Derek Hughes
Assignment #3
Predict 410 – Sec 56

INTRODUCTION:

Here we are using our results from our EDA analysis and simple regression model from Assignment #1 and #2 to determine if this base model is the best overall model compared to when we use heuristic variable selection procedures. We then selected the best model by comparing the metrics of Residual Mean Squares, Adjusted R-Square, AIC, and BIC of each model. After we determine the best model, we test its validity by examining various plots of the data with the predicted and residual values from the model. This is to ensure the normal assumptions required for a linear regression model are upheld and, thus, we can be assured that the model will be unbiased and valid. Finally we test for outliers that may affect the results by using Cook's Distance and determine if the predictor variables are highly correlated (multicollinearity) or not. By analyzing the results of all of these graphs, plots, and metrics we can draw conclusions about the nature of our data and how well the fitted model predicts the observations (sales price of house).

RESULTS:

There were variances in models selected when using the heuristic variable selection processes Forward, Backward, and Stepwise.

The original model from assignment #2. Our original single variable regression was y = x1.:

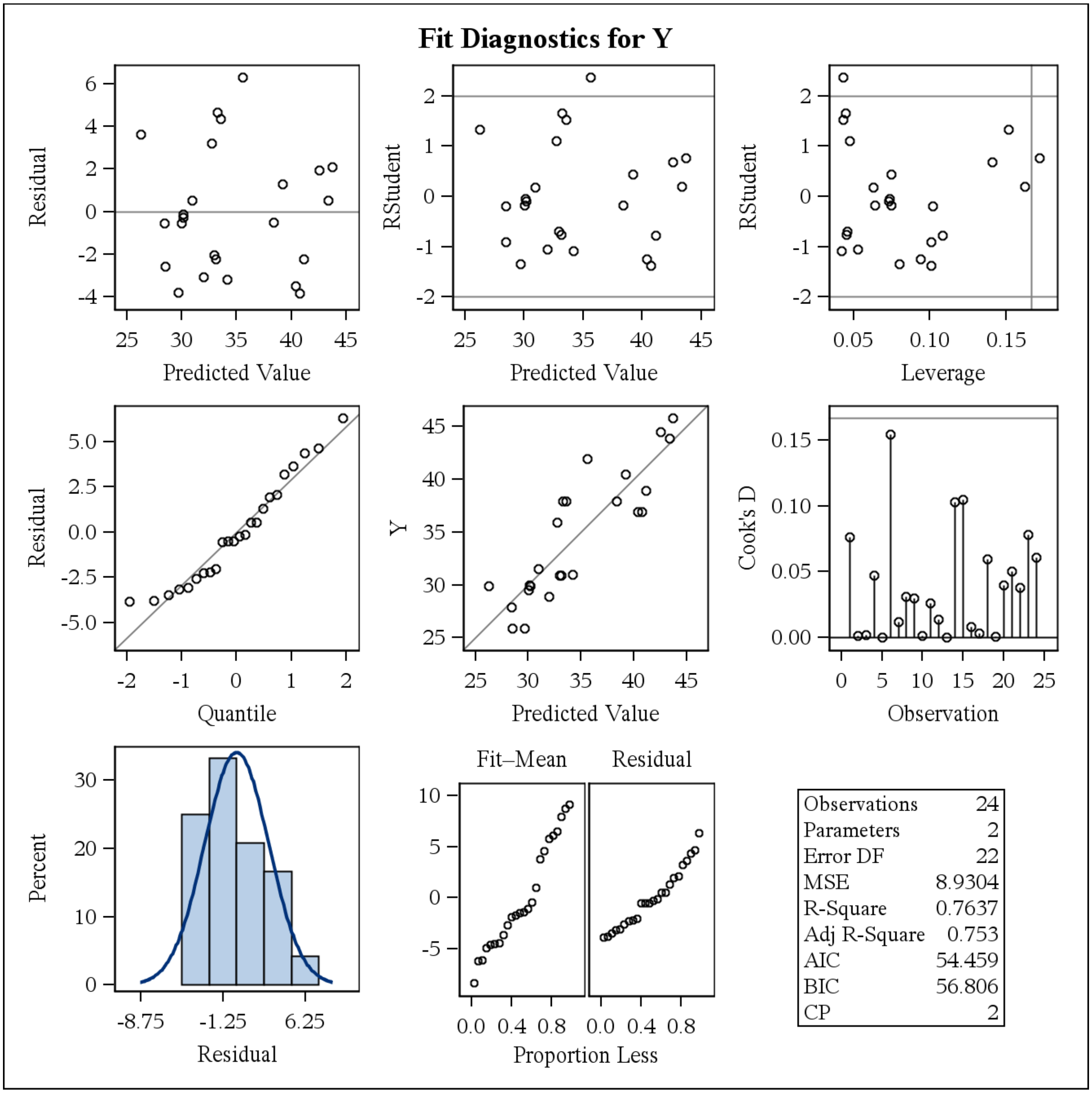
Number of Observations Read	24
Number of Observations Used	24

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

The REG Procedure
Model: original
Dependent Variable: Y



Observations	24
Parameters	2
Error DF	22
MSE	8.9304
R-Square	0.7637
Adj R-Square	0.753
AIC	54.459
BIC	56.806
CP	2

The **forward** model is shown below. It selected the model y = x1 + x2 + x4 + x5 + x6 + x8 + x9. The last step of each heuristic variable selection process is shown below in the interest of saving space.

Forward Selection: Step 7

Variable X4 Entered: R-Square = 0.8491 and C(p) = 6.2005

The REG Procedure
Model: forward
Dependent Variable: Y

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

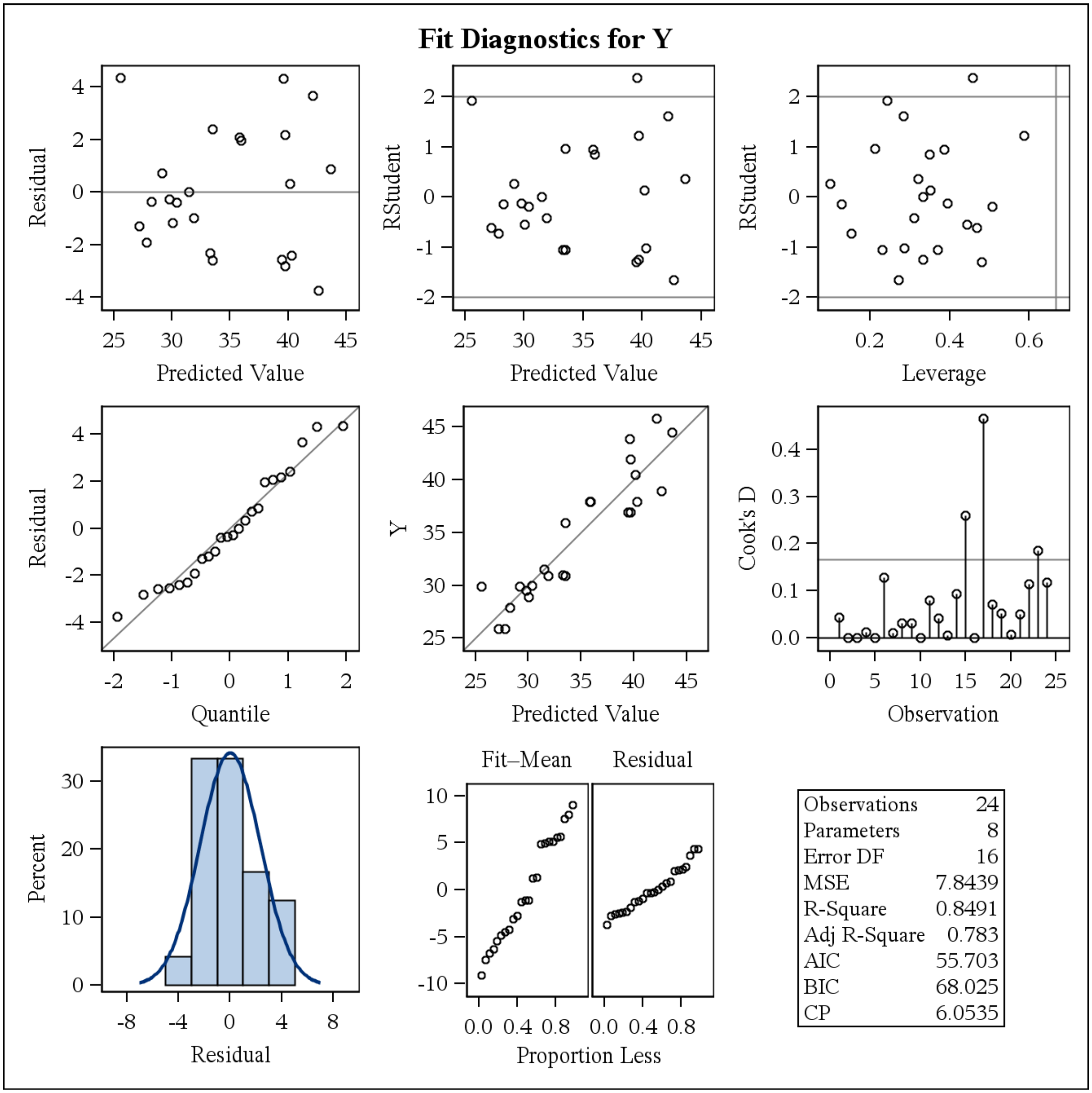
Bounds on condition number: 4.7443, 132.75

No other variable met the 0.5000 significance level for entry into the model.

