Assignment #3

James Gray

Introduction:

This study is composed of three parts using the building_prices data set variables shown below.

Y = Sales price of the house (thousands of US dollars)
X1 = Taxes (thousands of dollars)
X2 = Bathrooms (number)
X3 = Lot size (thousands of feet)
X4 = Living space (thousands of feet)
X5 = Garage stalls (number)
X6 = Rooms (number)
X7 = Bedrooms (number)
X8 = Age of the home (years)
X9 = Fireplaces (number)

The purpose of part 1 is to fit an optimal multiple regression model to the building_prices data set using automated variable selection procedures. Forward, backward and stepwise procedures are used to find the optimal multiple regression model. The model selection results are then evaluated to explain why each procedure may select a different model. These multiple regression models are then compared to a simple regression model based on X1 to determine if the multiple regression models generated by automated variable selection are more predictive or not. The purpose of part 2 is to assess the adequacy of the optimal model from part 1 using various diagnostic techniques. The purpose of part 3 is to evaluate a model with only two predictors (taxes (X1) and number of bathrooms (X2)) and determine if the X2 variable is best modeled as continuous or discrete variable.

Results:

Part 1 – Selecting an Optimal Regression Model using Variable Selection

In a previous assignment an optimal simple regression model was calculated using the X1 predictor. The simple regression model is as follows using the calculated coefficients from Figure 1:

$$Y = 13.3553 + 3.215X1 \quad (1)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.04186	635.04186	71.11	<.0001
Error	22	196.46772	8.93035		
Corrected Total	23	831.50958			

Root MSE	2.98837	R-Square	0.7637
Dependent Mean	34.62917	Adj R-Sq	0.7530
Coeff Var	8.62963		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.35530	2.59548	5.15	<.0001
X1	1	3.32151	0.39388	8.43	<.0001

Figure 1 - Simple regression model coefficients using X1

Three automated variable selection methods were used to choose an optimal regression model.

- Forward method – starts with the single best variable (the one that yields the largest F statistic and adds variables on at a time until the p-value for the variable being entered is larger than the specified value (Cody, 2011). This method is preferred when there are a large number of predictors.
- Backward elimination method – starts with all variables in the model and removes them one at a time (the one with the largest p-value leaves first) until all variables being considered for removal have p-values smaller than a given value (Cody, 2011). This method is preferred when there is a small set of variables.
- Stepwise selection method – very similar to the forward method except that a variable that has already been added to the model at a previous step might be removed later Cody, 2011).

The SAS default values for SLENTRY and SLSTAY in the PROC REG selection option were used in all calculations. The SLENTRY values specifies the significance level of variables entering the model and SLSTAY specifies the significance level of variables that stay within the model.

The forward selection process was executed and seven of the nine predictors were selected as shown in Figure 2. The additional six predictors increased the R-square (goodness of fit) from 0.7637 to 0.8491.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X1	1	0.7637	0.7637	2.2301	71.11	<.0001
2	X2	2	0.0343	0.7981	0.9991	3.57	0.0727
3	X9	3	0.0131	0.8112	1.7634	1.39	0.2520
4	X8	4	0.0119	0.8231	2.6410	1.28	0.2717
5	X5	5	0.0134	0.8365	3.3785	1.48	0.2398
6	X6	6	0.0074	0.8440	4.6798	0.81	0.3809
7	X4	7	0.0051	0.8491	6.2005	0.54	0.4730

Figure 2 - Forward variable selection summary

The multiple regression model formed by forward selection is as follows using the coefficients from Figure 3:

$$Y = 16.5901 + 2.2187X_1 + 6.1408X_2 + 2.867X_4 + 1.8553X_5 - 1.3164X_6 - 0.04656X_8 + 2.2518X_9 \quad (2)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	706.00703	100.85815	12.86	<.0001
Error	16	125.50255	7.84391		
Corrected Total	23	831.50958			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	16.59015	4.87745	90.74999	11.57	0.0036
X1	2.21867	0.80405	59.72386	7.61	0.0140
X2	6.14082	3.80521	20.42811	2.60	0.1261
X4	2.86700	3.90116	4.23644	0.54	0.4730
X5	1.85534	1.23618	17.66910	2.25	0.1529
X6	-1.31636	1.21900	9.14690	1.17	0.2962
X8	-0.04656	0.06067	4.61921	0.59	0.4540
X9	2.25175	1.43232	19.38610	2.47	0.1355

Figure 3 - Forward selection model fit

The backward elimination method was executed and seven of the nine variables were removed as shown in Figure 4. The R-square coefficient was 0.7981 as the last variable was removed and the addition of the X2 variable makes this model more predictive that the simple regression model.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X6	8	0.0006	0.8506	8.0537	0.05	0.8200
2	X3	7	0.0009	0.8497	6.1430	0.10	0.7618
3	X8	6	0.0041	0.8456	4.5242	0.43	0.5207
4	X4	5	0.0060	0.8396	3.0912	0.66	0.4265
5	X9	4	0.0075	0.8321	1.7954	0.84	0.3715
6	X5	3	0.0251	0.8071	2.1530	2.84	0.1085
7	X7	2	0.0090	0.7981	0.9991	0.93	0.3458

Figure 4 - Backward elimination selection summary

The multiple regression model formed by backwards selection is as follows using the coefficients from Figure 5:

$$Y = 10.1120 + 2.7170X_1 + 6.0985X_2 \quad (3)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.11203	2.99614	91.07817	11.39	0.0029
X1	2.71703	0.49115	244.69696	30.60	<.0001
X2	6.09851	3.22705	28.55593	3.57	0.0727

