checked** (click the *Check All* button)
Report Options (*in the Toolbar*)	
Variable Labels	**Column Names**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Sale Price Column	Sale_Price	Rows Processed	128
Number of Independent Variables (X)	12	Rows Used in Estimation	125
R^2	0.8914	Rows with X's Missing	0
Adjusted R^2	0.8798	Rows with Sale Price Missing	3
Coefficient of Variation	0.1393		
Mean Square Error (MSE)	1.012713E+09		
Square Root of MSE	31823.14		
Average |Percent Error|	11.3815		
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

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

Correlation Matrix

Section 1

	Main_SF	Baths	BS_SF_Fin	Age	Garage	Lot_SF
Main_SF	1.0000	0.4748	0.5861	-0.0296	0.2345	0.4518
Baths	0.4748	1.0000	0.6524	-0.2847	0.2521	0.1386
BS_SF_Fin	0.5861	0.6524	1.0000	-0.2002	0.2211	0.2937
Age	-0.0296	-0.2847	-0.2002	1.0000	-0.4998	-0.1569
Garage	0.2345	0.2521	0.2211	-0.4998	1.0000	0.2289
Lot_SF	0.4518	0.1386	0.2937	-0.1569	0.2289	1.0000
(Walls_Type="Brick")	0.4176	0.2396	0.3224	-0.0467	0.1479	0.3619
(Walls_Type="Wood")	0.0243	-0.2583	0.0176	0.3254	-0.1680	0.0570
(Pool=1)	-0.0660	-0.0690	-0.0698	0.0090	-0.0138	0.0464
(Lake_Front="Yes")	0.0125	0.0319	-0.1098	0.2949	-0.4051	-0.1610
(NBHD="Glen Lake")	0.6241	0.4546	0.5901	-0.2728	0.3369	0.4651
(NBHD="Wood Village")	-0.2464	0.0721	-0.0677	-0.8336	0.3032	-0.0548
Sale_Price	0.6660	0.5042	0.6603	-0.4721	0.4780	0.4810

.
.
.

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 Coefficient T-Tests

Regression Coefficient T-Tests

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Test of H0: β(i) = 0 T-Statistic	P-Value	Reject H0 at α = 0.05?
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 Coefficient Confidence Intervals

Regression Coefficient Confidence Intervals

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	95% Confidence Interval Limits for β(i) Lower	Upper
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 the confidence interval limits was 1.981.

The confidence interval 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 (Reading Form)

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)

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 model can be copied and pasted as a transformation to the Column Info portion of the Data Window to give property value estimates.

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 Summary

Analysis of Variance Summary

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

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.

Residual Normality Tests

Residual Normality Tests

	Test of H0: Residuals Normally Distributed		
Test Name	Test Statistic Value	P-Value	Reject H0 at α = 0.2?
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

Multicollinearity Report

Independent Variable (IV)	Variance Inflation Factor	R^2 Versus Other IV'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

Estimated Property Values and Confidence Interval Limits

	Sale_Price		Standard Error of	95% Confidence Interval Limits	
Row	Actual	Estimated	Estimated	Lower	Upper
1	147900	204313.3	33107.13	138715.80	269910.8
2	184000	158453.3	32532.10	93995.14	222911.5
3	225000	189488.6	34125.84	121872.70	257104.6
4	108561	152554.5	32509.92	88140.30	216968.8
5	165500	172195.8	32522.43	107756.80	236634.8
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
121	260000	238226.6	32909.87	173019.90	303433.3
122	268900	269901.3	33429.27	203665.50	336137.1
123	281000	234967.1	33459.74	168670.90	301263.3
124	300000	288083.7	33387.28	221931.00	354236.3
125	225000	226431.8	33038.35	160970.60	291893.1
126		227327.7	33284.24	161379.30	293276.2
127		223182.1	34296.47	155228.00	291136.2
128		195862.8	33108.41	130262.80	261462.9

This report gives the estimated property value and confidence interval 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

Residuals (Actual - Estimated) and Percent Error Report

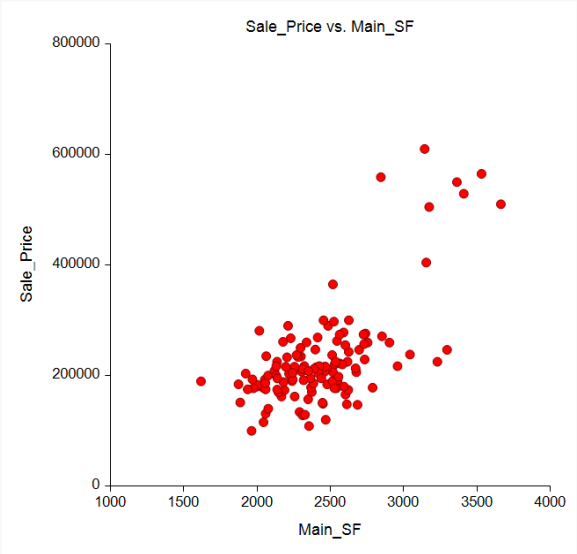
	Sale_Price			
Row	Actual	Estimated	Residual	Absolute Percent Error
1	147900	204313.3	-56413.3000	38.1429
2	184000	158453.3	25546.6700	13.8841
3	225000	189488.6	35511.3700	15.7828
4	108561	152554.5	-43993.5400	40.5243
5	165500	172195.8	-6695.8150	4.0458
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
121	260000	238226.6	21773.3700	8.3744
122	268900	269901.3	-1001.2780	0.3724
123	281000	234967.1	46032.9100	16.3818
124	300000	288083.7	11916.3400	3.9721
125	225000	226431.8	-1431.8220	0.6364
126		227327.7		
127		223182.1		
128		195862.8		