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

The REG Procedure
Model: forward
Dependent Variable: Y

Backward Elimination: Step 6

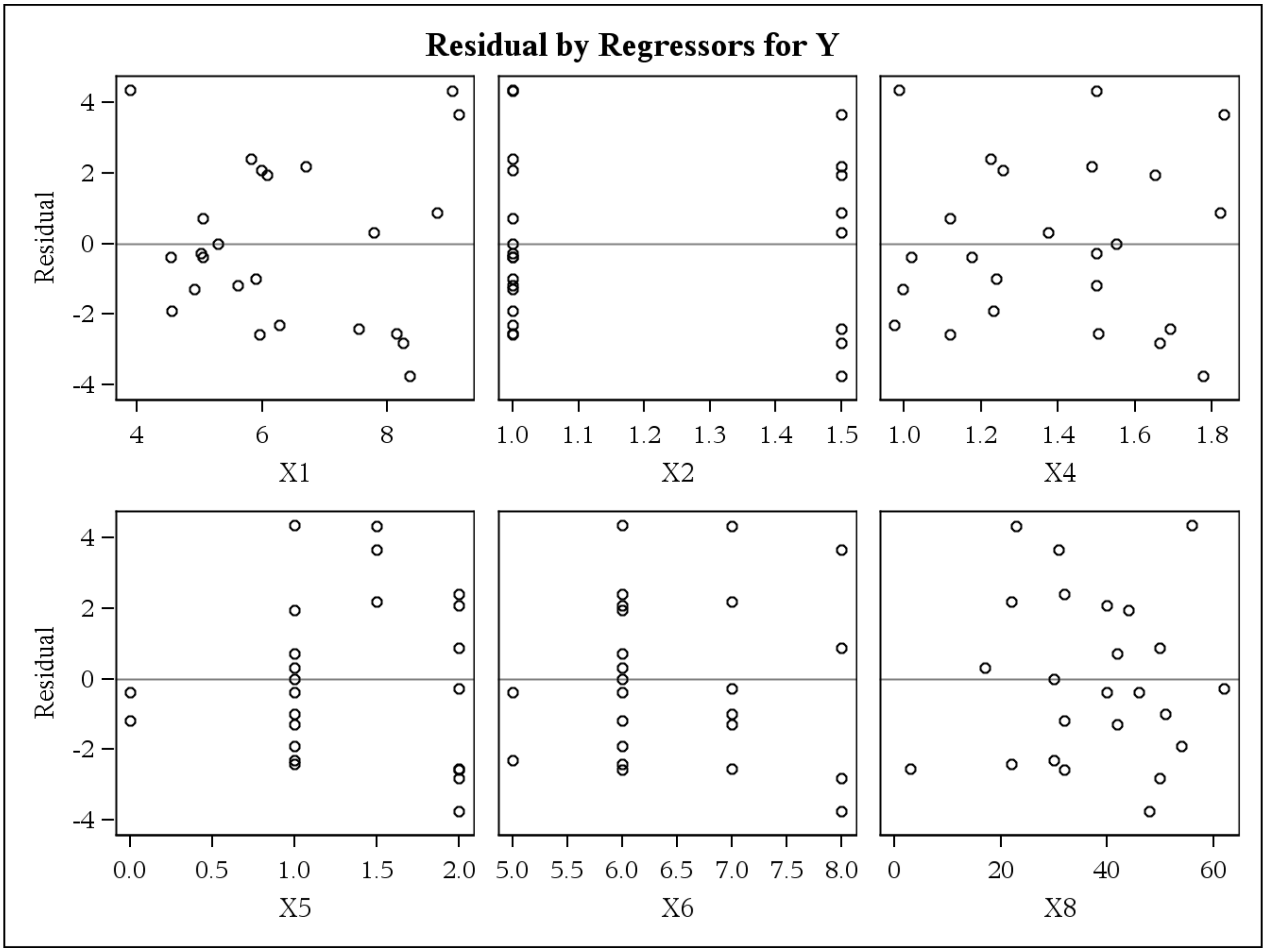


Fit Diagnostics for Y

Observations	24
Parameters	8
Error DF	16
MSE	7.8439
R-Square	0.8491
Adj R-Square	0.783
AIC	55.703
BIC	68.025
CP	6.0535

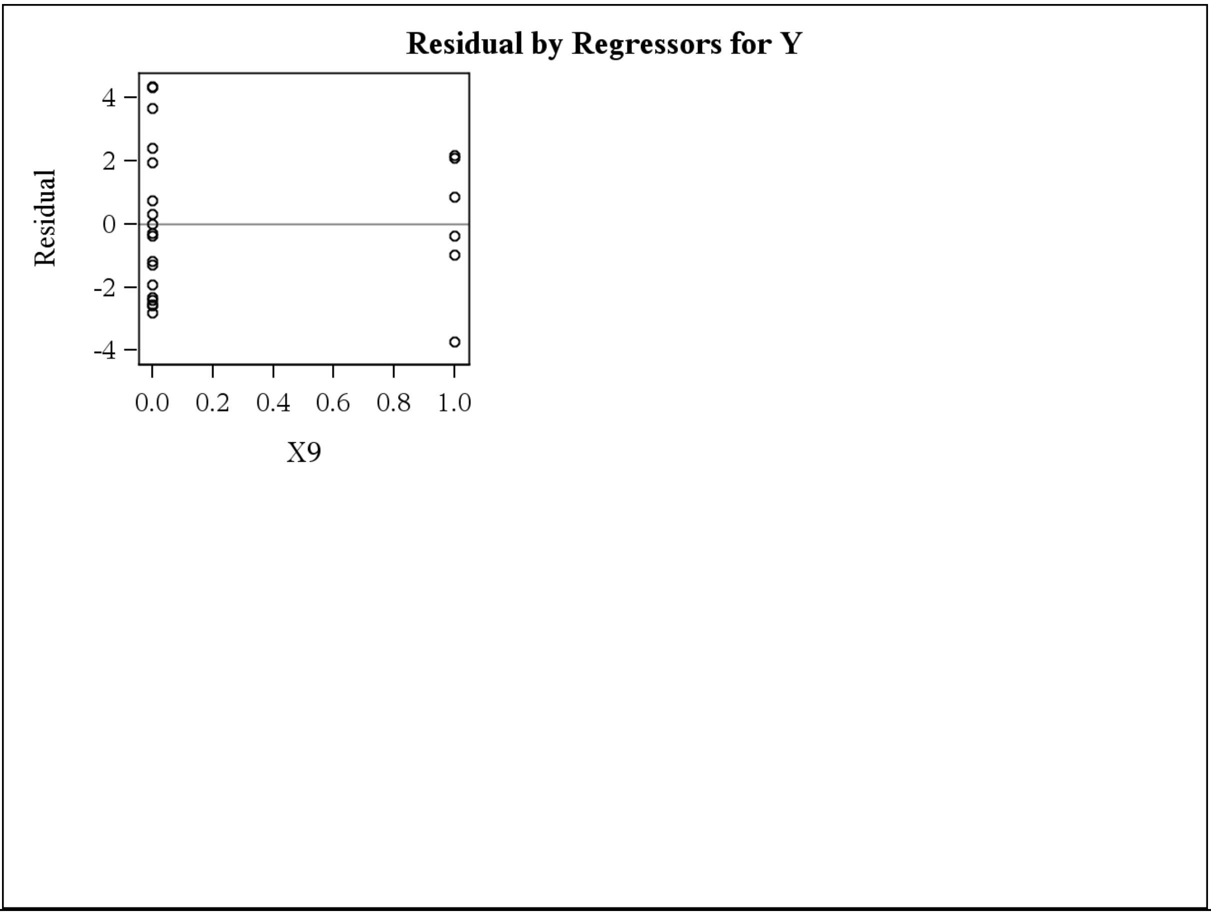
The REG Procedure
Model: forward
Dependent Variable: Y

Backward Elimination: Step 6



The REG Procedure
Model: forward
Dependent Variable: Y

Backward Elimination: Step 6



Stepwise and **backward** both selected the same model, y = x1 + x2. Only the last step of each process is shown below in the interest of saving space.