Figure 5 - Backward Elimination model fit

The stepwise selection process was executed and two predictors were selected as shown in Figure 6.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X1		1	0.7637	0.7637	2.2301	71.11	<.0001
2	X2		2	0.0343	0.7981	0.9991	3.57	0.0727

Figure 6 - Stepwise selection summary

The multiple regression model formed by stepwise selection is as follows using the coefficients from Figure 7:

$$Y = 10.1120 + 2.7170X_1 + 6.0985X_2 \quad (4)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.11203	2.99614	91.07817	11.39	0.0029
X1	2.71703	0.49115	244.69696	30.60	<.0001
X2	6.09851	3.22705	28.55593	3.57	0.0727

Figure 7 - Stepwise model fit

Figure 8 summarizes the regression models produced by each variable selection method. The main reason for differences in these equations is the significance level set for variables entering and exiting the model.

Method	Optimal Regression Model
Forward	Y = 16.5901 + 2.2187X_1+ 6.1408X_2+2.867X_4+ 1.8553X_5-1.3164X_6-0.04656X_8+2.2518X_9
Backward	Y = 10.1120 + 2.7170X_1 + 6.0985X_2
Stepwise	Y = 10.1120 + 2.7170X_1 + 6.0985X_2

Figure 8 -Automated Variable Selection Models

In the forward procedure the p-value for the F statistic is set to 0.50 so this will allow many variables into the model. In the backward elimination procedure the p-value for the F statistic is set to 0.10 and therefore more variables will be removed. The regression models for forward and backward selection would be identical if the same p-value was used. In the stepwise procedure the p-value for the F statistic is set to 0.15 for both entry and stay.

The predictive strength of each of the four models is summarized in Table 1 using the information above. The R-square calculations confirm that all three multiple regression models (2, 3, 4) are more predictive that the simple regression model (1).

Model	R-Square
Simple Regression	0.7637
Forward	0.8491
Backward	0.7981
Stepwise	0.7981

Table 1 - R-square summary for all four models

The Mallows Cp statistic is used to evaluate the multiple regression models produced by the automated variable selection processes to select the best model. When using the Mallows Cp statistic the optimal model is the first model where Cp is less than or equal to the number of predictors.

Model	Cp	Predictors (#)	Predictor - Cp
Forward	6.2005 (Figure 2)	7	0.7995
Backward	0.991 (Figure 4)	2	1.009
Stepwise	0.991 (Figure 6)	2	1.009

Table 2 - Mallow Cp Statistic for Comparing Regression Models

The regression model produced by the forward procedure is considered the optimal model due that Cp is less than the number of predictors and the difference between the number of predictors and Cp is the smallest when compared to the backward and stepwise processes.

Part 2 – Evaluating Adequacy of the Optimal Regression Model

The adequacy of regression models are predicated on a set of assumptions and the application of the model may result in errors if these assumptions are violated. We will use a set of visualizations and diagnostics to test model adequacy.

Linearity Assumption

The plot of the fitted regression model over the scatterplot of the observations shown in Figure 9 confirms linearity by the data points that are evenly dispersed around the regression line.

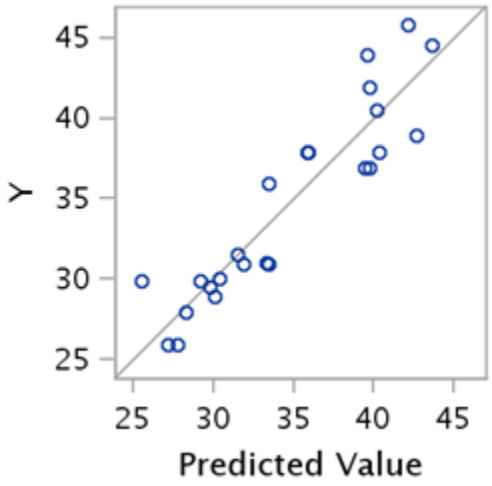

Figure 9 - Fitted Regression Model and Scatterplot

Normality Assumption

An assumption of Ordinary Least Squares (OLS) regression is that the residuals are normally distributed. The Quantile-Quantile plot shown by Figure 10 is a graphical method that confirms that data come from a normal distribution when the data points cluster closely to the straight line. The residual errors not

explained by the regression model are tightly dispersed around the line thereby confirming the errors have a normal distribution. This confirms the normality assumption.

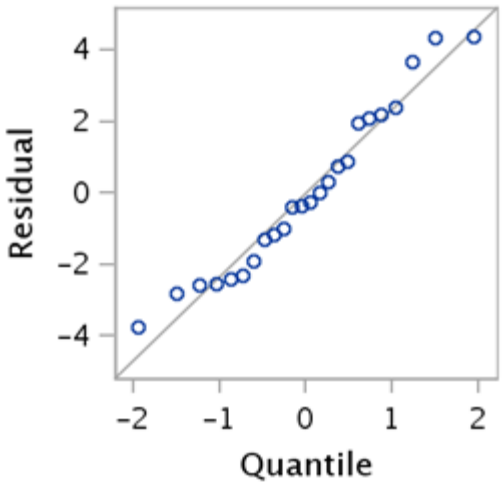

Figure 10 - Quantile-Quantile Plot for Normality

Normality is also confirmed by the frequency histogram (Figure 11) that shows a normal distribution of the residuals.