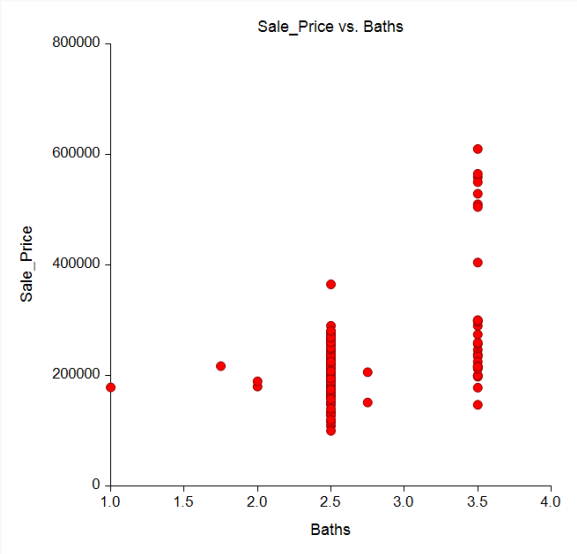
This section shows the distance between the actual sale prices and the estimated sale prices. This difference is also shown as a percent difference. No residuals or errors are given for the properties without a known sale price.

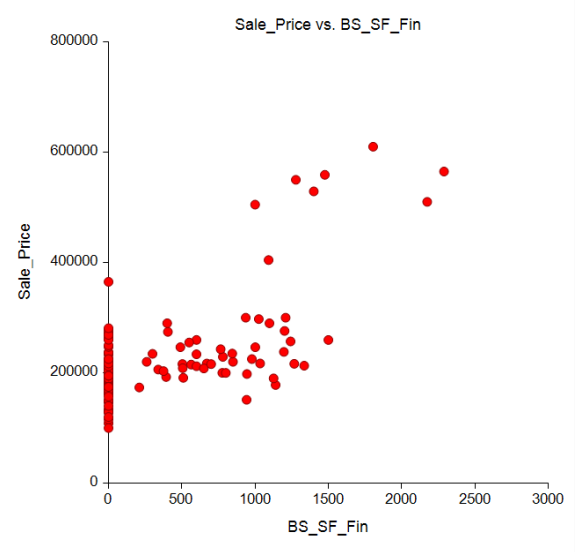
Sale Price vs X's Plots

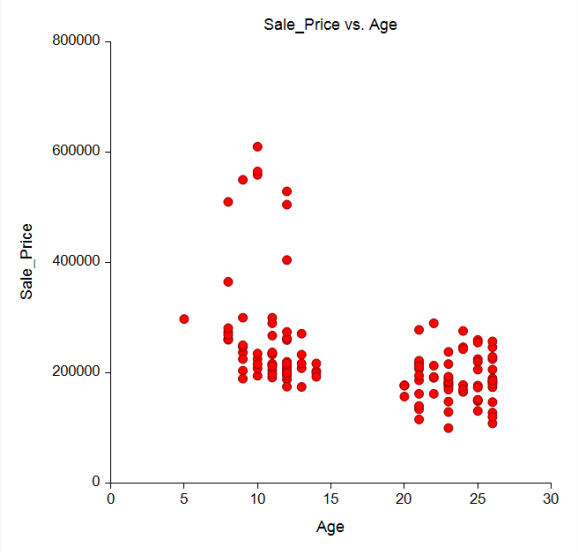
It is often useful to examine sale price scatter plots as a preliminary step to forming the model. These plots can be used to identify trends, outliers, or other anomalies in the data.

Sale Price vs X Plots









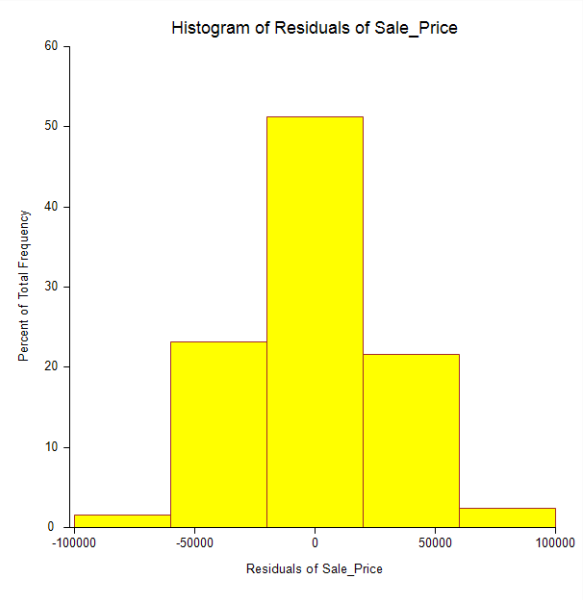
(additional plots follow)

Residual Distribution Plots

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.

Residual Distribution Plots



Normal 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.

Residual Distribution Plots