Backward Elimination: Step 7

Variable X7 Removed: R-Square = 0.7981 and C(p) = 0.9991

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			

The REG Procedure
Model: backward
Dependent Variable: Y

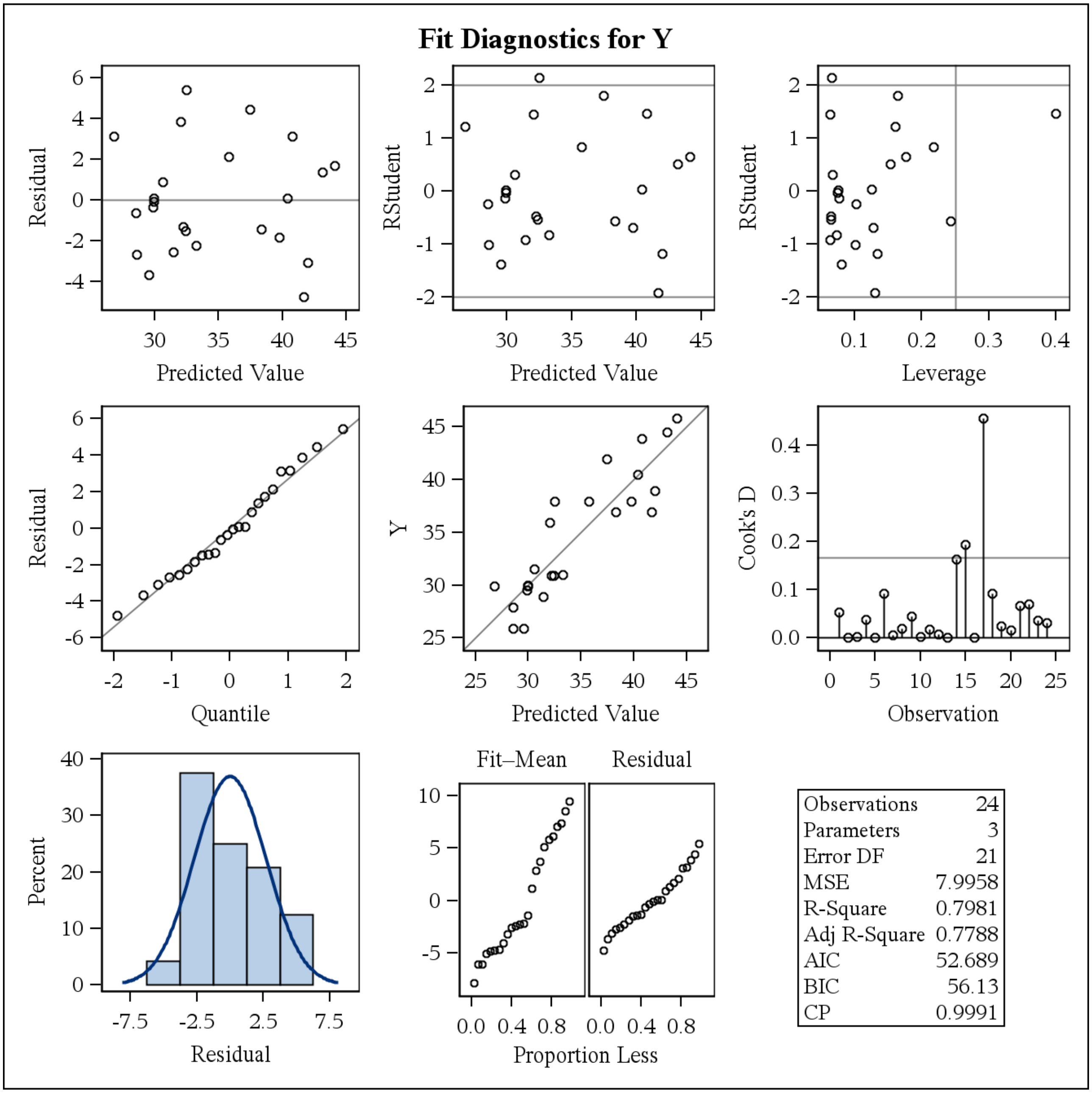
Backward Elimination: Step 7

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

Bounds on condition number: 1.7366, 6.9462

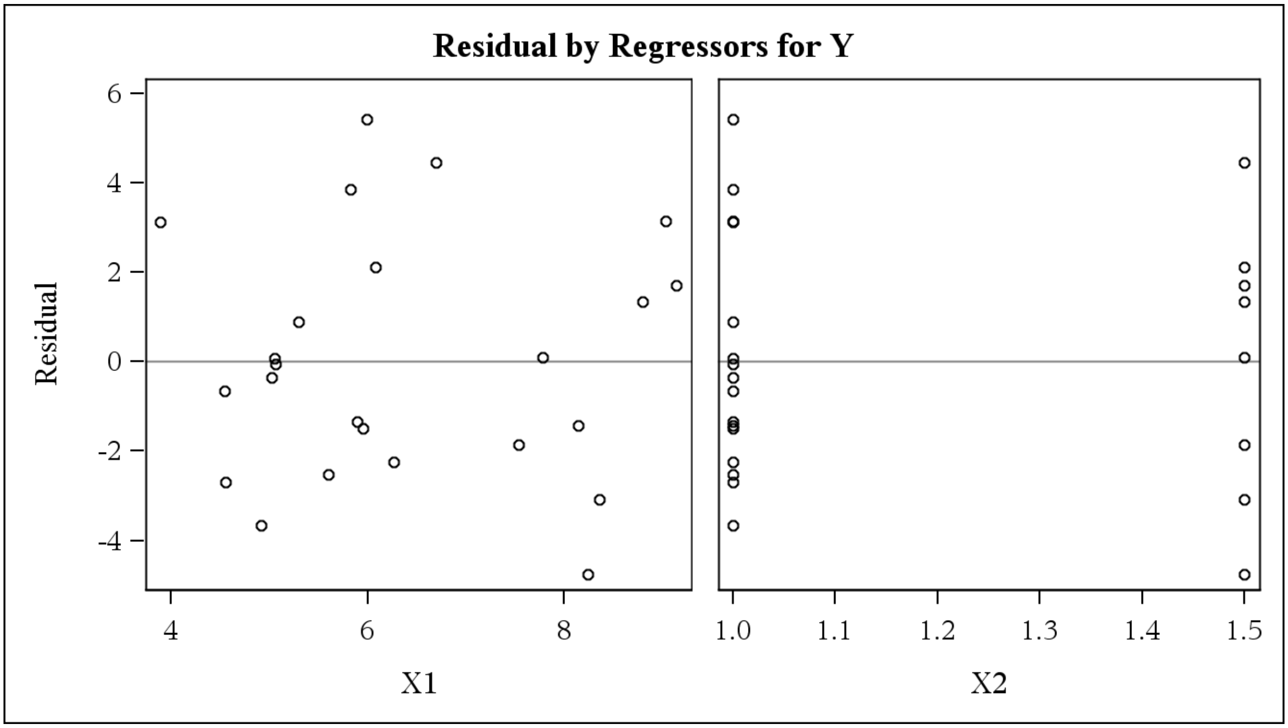
All variables left in the model are significant at the 0.1000 level.

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



Fit Diagnostics for Y

Observations	24
Parameters	3
Error DF	21
MSE	7.9958
R-Square	0.7981
Adj R-Square	0.7788
AIC	52.689
BIC	56.13
CP	0.9991



Stepwise process:

Stepwise Selection: Step 2

Variable X2 Entered: R-Square = 0.7981 and C(p) = 0.9991

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			

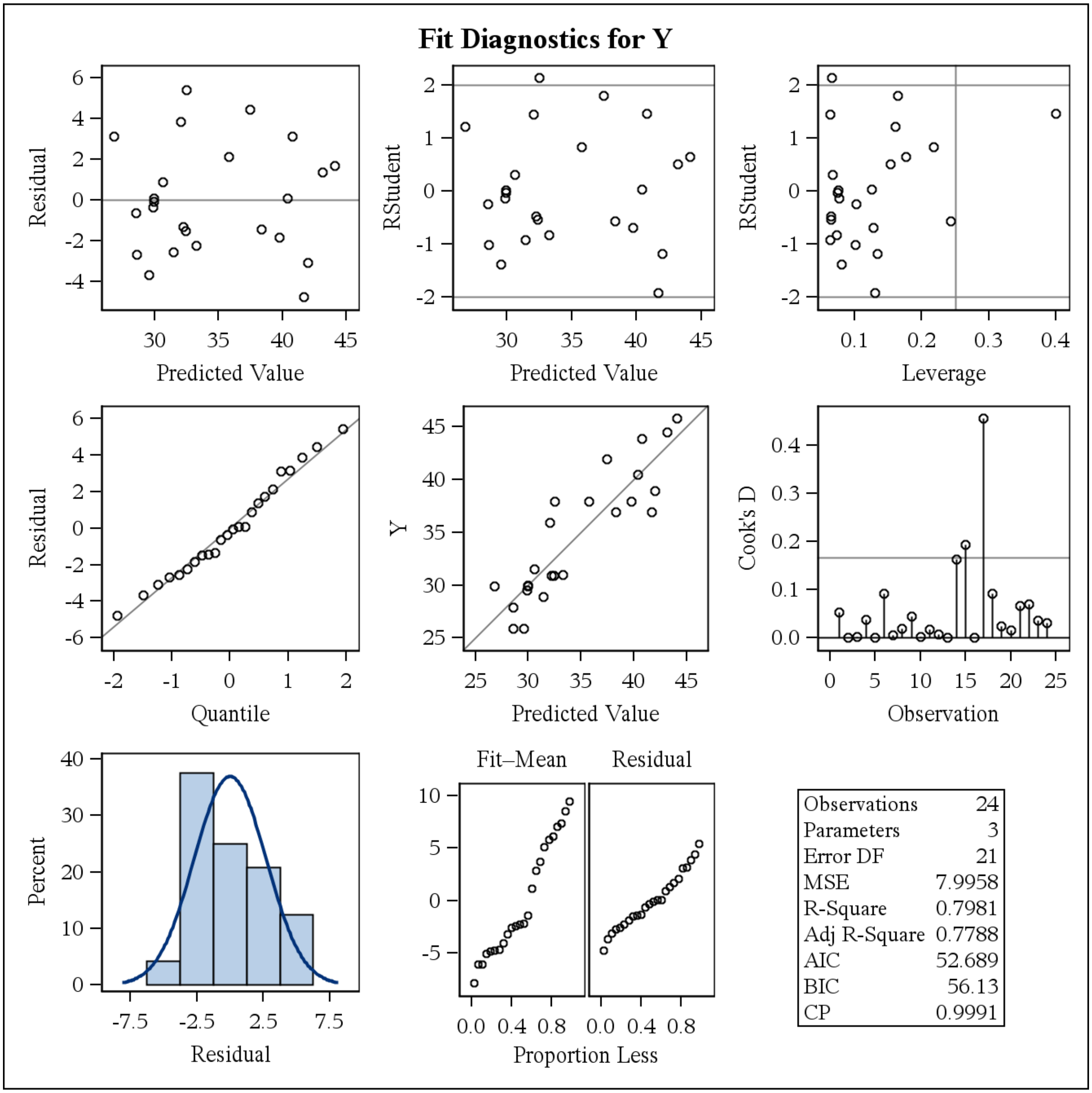
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	10.11203	2.99614	91.07817	11.39	0.0029
X1	2.71703	0.49115	244.69696	30.60	<.0001
X2	6.09851	3.22705	28.55593	3.57	0.0727