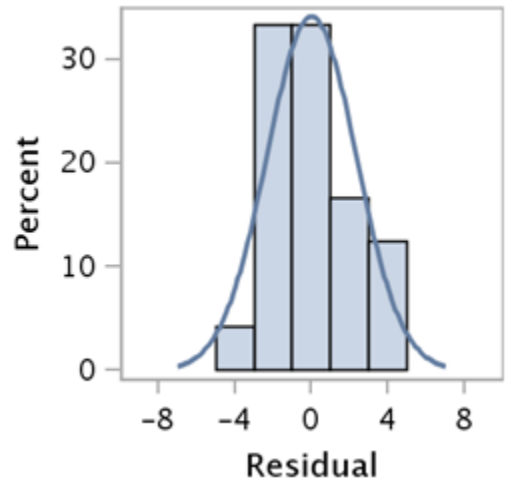

Figure 11 - Frequency Histogram of Residuals

Constant Variance Assumption

Under normal conditions the standardized residuals are uncorrelated with the predictors and the plot should represent a random scatter of data points. In Figure 12 below the residuals are plotted against each of the predictor. The standardized residuals are not correlated to the predictors X1, X4 and X8 given the random nature of the scatter. The residuals are correlated to the predictors X2, X5, X6 and X9 shown by the straight lines in those plots. Predictors X2, X5, X6 and X9 violate the normality assumption.

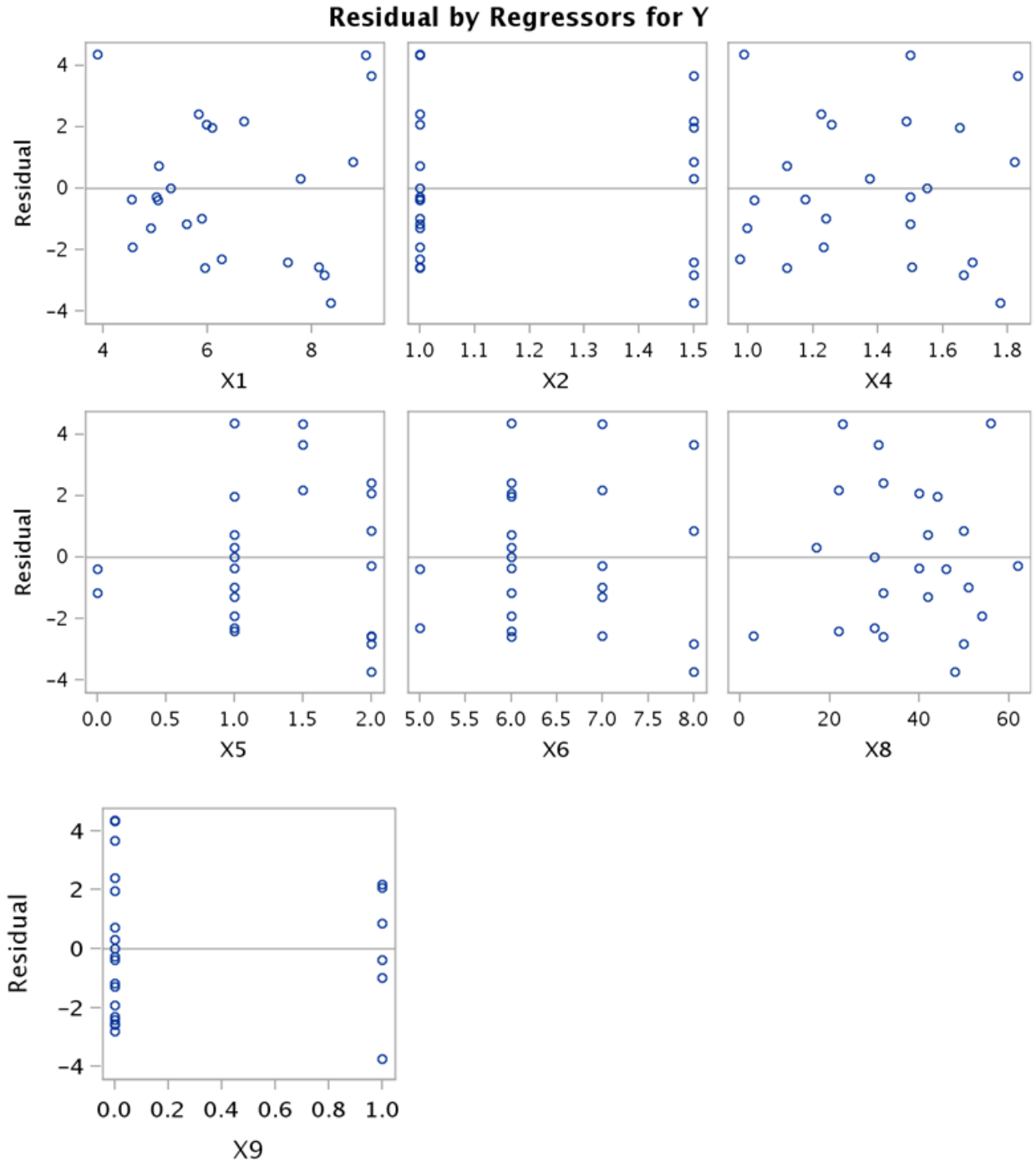

Figure 12 - Standardized Residuals by Predictor X1 and X2

Outliers and Influential Observations

Analysis is required to ensure that model fitting is not heavily influenced by one or a few data points. The Cooke's distance metric measures the influence of each observation and identifies any highly influential observations above a critical line (Figure 13). Observations 15, 17 and 23 are flagged as highly influential data points and should be examined in more detail. The regression model should be refitted without these observations to determine the impact on the model.