Bounds on condition number: 1.7366, 6.9462

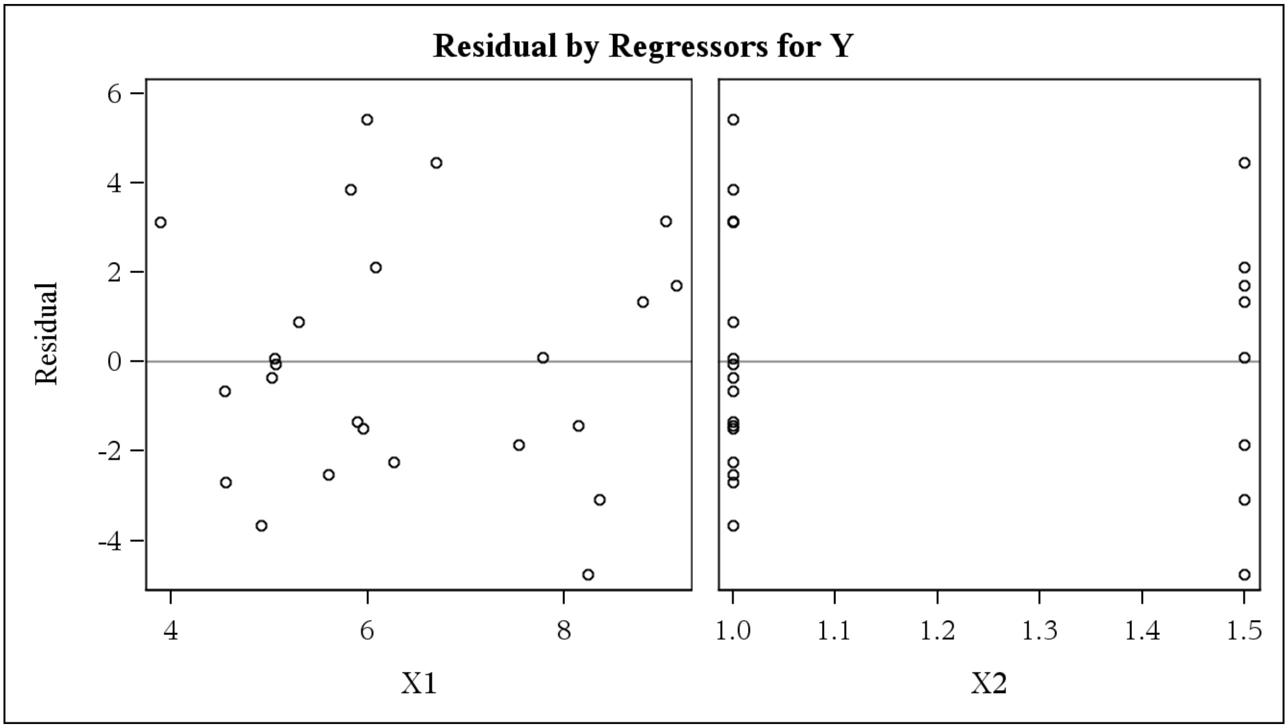
All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

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



Observations	24
Parameters	3
Error DF	21
MSE	7.9958
R-Square	0.7981
Adj R-Square	0.7788
AIC	52.689
BIC	56.13
CP	0.9991



Using the default values for SLENTRY and SLSTAY, which control the significance "limit" being used to determine if a process should add or remove a variable, with this data set cause the selected model to differ. One reason is that the default forward entry limit (SLENTRY) is set to 0.5, thus allowing any variable with a significant correlation to the Y at 0.5 or less to be added to the selection. With the backward selection process the default significance value (SLSTAY) to remove a variable is 0.1, which is more restrictive and should select fewer variables in its model. Stepwise default value, which uses a value for SLENTRY to add a variable and SLSTAY to remove a variable, is 0.15 for both values. Furthermore, Stepwise is unique in the sense that re-evaluates each variable in the model for significance after every new variable is added to it. This approach takes in account the effects of a new variable on the model and the new predictive strength of each variable in the model. Thus, some previously significant variables can be removed later. The forward and backward processes do not account for such effects and therefore can produce different models at times.

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	X1	Y	X2	X3	X4	X5	X6
1	original	PARMS	Y	2.98837	13.3553	3.32151	-1	.	.	.	.	.
2	forward	PARMS	Y	2.80070	16.5901	2.21867	-1	6.14082	.	2.86700	1.85534	-1.31636
3	backward	PARMS	Y	2.82768	10.1120	2.71703	-1	6.09851	.	.	.	.
4	stepwise	PARMS	Y	2.82768	10.1120	2.71703	-1	6.09851	.	.	.	.

Obs	X7	X8	X9	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_ADJRSQ_	_CP_	_AIC_	_BIC_
1	.	.	.	1	2	22	8.93035	0.76372	0.75298	2.00000	54.4587	56.8058
2	.	-0.046560	2.25175	7	8	16	7.84391	0.84907	0.78303	6.20050	55.7025	67.7914
3	.	.	.	2	3	21	7.99580	0.79806	0.77883	0.99906	52.6892	56.1300
4	.	.	.	2	3	21	7.99580	0.79806	0.77883	0.99906	52.6892	56.1300

Determining how predictive each model is compared to the other is a matter of evaluating the results of a few metrics designed to compare models. Our base model, y = x1 found in assignment #2, used R-squared to determine how well each predictor variable accounted for the variance in the model. We cannot use R-squared when comparing simple and multiple linear regression models because R-squared always increases with more variables; thus, we can use adjusted R-square as one metric because it compensates for more variables in the model.

We can also use Mallow's C(p), Residual Mean Square, and information criteria such as AIC and BIC to compare models.

Residual Mean Square is essentially the average squared error of the residuals of the model. The sum of squares between the predicted and observed values (the residual) divided by the number of observations minus the number of predictor variables produces the Residual Mean Square value. The lower this number is the better predictive ability of the model.

Mallow's C(p) is used because it takes into account both the bias of using an equation based on a subset of variables and the variance of the model. It is found by, again using the sum of squares of the error for the predictor variable divided by the estimated variance (often obtained from the linear model with the full set of variables) plus a bias term (2p - n). C(p) values that are most desireable are those that are closest to the number of variables in the model. For example, if the model has three variables, a C(p) value closer to 3 is more desired.

AIC and BIC are information criteria metrics that try to balance the conflicting demands of accuracy (fit) and simplicity (small number of variables). We can always add more and more variables to increase accuracy, even if it is a minor increase in accuracy, but additional variables create more complex models that take longer to perform as well. Performance and efficiency are real issues in the real world so AIC and BIC try to rank models on the basis fit and simplicity at the same time, with the BIC process applying a stronger penalty for the amount of variables in the model. The actual value of AIC or BIC is irrelevant except for comparison purposes as lower values are more desirable. Also, if the difference of AIC between models does not differ by more than 2 then those models should be treated as equally adequate.

Of the four models, I selected the model found with the stepwise and backward process (y= x1 + x2) as the best model. Using adjusted R-square, this model was the second highest (.7788) while the forward process model (y= x1 + x2 + x4 + x5 + x6 + x8 + x9) had the highest adjusted R-square at .783. The stepwise and backward process model also had a residual mean square error of 7.996 compared to 7.844 from the forward model. The original model had the lowest adjusted R-square value (.7637) and largest residual mean square error (8.930) of the models. Finally, the AIC and BIC values for the stepwise and backward model were 52.689 and 56.130 respectively. For the original model the AIC was 54.459 and BIC 56.806 while the forward model produced an AIC of 55.703 and BIC of 67.791.

While the stepwise/backward model did not have the highest adjusted R-square and residual mean square error, it was only marginally lower than the highest adjusted R-square value (.783 to .779) and marginally higher MSE (7.996 vs 7.844). However, the AIC and BIC values really tell the story between these models as the stepwise/backward model has significantly better AIC value (52.6892 vs. 55.7025) and an exceptionally better BIC value (56.130 vs. 67.7914). Any AIC/BIC values greater than 2 units is considered a significant difference between the quality of the models that are trying to balance the traits of accuracy (fit) and simplicity (small number of variables). These values show us that the marginally better values of the forward model for adjusted R-square and MSE are most likely due to having a few extra variables contributing to allow it to have slightly better accuracy (fit). Yet, those additional variables make the model more complex and divergent from the principle of parsimony. The original model does much better with its AIC (54.4587) and BIC (56.8058) values and they are technically within 2 units of the stepwise/backward model but not by much and, combined with the original models lower adjusted R-square and higher MSE values, I felt the stepwise/backward model was the best overall model according to the metrics of MSE, adj R-Square, AIC, and BIC.

******* PART 2 ***** - TESTING BEST MODEL SELECTED**

Number of Observations Read	24
Number of Observations Used	24

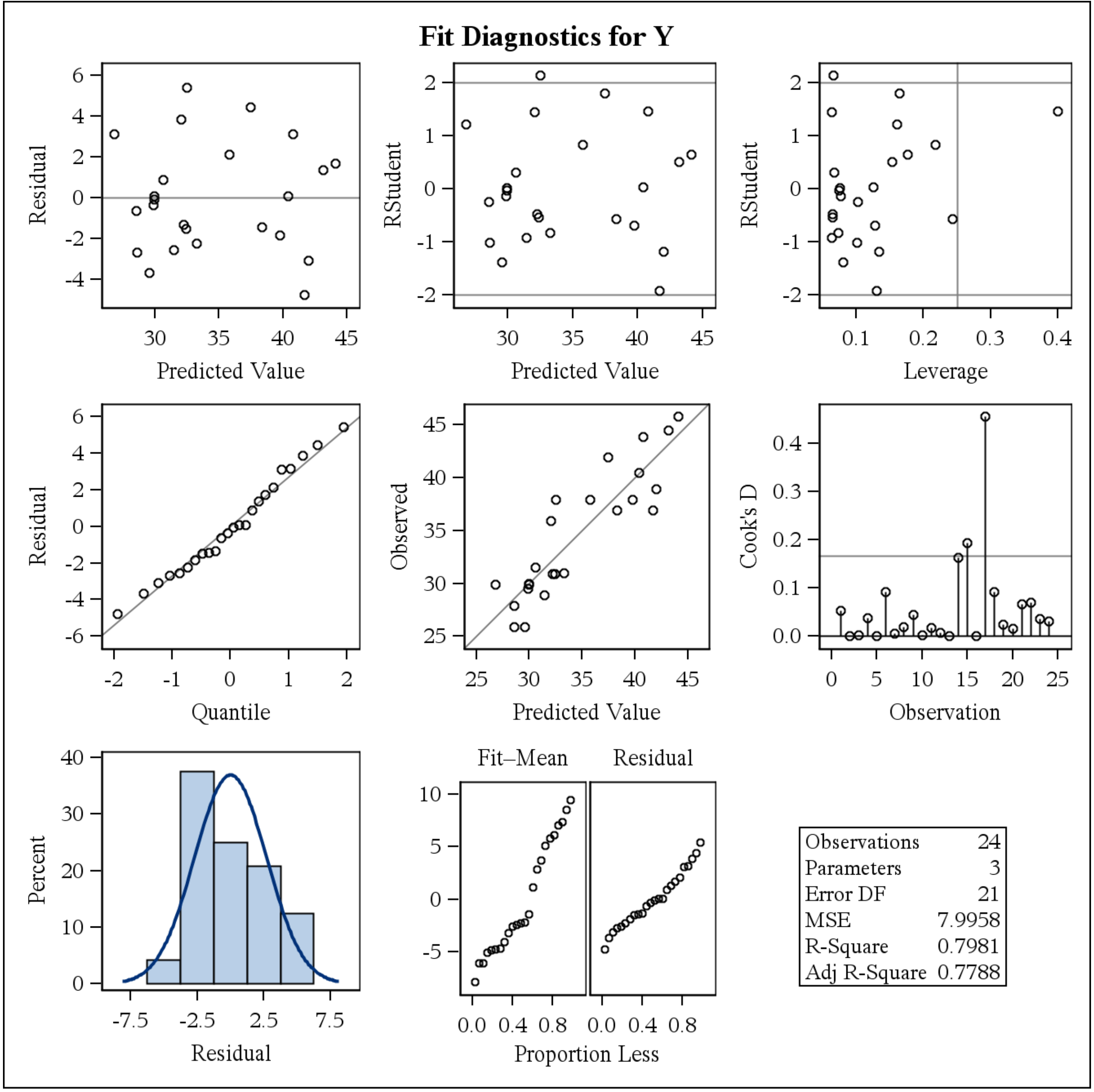
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	663.59779	331.79890	41.50	<.0001
Error	21	167.91179	7.99580		
Corrected Total	23	831.50958			