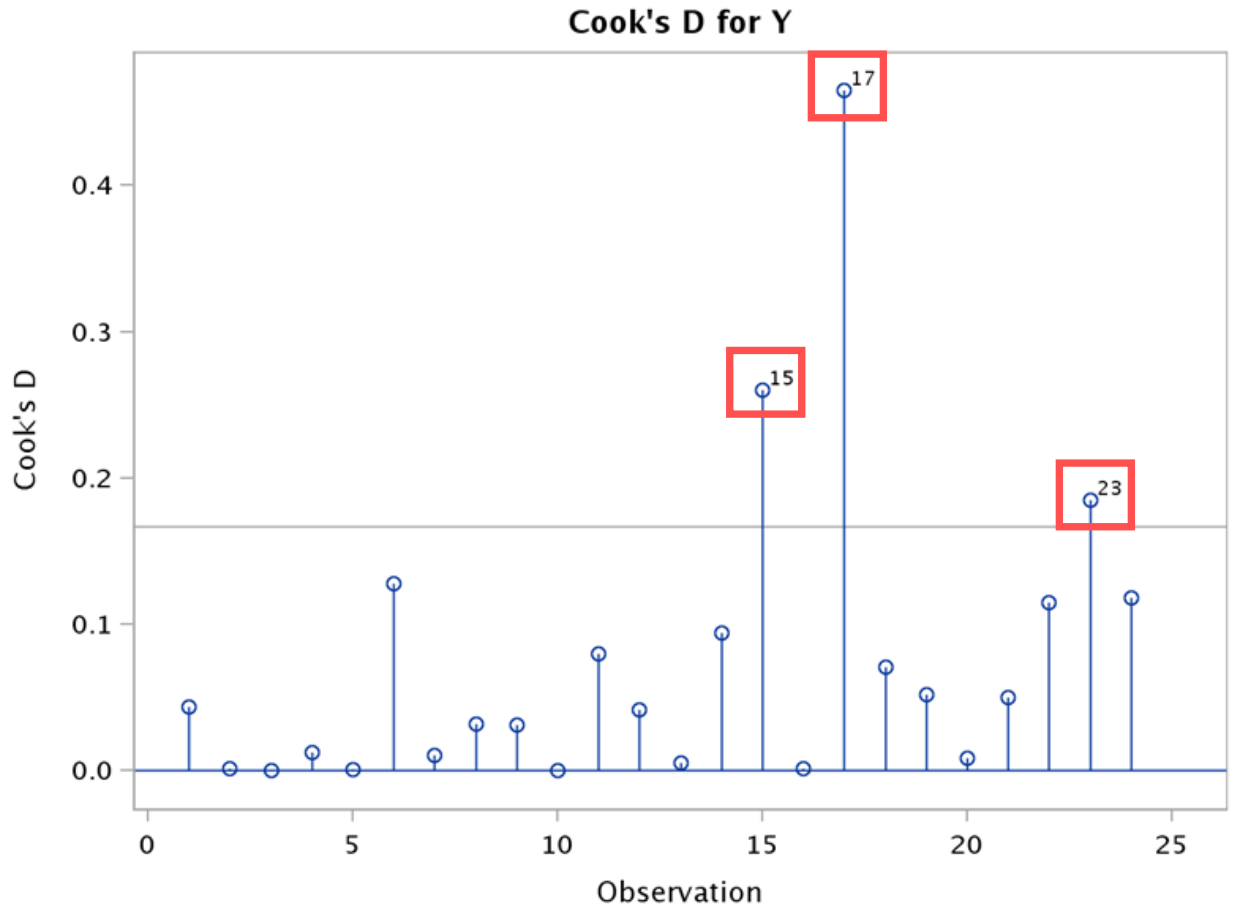

Figure 13 - Cook's Distance for Y

Additional residual analysis was conducted by plotting studentized residuals by leverage (Figure 14). Observation #17 was again identified as highly influential. The regression model may improve fit if this observations is removed from the model.

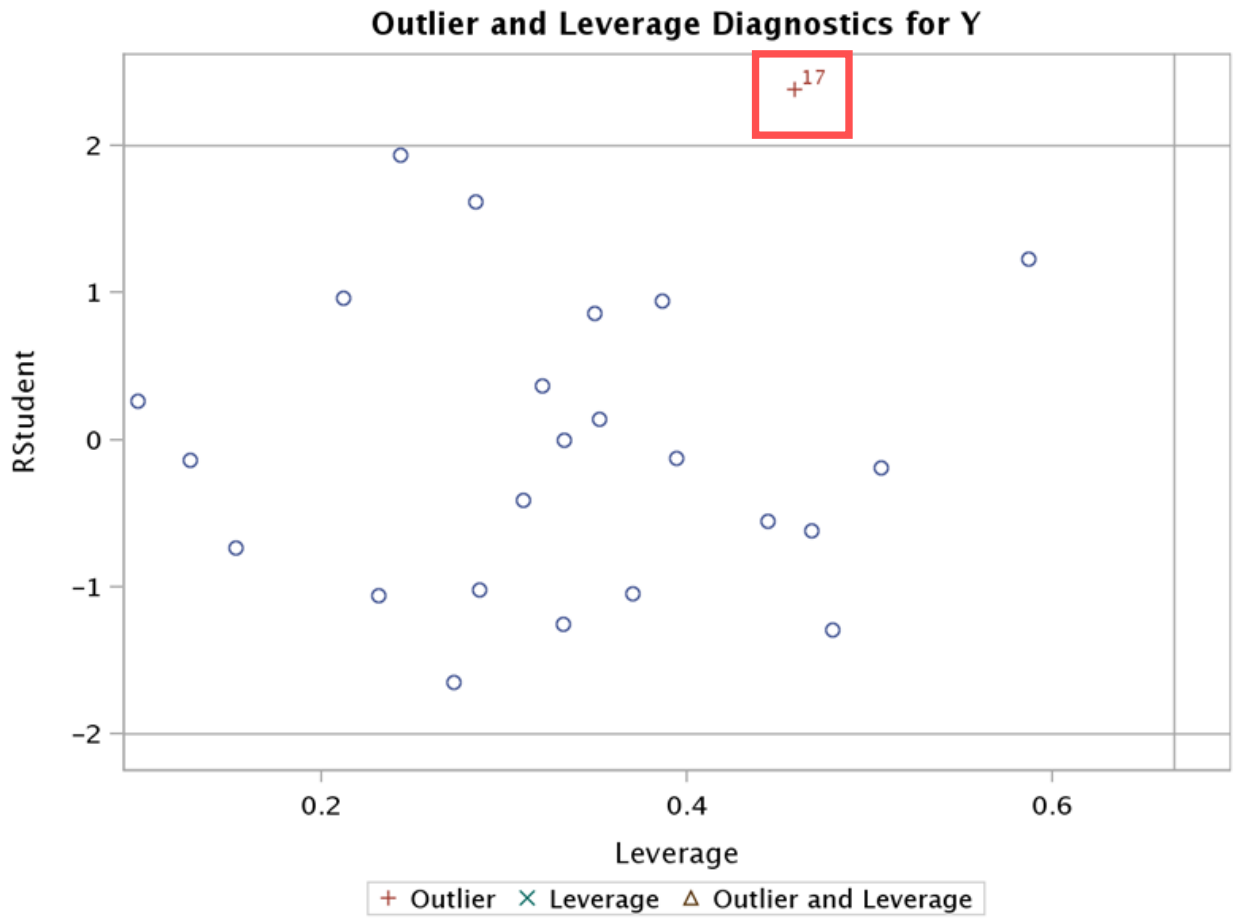

Figure 14 - Outlier and Leverage Diagnostics for Y

Multicollinearity

Multicollinearity exists when two or more predictors are highly correlated and this can result in inaccurate models. The variance inflation factor (VIF) is a method to evaluate multicollinearity. Values greater than 10 are considered large and values between 5 and 10 should be evaluated (Cody, 2011). If these conditions exist, then the violating predictor may need to be removed. All Variance Inflation Factors in Figure 15 are well under 10 and therefore multicollinearity does not appear to exist.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	16.59015	4.87745	3.40	0.0036	0
X1	1	2.21867	0.80405	2.76	0.0140	4.74426
X2	1	6.14082	3.80521	1.61	0.1261	2.46129
X4	1	2.86700	3.90116	0.73	0.4730	3.40594
X5	1	1.85534	1.23618	1.50	0.1529	1.63770
X6	1	-1.31636	1.21900	-1.08	0.2962	3.40995
X8	1	-0.04656	0.06067	-0.77	0.4540	2.12779
X9	1	2.25175	1.43232	1.57	0.1355	1.17696

Figure 15 - Model Coefficients and Variance Inflation Factor

Part 3 – Evaluating a Discrete Predictor

This next part of the analysis will fit a model with only predictors X1 (taxes in thousands of US dollars) and X2 (number of bathrooms). The model is fit by the equation below:

$$Y = 10.1120 + 2.7170X_1 + 6.09851X_2 \quad (5)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

Root MSE	2.82768	R-Square	0.7981
Dependent Mean	34.62917	Adj R-Sq	0.7788
Coeff Var	8.16562		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.11203	2.99614	3.38	0.0029	0
X1	1	2.71703	0.49115	5.53	<.0001	1.73656
X2	1	6.09851	3.22705	1.89	0.0727	1.73656

Figure 16 - Regression results using X1 and X2

The regression model assumptions were then evaluated by analyzing the residuals for the X1 and X2 variables as shown in Figure 17. The scatterplot for X1 shows that the residual error is uncorrelated with this predictor due to the random scatter of data points. This confirms that there is a linear relationship between X1 and Y. The scatterplot for X2 shows the residual error as a straight line when the predictor value is 1.0 and 1.5. This violates the linearity assumption indicating a non-linear relationship between X2 and Y. Based on this residual distribution, it appears that X2 is functioning as a discrete variable.