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

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	1	10.11203	2.99614	3.38	0.0029	0
X1	Taxes in thousands of dollars	1	2.71703	0.49115	5.53	<.0001	1.73656
X2	Number of bathrooms	1	6.09851	3.22705	1.89	0.0727	1.73656

We always want to test the statistical significance of the model and the coefficients of the predictor variables if they equal zero. We can see the F value to be 41.50 with an alpha significance of much less than .0001. This allows us to reject the null hypothesis that the model’s correlation is equal to zero. Since we reject this hypothesis it indicates the model is significantly correlated to the response variable values. We can do the same with the predictor variables. With X1 the t-value is 5.53 with a p-value of less than .0001. This allows us to reject the null that X1’s correlation with Y is significant.

For X2, the t test value is 1.89 with another 0.0727 alpha value that allows us to fail to reject the null for X2 and indicate that X2 does not have a significant correlation with the response variable when used within this model. It should be noted that it is important to check variables selected by the heuristic processes when using the default values in some programs. In this case, the default “cutoff” value for selecting variables for the Backward and Stepwise processes was 0.1 and 0.15 (for adding and removing) respectively. If we used a more restrictive cutoff value, such as the standard alpha value of 0.05 then the model may have just selected X1 as the only predictor variable in both processes. We use a higher than normal cutoff value so we have a higher chance of evaluating more variables from the data set when conducting the heuristic variable selection processes.



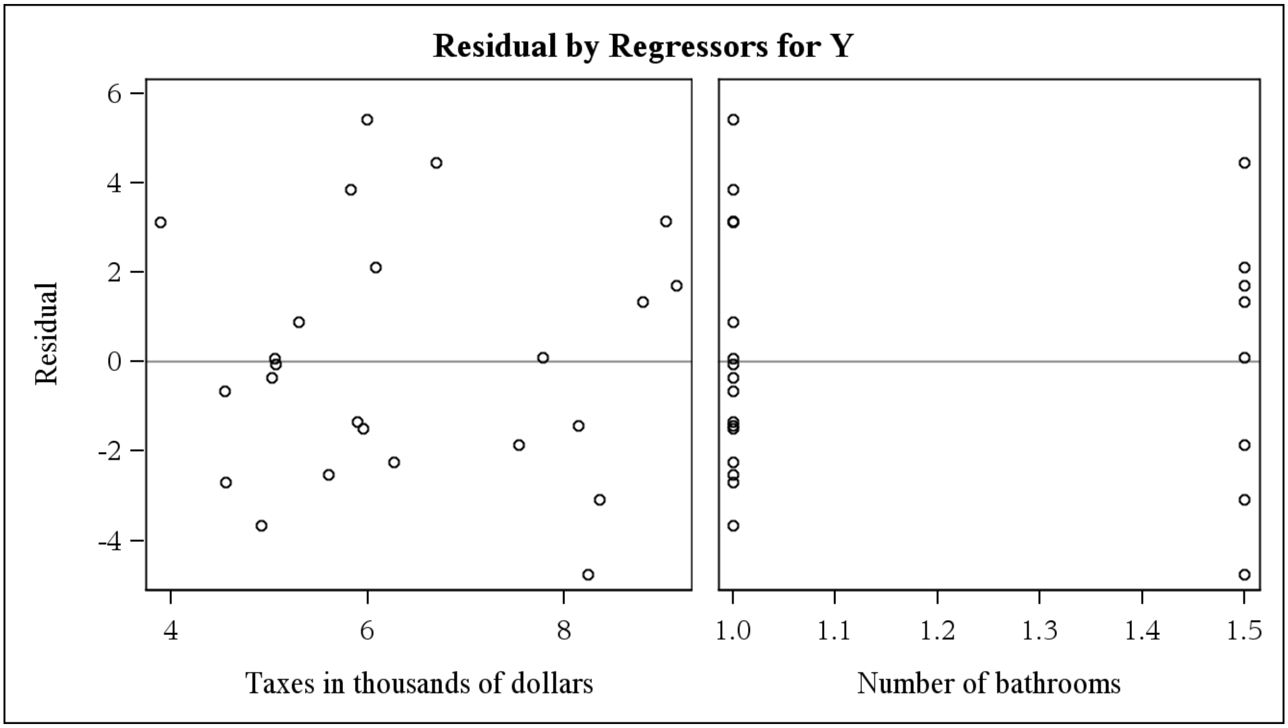
Observations	24
Parameters	3
Error DF	21
MSE	7.9958
R-Square	0.7981
Adj R-Square	0.7788

a) The regression line for the model can be overlaid on a scatterplot of the predicted values versus the observed values. By overlaying the regression line we can determine if the plots are scattered in a linear fashion around the line.

Furthermore, we can see the precision of the model by observing how far away the observed plots are from the regression line (predicted values). The closer they are to the line, the better the model accounts for the variation between the observed and predicted values on average. This is reflected in the adjusted R-square value where a higher adjusted R-square value is a good indicator the model is a better fit for the data.

Here we can see that the plots follow a wide but linear pattern as the plots stay fairly evenly distributed around the regression line at lower and higher predicted values. This is reflected by the adjusted R-square value of .7788 for the model, which indicates that the variables account for 77.88% of the variance in the model. While the adjusted R-square value is fairly strong, we need to ensure our model and its variables meet all of the assumptions for a linear regression model.

b) The Q-Q plot for the residual of the response variable allows us to verify the normal assumption for the errors. If the residual plots are very close to the line then this indicates that they come from the same normal distribution and can be used for further analysis. Residual plots not following the straight line may indicate another model type is needed such as quadratic or logarithmic. In this case, the residuals plotted against the quantile values indicate a fairly straight line, but we can see that the most of the middle quantile values tend to all be slightly below the line while the values in the far left and right quantiles tend to all be above the line. The overall distribution is acceptable for verifying the normality assumption has been met, but the aforementioned pattern, although subtle, may be worth remembering as it may indicate an interesting interaction of the variables.



c) We can plot the residuals of the response variable against the predictor values to see if any non-random patterns appear. With the first variable (X1 - taxes in thousands of dollars), the residual plots are scattered randomly about the zero mean value to indicate uniform error variances. In layman terms, the X1 value as a predictor predicts every value equally well, not just some of the values.

In the case of the second variable, the plots appear to be scattered vertically around the values of 1.0 and 1.5. This indicates the values of the predictor variable are probably discrete or categorical and only appear in .5 increments. With this discrete variable (being treated as a continuous variable), we can see that the residuals still tend to be distributed symmetrically around the zero mean of the residuals (although the distribution is a bit heavier below zero at the 1.0 number of bathrooms), only at the discrete values of the predictor variable. Symmetrical distribution of residuals around zero mean is what we want because it indicates a uniform randomness of error for the model for each value of the predictor variable. Thus, x2 (number of bathrooms) as a predictor, shows it can be used with our model as well.

d) Cook’s distance is also displayed. This test allows us to determine how much a single observation affects the response variable if that observation was to be removed. Some values can have a strong influence because they are abnormally large or small so it is important to check if any values stand out as outliers. Generally speaking, any value greater than 1 is considered to be influential and worth exploring further (i.e. - is it an outlier? mistaken entry? abnormal situation?). We can see that one observation (17) has a much higher Cook’s D values than the others. This means it will have a stronger influence if removed, but despite being more extreme than the other observations, it is not so different that it has a strongly significant influence because it is still well below 1.
The #14 and #15 observations are also slightly more influential than the remaining observations, but, again, neither is more than a Cook's distance of 1.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	10.11203	2.99614	3.38	0.0029	0
X1	Taxes in thousands of dollars	1	2.71703	0.49115	5.53	<.0001	1.73656
X2	Number of bathrooms	1	6.09851	3.22705	1.89	0.0727	1.73656

e) Multicollinearity is when there is a strong correlation between variables. When it exists, it inflates the coefficients of the relevant variables in the model. This is not desired because it means we are using two variables that are very similar to each other to predict the same thing. In other words, one of them is not necessary. Multicollinearity describes this and it is our responsibility to determine which, if any, variables are related to each other in this way and remove the variable that has the least predictive power of the two variables. We test this by using the Variance Inflation Factor (VIF). Traditionally any variable with a VIF of greater than 10 was said to have collinearity, but often a VIF value of 5 or as low as 3 is used. Here we will use VIF of 5 as our cutoff value to indicate collinearity of the variable.