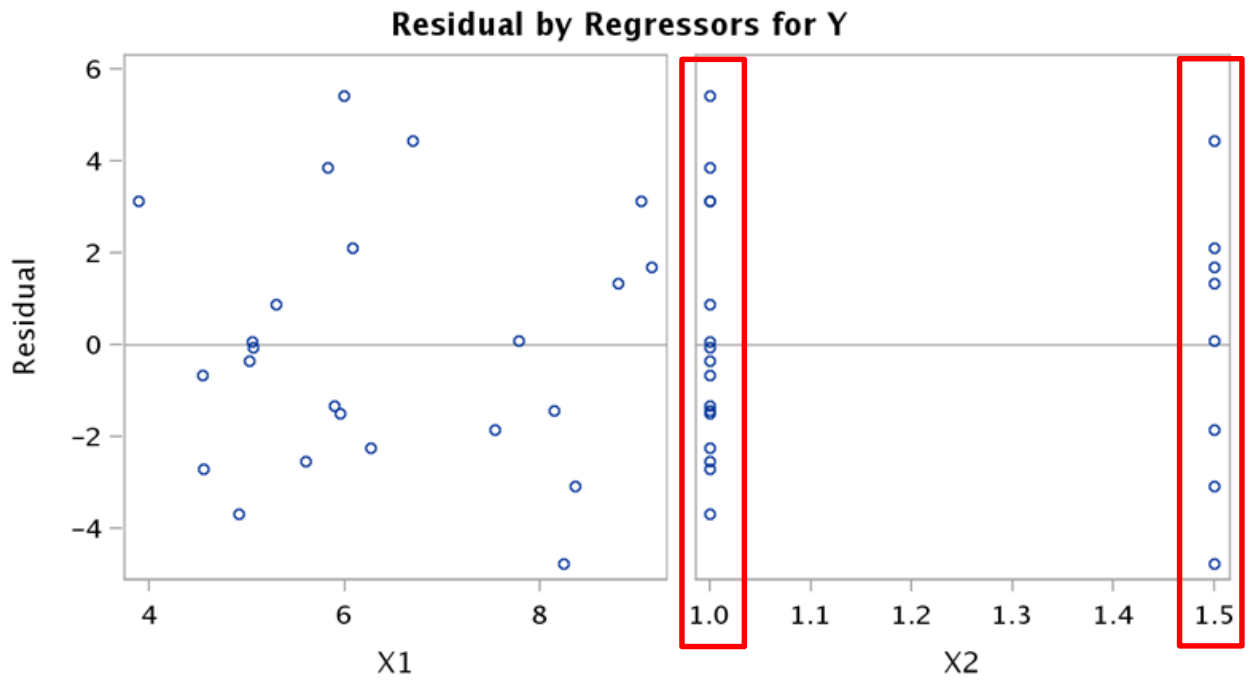

Figure 17 - Residual by Predictor X1 and X2

Given the discrete nature of predictor X2, a regression model was fit recasting X2 as a discrete variable. Handling the number of bathrooms as a discrete variable is reasonable given that we know this variable is not continuous in practice. Analyzing the building_prices data set confirms that the number of bathrooms is either 1.0 or 1.5. Under this scenario if a house has 1.5 bathrooms a dummy variable is set to 1 and then set to 0 for all other values (1 bathroom in this data set).

The comparison between the regression results is shown in Figure 18 with X2 handled as a discrete variable on the left and X2 handled as a continuous variable on the right. When a regression equation is used for prediction, the variables are selected with an eye toward minimizing the Mean Square Error (MSE) of prediction (Chatterjee & Hadi, 2013). Although the mean square error and adjusted R-square calculations are identical, the model that treats X2 as a discrete variable has less standard error for this variable (1.6135 vs. 3.2270). The intercept variable is also significant when X2 is handled as a discrete variable. The model that uses X2 as a discrete variable is thus preferred if choosing between the two.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

Root MSE	2.82768	R-Square	0.7981
Dependent Mean	34.62917	Adj R-Sq	0.7788
Coeff Var	8.16562		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	16.21054	2.88345	5.62	<.0001	0
X1	1	2.71703	0.49115	5.53	<.0001	1.73656
bath_dummy	1	3.04925	1.61353	1.89	0.0727	1.73656

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

Root MSE	2.82768	R-Square	0.7981
Dependent Mean	34.62917	Adj R-Sq	0.7788
Coeff Var	8.16562		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.11203	2.99614	3.38	0.0029	0
X1	1	2.71703	0.49115	5.53	<.0001	1.73656
X2	1	6.09851	3.22705	1.89	0.0727	1.73656

Figure 18 - Regression when X2 is handled as a discrete variable compared to continuous variable (X2)

A model is fit with only X2 (the continuous variable) and diagnostic plots are analyzed to confirm violations to the OLS assumptions. The regression model is formed using the regression results from Figure 19:

$$Y = 13.950 + 17.7250\, X_2 \qquad (6)$$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	418.90083	418.90083	22.34	0.0001
Error	22	412.60875	18.75494		
Corrected Total	23	831.50958			

Root MSE	4.33070	R-Square	0.5038
Dependent Mean	34.62917	Adj R-Sq	0.4812
Coeff Var	12.50593		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.95000	4.46398	3.13	0.0049
X2	1	17.72500	3.75049	4.73	0.0001

Figure 19 - Regression Results with X2 Only

This simple regression model violates a number of OLS assumptions. The linearity assumption is violated as shown in Figure 20 where the predicted value does not track the regression line Y.

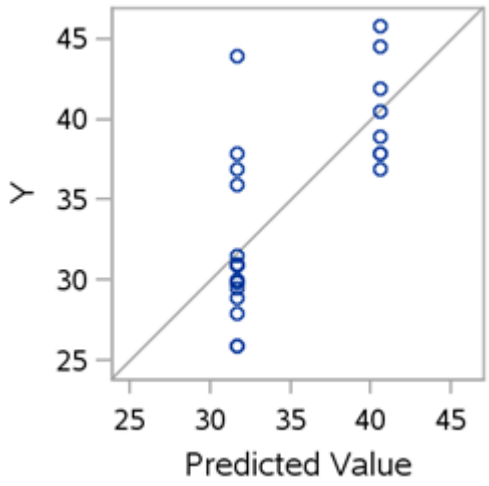

Figure 20 - Linearity Diagnostic for X2 only

The residuals by predicted value in Figure 21 also show a non-random scatter that confirms a non-linear relationship exists.