Our two predictor variables in the model both show a VIF of 1.737 which means their coefficients are inflated by a factor of 1.737 because of the amount they are correlated with each other. This is a perfectly acceptable VIF value and shows there is an insignificant amount of collinearity between the two variables. If there was a significant VIF of a few variables indicating multicollinearity then we could review the correlations between each variable in the model to determine which variables are highly correlated. Afterwards, we would decide between the highly correlated variables which to removed based on scientific or practical means.

PART 3* - Evaluation of X2 and bath_dummy

a) Analysis with continuous **X1 and X2 variable**

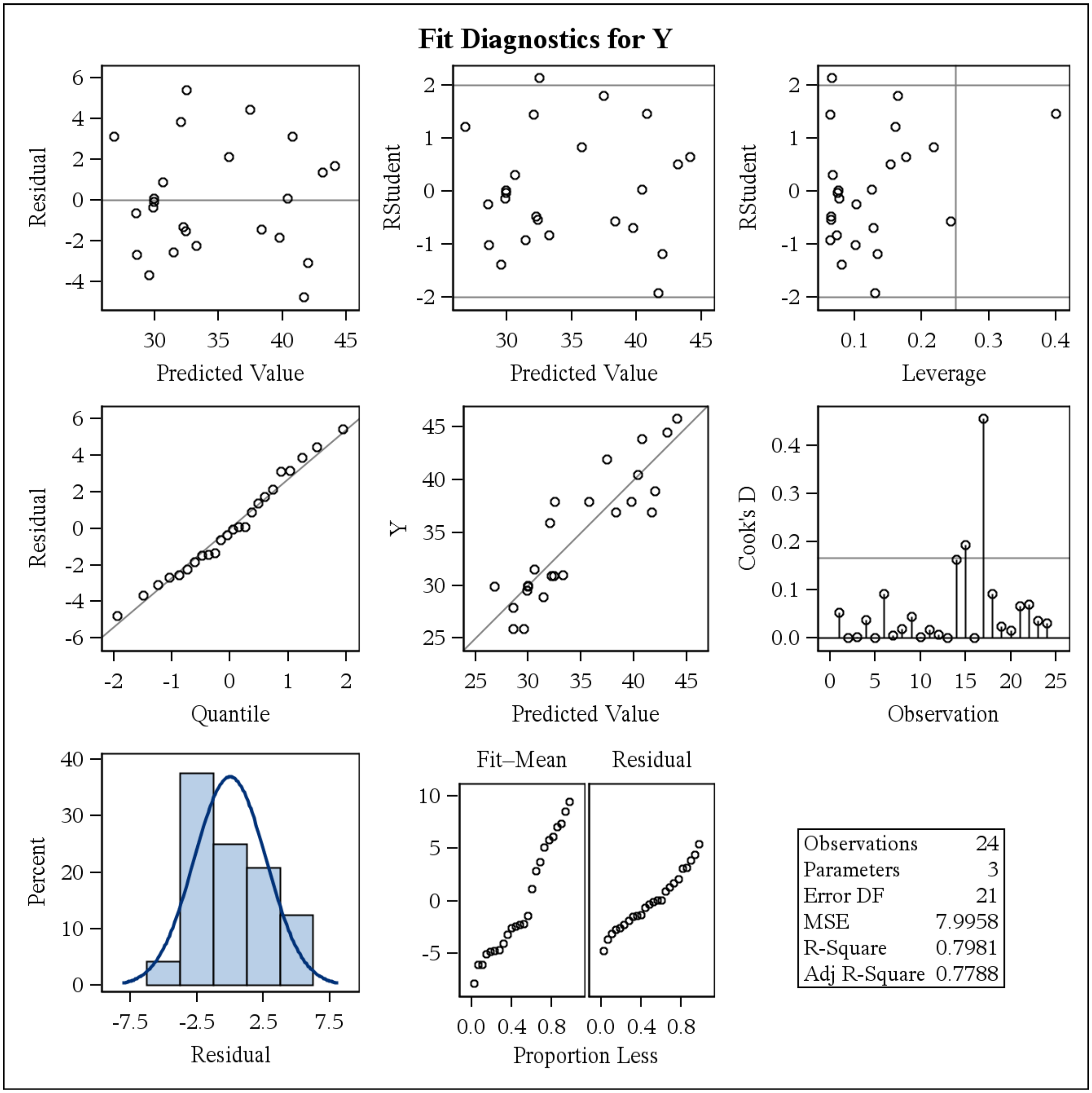
Number of Observations Read	24
Number of Observations Used	24

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

Simple Linear Regression with X2



X2 is a discrete or categorical variable based on the values of 1 or 1.5. Our models can handle a predictor variable being continuous or discrete in some situations, but a linear regression is mainly designed for continuous variables. It may require additional testing of both approaches (using X2 as continuous or discrete) to see if there is a significant difference in the resulting models.

Fitting a model just using X2 allows us to test the OLS assumptions that treating X2 as a continuous predictor variable could violate the OLS model assumption of linearity or normality. We will explore any differences between bath_dummy and X2 in a simple linear regression.

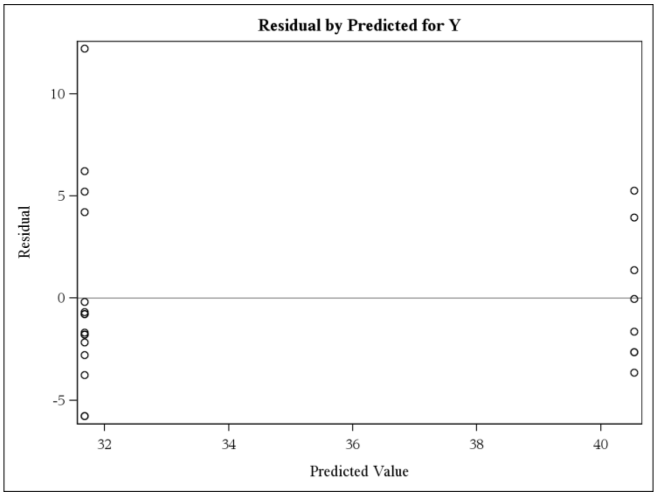
Simple Linear Regression with X2

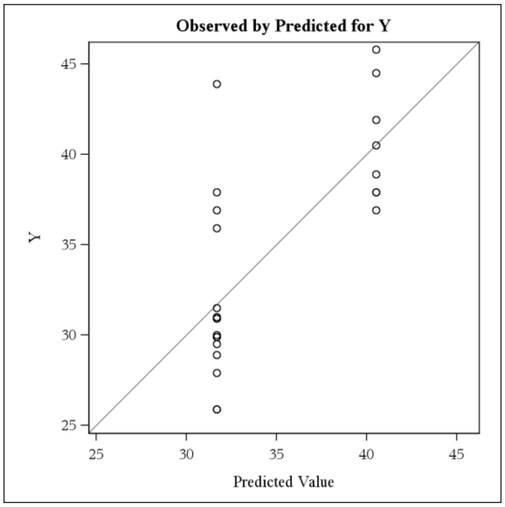
Number of Observations Read	24
Number of Observations Used	24

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			

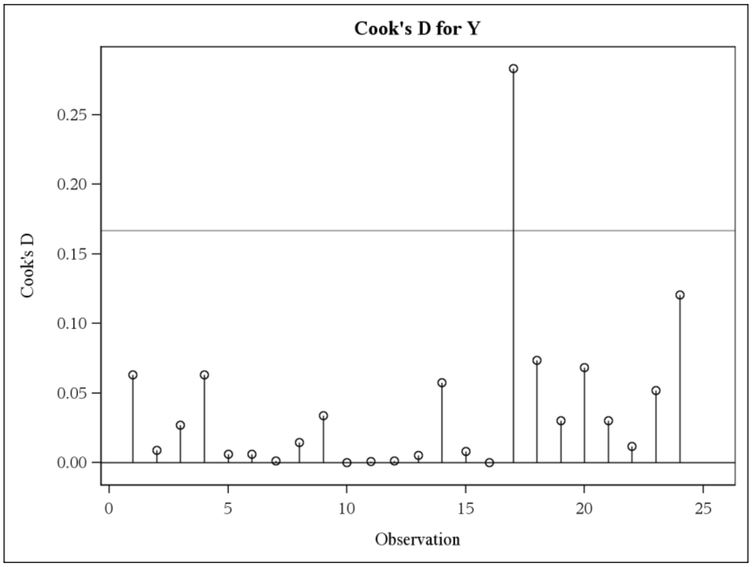
Root MSE	4.33070	**R-Square**	0.5038
Dependent Mean	34.62917	**Adj R-Sq**	0.4812
Coeff Var	12.50593		

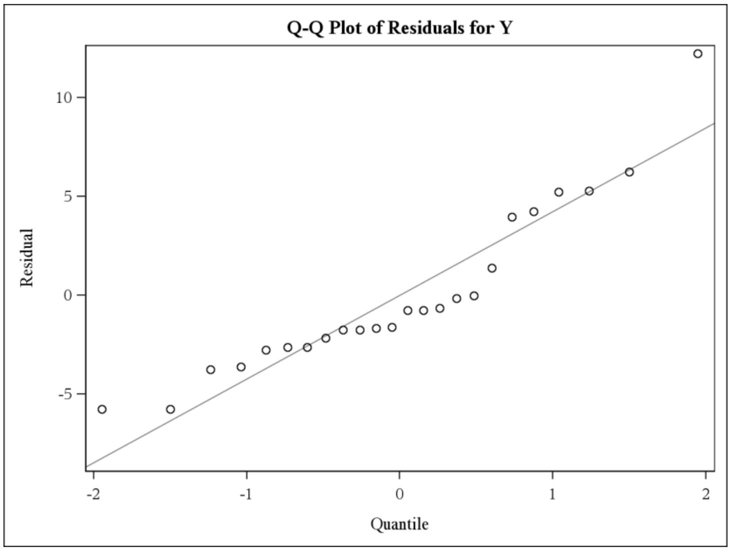
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	13.95000	4.46398	3.13	0.0049
X2	1	17.72500	3.75049	4.73	0.0001