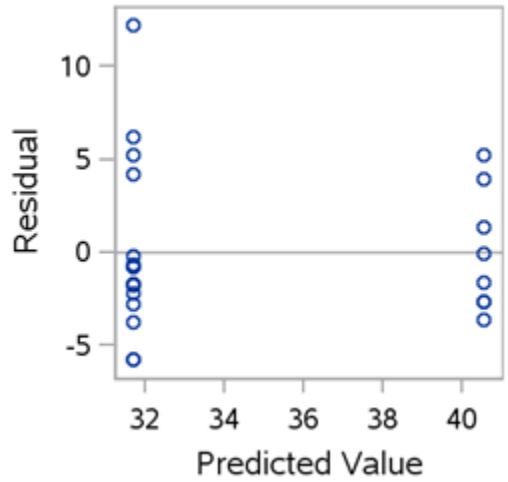

Figure 21 -Residual by Predicted Value

The residuals are also correlated to the predictor X2 shown in Figure 22 and this violates the constant variance assumption. Treating X2 as a continuous predictor variable clearly violates the OLS assumptions.

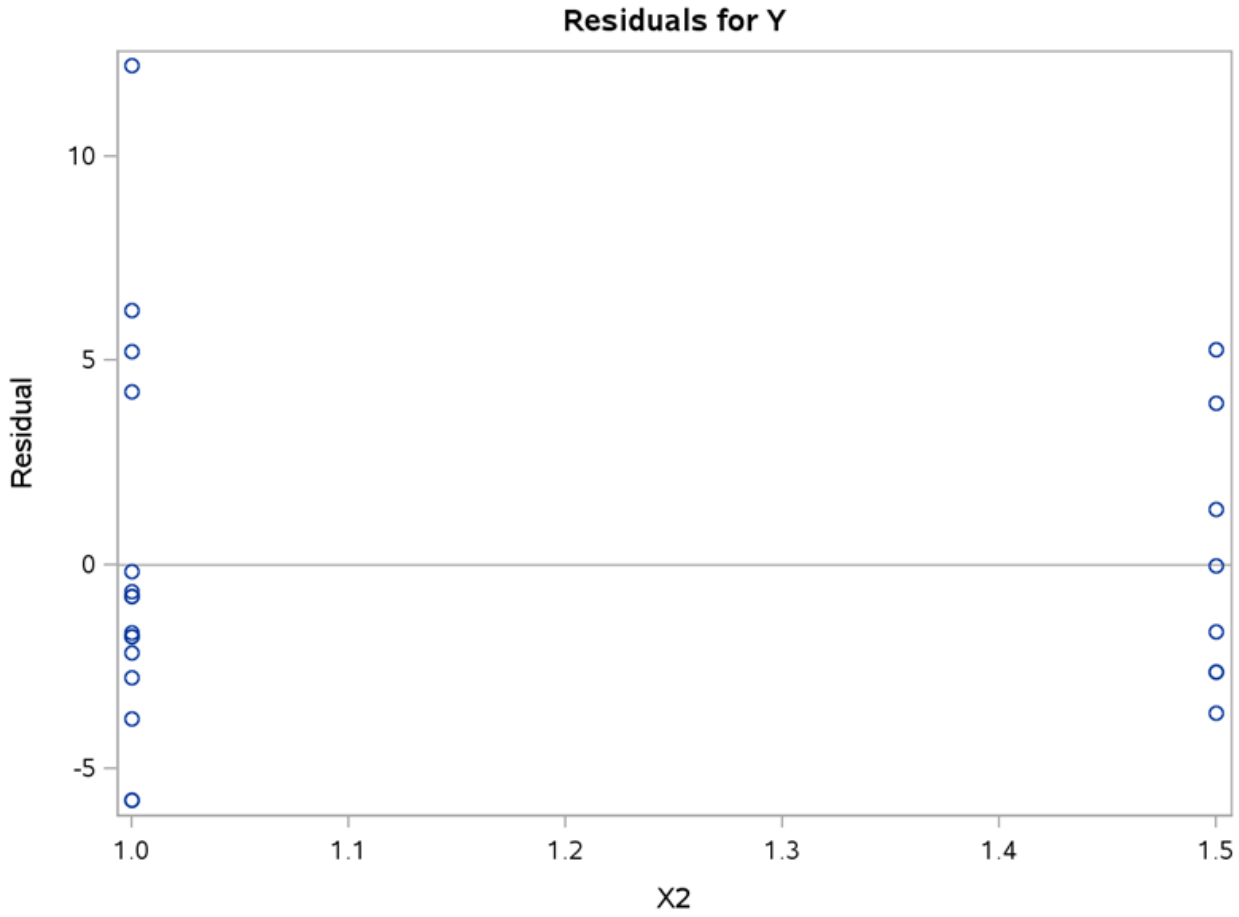

Figure 22 - Residuals by Predictor X2

Conclusions:

Part 1 of this analysis demonstrated that multiple regression models formed by automated variable selection procedures are more predictive than a simple regression model using the building_prices data set. The optimal model was selected using the Mallows Cp statistic and the additional predictors improved the R-square value from 0.7637 to 0.8491 when compared to the simple regression model.

In Part 2, the optimal model was then examined to evaluate adequacy using residual diagnostics and visualizations. The analysis confirmed normality of residuals although an assumption was violated for four predictors (X2, X5, X6, X9) that showed correlation to the residuals. A correlation between the predictors was not found. The analysis also identified three observations with a highly influential impact on the model. It is recommended that additional analysis be performed to understand the models results by removing those observation from the data set.

In Part 3, a model using X1 and X2 was examined to understand the discrete nature of the X2 variable (number of bathrooms) and its impact on the model. The analysis confirmed that handling the number of bathrooms as a discrete variable instead of a continuous variable reduced the variable standard error

and the intercept became statistically significant. A simple regression model of X2 confirmed that using X2 as a continuous violated the OLS assumptions.

Code:

```
/*      James Gray
        2013.07.11
        graymatter@u.northwestern.edu
        Assignment3_JG.sas
*/

/*      This code is for PREDICT 410 Assignment #3 - Multiple Regression Model. */

*************************************************************************************************;
* Get the data on the SAS server - mydata.building_prices - Regression by Example pg. 328-9
* Y = Sales price of the house (thousands of dollars)
* X1 = Taxes (thousands of dollars)
* X2 = Number of bathrooms
* X3 = Lot size (thousands of feet)
* X4 = Living space (thousands of feet)
* X5 = Garage stalls (#)
* X6 = Rooms (#)
* X7 = Bedrooms (#)
* X8 = Age of the home (years)
* X9 = Fireplaces (#)
*************************************************************************************************;
libname mydata '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
run;

*************************************************************************************************;
* Review building_prices dataset metadata and 5 observations;
*************************************************************************************************;
proc contents data=mydata.building_prices; run; quit;
proc print data=mydata.building_prices(obs=5); run; quit;

*************************************************************************************************;
* Fit a regression model using X1(Taxes) as the reference model;
*************************************************************************************************;
proc reg data=mydata.building_prices;
        model Y = X1; * generate the reference model using X1 as the predictor;
run;

*************************************************************************************************;
* Run an automated variable selection using the Forward procedure;
*************************************************************************************************;
proc reg data=mydata.building_prices;
model y = x1-x9 / selection=forward slentry=0.5;
run;
* Default value for SLENTRY in SAS is 0.50.  This is the p-value of the F statistic;
* Forward selection will fit "larger" models due to the entry criteria;

*************************************************************************************************;
```

```
* Run an automated variable selection using the Backward Elimination procedure;
*************************************************************************************************;
proc reg data=mydata.building_prices;
model y = x1-x9 / selection=backward slstay=0.1;
run;
* Default value for SLSTAY is 0.10.  This is the p-value of the F statistic;

*************************************************************************************************;
* Run an automated variable selection using the Stepwise procedure;
*************************************************************************************************;
proc reg data=mydata.building_prices;
model y = x1-x9 / selection=stepwise slentry=0.15 slstay=0.15;
run;
* Default values are 0.15 for both slentry and slstay;

*************************************************************************************************;
* Calculate metrics to select the optimal model;
* Mallows Cp - choose the first model where Cp is less than or equal to # of predictors;
* Akaike Information Criterion (AIC);
* Bayes Information Criterion (BIC);
* Models with smaller AIC and BIC are preferred (Regression by Example page 305)
*************************************************************************************************;
proc reg data=mydata.building_prices;
model y = x1-x9 / selection=cp aic bic best=4;
run;

*************************************************************************************************;
* Assess model adequacy for the optimal regression model produced by variable selection;
* The forward selection produced by the best model, fit the model;
* Using diagnostic plots and Variance Inflation Factors;
*************************************************************************************************;
ods graphics on;
* fit a regression and generate plots;
proc reg data=mydata.building_prices plots = (diagnostics residuals cooksd(label)
        RStudentLeverage(label));
        id Obs;
        * fit the model using the best model;
        model Y = X1 X2 X4 X5 X6 X8 X9 / VIF;
run;
ods graphics off;

*************************************************************************************************;
* Fit a model with X1 and X2 only, with X2 as a continuos variable;
* Then treat X2(bathrooms) as a discrete predictor value and run regression model;
*************************************************************************************************;
ods graphics on;
proc reg data=mydata.building_prices plots = (diagnostics residuals cooksd(label)
        RStudentLeverage(label));
        id Obs;
        * fit the model using the best model produced by variable selection X1,X2;
        model Y = X1 X2 / VIF;
run;
ods graphics off;
```

```
* Handle X2 as a discrete variable by converting it to a dummy variable;
data temp;
        set mydata.building_prices;
        if (X2=1.5) then bath_dummy=1;
                else bath_dummy=0;
run;

/* proc contents data=temp; run; quit; */
/* proc print data=temp (obs=24); */

ods graphics on;
* Fit model using X2 as a discrete var (bath_dummy);
proc reg data=temp plots = (diagnostics residuals cooksd(label)
        RStudentLeverage(label));
        id Obs;
        * fit the model using the dummy var;
        model Y = X1 bath_dummy / VIF;
run;
ods graphics off;

* Fit model with only X2;
ods graphics on;
proc reg data=mydata.building_prices plots = (diagnostics residuals cooksd(label)
        RStudentLeverage(label));
        id Obs;
        * fit the model using X2 only;
        model Y = X2 ;
run;
ods graphics off;

quit;
*************************************************************************************************;
* END
*************************************************************************************************;
```

References:

Cody, R. (2011). *Sas statistics by example*. (p. 145). Cary: SAS Institute.

Chatterjee, S., & Hadi, A. (2013). *Regression analysis by example*. (p. 302). Hoboken: John Wiley & Sons.