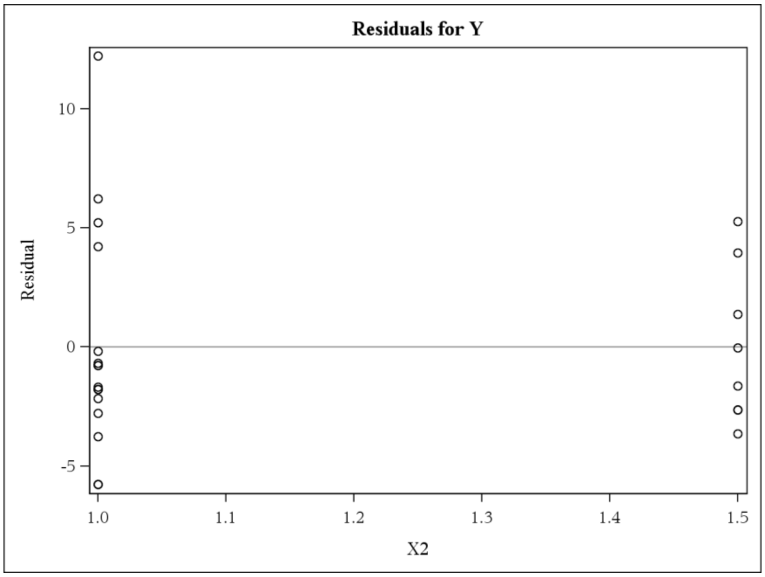


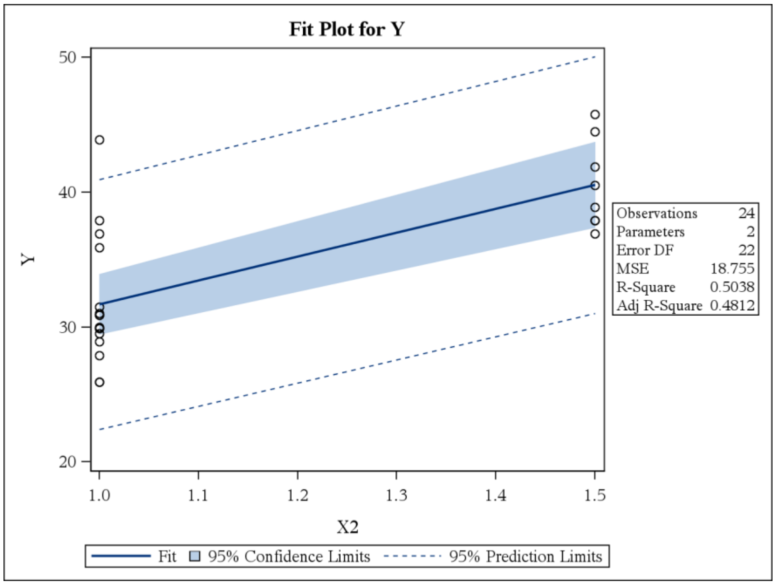


Simple Linear Regression with X2









Number of Observations Read	24
Number of Observations Used	24

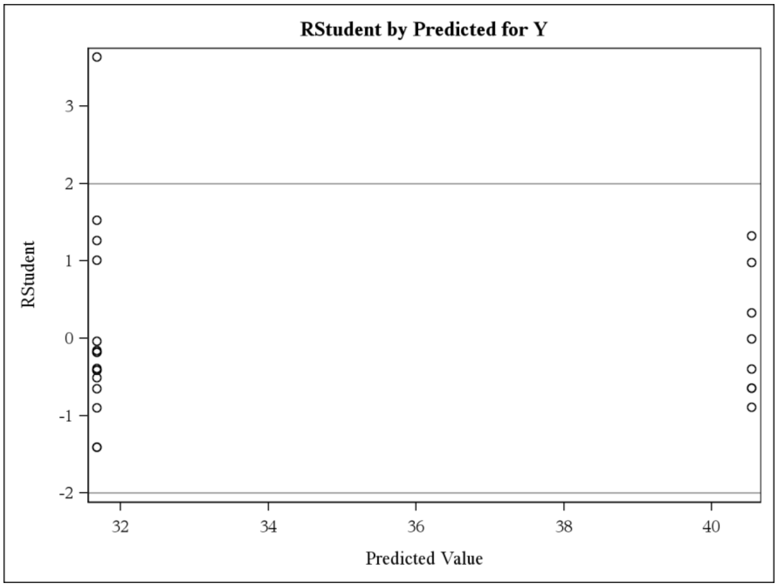
Analysis of Variance

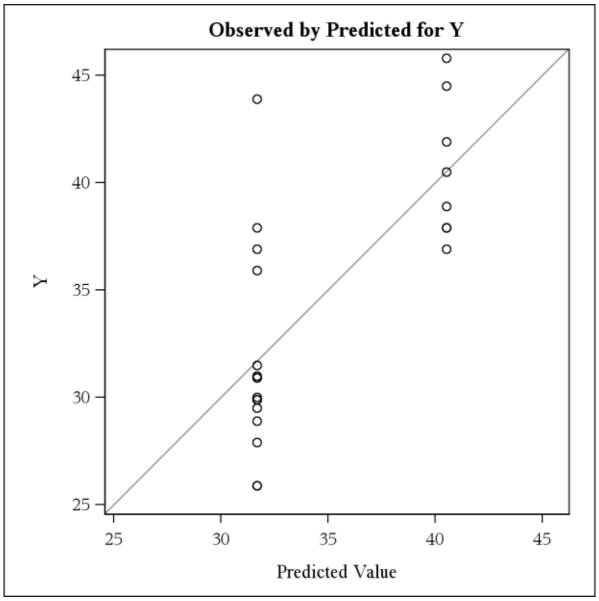
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	418.90083	418.90083	22.34	0.0001
Error	22	412.60875	18.75494		
Corrected Total	23	831.50958			

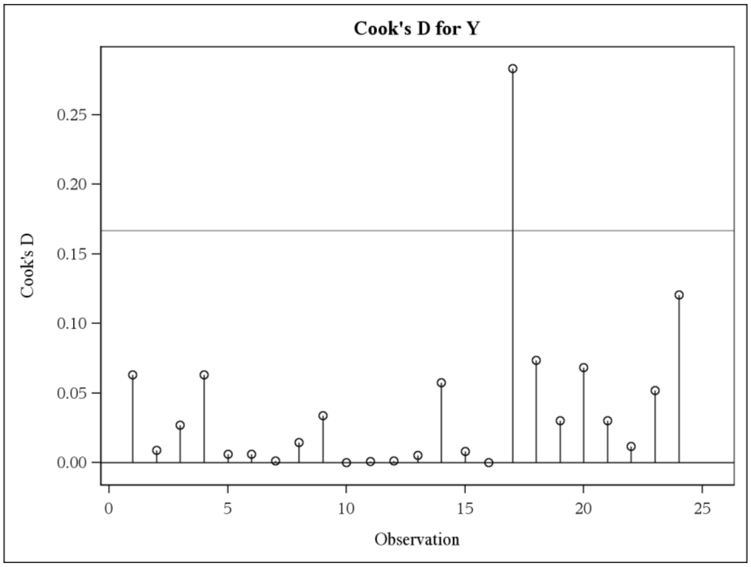
Root MSE	4.33070	**R-Square**	0.5038
Dependent Mean	34.62917	**Adj R-Sq**	0.4812
Coeff Var	12.50593		

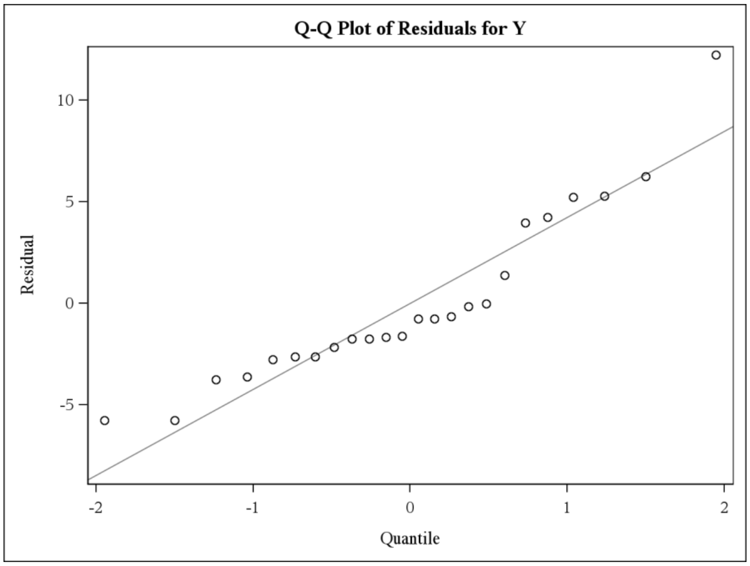
Parameter Estimates

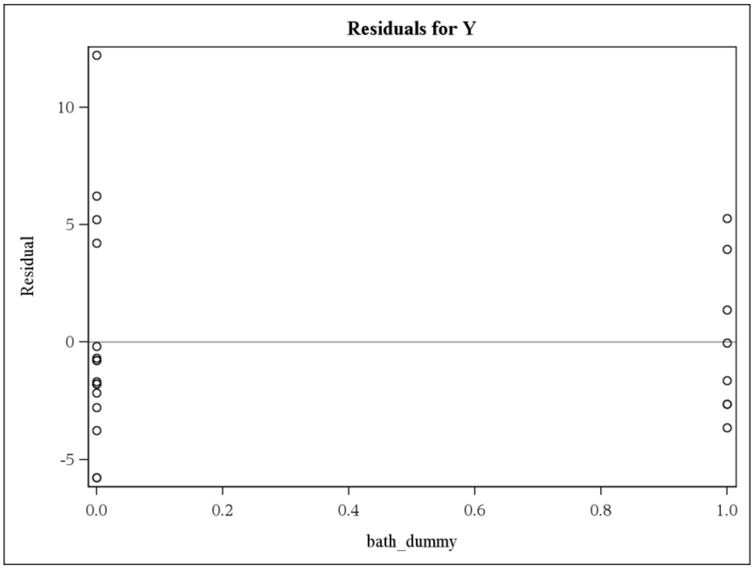
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	31.67500	1.08267	29.26	<.0001
bath_dummy	1	8.86250	1.87525	4.73	0.0001

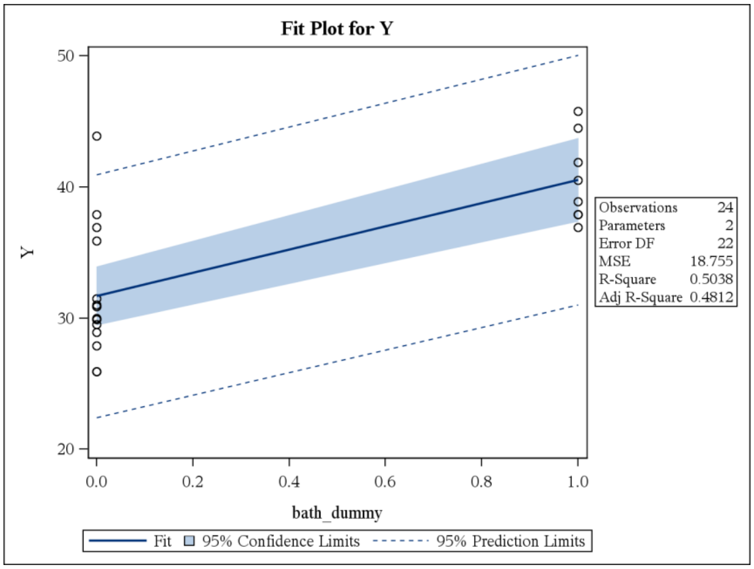












When observing the plot of residuals of Y regressed against the discrete bath_dummy or the continuous X2 value, we can see it does not follow the same plot distribution that we see with a continuous variable (random scattering of plots on the graph) that meets the assumption of linearity. Still, even with the limited values of X2, we can see there is a linear relationship with the response variable.

The plots are the same for the continuous (x2) and discrete (bath_dummy) model versions, with only the intercept and coefficient of X2 or bath_dummy varying between the models. Otherwise the comparison tests (F-test and t-test of the variable coefficients being not equal to zero) are the same using either the discrete or continuous approach.

Again, the plots and the models are the same except for the coefficients for when X2 is used or when the bath_dummy is used. In the QQ-plot, the residuals plotted against the quantile values indicate a fairly straight line, but we can see that the middle quantile values tend to all be below the line while the values in the far left and right quantiles tend to all be above the line. This is similar to an S-shape around the 45 degree line. Furthermore, we can see one outlier observation in the upper right corner. This may indicate that the errors are not normally distributed. The tests for linearity of X2 or the bath_dummy when observing the regression line over the scatterplot remain the same as well (indicate linearity).

CONCLUSION:

Based on our evaluation of single variables in Assignment #2 we came up with a simple regression model to use to compare to multiple regression models from the forward, backward, and stepwise variable selection processes. By having very similar adjusted R-square and MSE values along with significantly improved AIC and BIC values, we selected the model chosen by the backward and stepwise processes as the best model.

Y = 10.112 + 2.71703(X1) + 6.09851(X2)

We checked that the normality assumptions holds by comparing the residuals of the model to a quantile distribution and found that the plots were close to the line and didn't show any overt patterns. We overlaid the regression line on the scatterplot to ensure that the correlation was a linear correlation. Again, we could see that the values approximated a straight line with a positive correlation (line starts in the bottom left and goes to the top right). This indicates a linear relationship and that the assumptions of linearity are met. We checked the residuals against the predictor variables, X1 and X2, to determine if there were any other patterns of information that could be identified by graphically examining the pattern of the residuals. Again, we could see a fairly (but not perfect) symmetric distribution about the mean error value of zero for both variables (even if X2 values were centered around the values of 1.0 and 1.5). Finally, we used Cook's distance for each observation to determine if there are any outstanding or highly influential observations that may be affecting the model in a significant manner.

Overall the results showed that X1 with X2, the amount of taxes and number of bathrooms, were the best predictor variables to use to predict the sales price of the house. However, it should be noted that when evaluating the coefficient of X2 (number of bathrooms) versus Y (sale price of house) with the t-test it was not significant using the standard alpha value of 0.05.

We also tested the differences when using X2 (continuous) or a bath_dummy (discrete) in the model. Interestingly, the model's significance was the same for both approaches with only the coefficients changing for the intercept and X2/bath_dummy. When testing X2 or bath_dummy in a simple linear regression model, this predictor variable may be suspect to errors and the violation of normality assumptions if it is used. Consideration to transform some of variables may improve this model.

CODE:

```
***PART 1***;

* creating the macro for data set;
%let PATH = /courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/;
%let NAME = MYDATA;
%let LIB = &NAME..;
```

```
libname &NAME. "&PATH." access=readonly;

%let INFILE = &LIB.building_prices;
%let TEMPFILE = TEMPFILE;

* setting varaible names with labels;
data &TEMPFILE.;
set &INFILE.;
label Y = "Sale price of house in thousands of dollars";
label X1 = "Taxes in thousands of dollars";
label X2 = "Number of bathrooms";
label X3 = "Lot size in thousands of square feet";
label X4 = "Living space in thousands of square feet";
label X5 = "Number of garage stalls";
label X6 = "Number of rooms";
label X7 = "Number of bedrooms";
label X8 = "Age of home in years";
label X9 = "Number of fireplaces";
run;


/* view data set */
proc print data=&TEMPFILE.;
run;


* TEST of multi--option model;


ods graphics on;

proc reg data  = MYDATA.building_prices
   outest = OUTFILE AIC BIC CP ADJRSQ MSE
   plots = diagnostics(stats=(default aic bic CP ADJRSQ MSE))
   ;
original:  model Y = X1;
forward: model Y = X1-X9 /selection=forward;
backward: model Y = X1-X9 /selection=backward;
stepwise: model Y = X1-X9 /selection=stepwise;
run;
quit;

ods graphics off;

proc print data=OUTFILE;
```

```
run;


***** PART 2 *****;
ods graphics on;
PROC REG data=&TEMPFILE. plots = (diagnostics residualplot);
        MODEL y=x1 x2 / vif;
        title "Analysis of best model selected";
        output out=fitted_model pred=yhat residual=resid ucl=ucl lcl=lcl;
run;
ods graphics off;



***** PART 3 *****;
data &TEMPFILE.;
set &INFILE.;
if (X2=1.5) then bath_dummy=1;
else bath_dummy=0;
run;

ods graphics on;
PROC REG data=&TEMPFILE. plots(only)=(DIAGNOSTICS FITPLOT);
        MODEL y=x1 x2 / vif;
        title "Analysis with discrete bath_dummy variable";
        output out=fitted_model pred=yhat residual=resid ucl=ucl lcl=lcl;
run;



**** Running X2 and bath_dummy as single linear regressions ****;
proc reg data=&TEMPFILE. plots(only)=(diagnostics(unpack) residualplot fitplot);
        model y=x2;
        title "Simple Linear Regression with X2";
run;

proc reg data=&TEMPFILE. plots(only)=(diagnostics(unpack) residualplot fitplot);
model y=bath_dummy;
title "Simple Linear Regression with bath_dummy";
run;
ods graphics off;
```

****BINGO BONUS******

1) USING MACROS (5pts)

```
* creating the macro for data set;
%let PATH = /courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/;
%let NAME = MYDATA;
%let LIB = &NAME..;
libname &NAME. "&PATH." access=readonly;

%let INFILE = &LIB.building_prices;
%let TEMPFILE = TEMPFILE;

* setting varaible names with labels;
data &TEMPFILE.;
set &INFILE.;
label Y = "Sale price of house in thousands of dollars";
label X1 = "Taxes in thousands of dollars";
label X2 = "Number of bathrooms";
label X3 = "Lot size in thousands of square feet";
label X4 = "Living space in thousands of square feet";
label X5 = "Number of garage stalls";
label X6 = "Number of rooms";
label X7 = "Number of bedrooms";
label X8 = "Age of home in years";
label X9 = "Number of fireplaces";
run;

/* view data set */
proc print data=&TEMPFILE.;
run;
```

2) DEPLOYING MODELS (5pts)

```
data data_with_sale_price;
        set &TEMPFILE.;
        sale_price = 10.112 + 2.71703*X1 + 6.09851*X2;
        title "Data set with deployed model for predicted sales price";
run;

proc print data=model_sale_price;
run; quit;
```

Obs	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y	bath_dummy	sale_price
1	4.918	1.000	3.472	0.998	1.000	7.000	4.000	42.000	0.000	25.900	0	29.5729
2	5.021	1.000	3.531	1.500	2.000	7.000	4.000	62.000	0.000	29.500	0	29.8527
3	4.543	1.000	2.275	1.175	1.000	6.000	3.000	40.000	0.000	27.900	0	28.5540
4	4.557	1.000	4.050	1.232	1.000	6.000	3.000	54.000	0.000	25.900	0	28.5920
5	5.060	1.000	4.455	1.121	1.000	6.000	3.000	42.000	0.000	29.900	0	29.9587

3) PROC VARCLUS (10pts)

```
proc varclus data=&TEMPFILE.;
        VAR x1-x9;
        title "Clustering with SAS!  :)  ";
run;
```

How many were in variable clusters were found?

Cluster Summary for 3 Clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	3	3	2.345616	0.7819	0.5271
2	4	4	2.904572	0.7261	0.6184
3	2	2	1.22578	0.6129	0.7742

Total variation explained = 6.475968 Proportion = 0.7196

3 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	X5	0.6123	0.1385	0.4500
	X6	0.8825	0.4341	0.2077
	X7	0.8508	0.2053	0.1877
Cluster 2	X1	0.8234	0.3068	0.2547
	X2	0.6750	0.1996	0.4060
	X3	0.6017	0.0812	0.4335

3 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
	X4	0.8045	0.3813	0.3160
Cluster 3	X8	0.6129	0.0913	0.4260
	X9	0.6129	0.0476	0.4065

I think this question is asking how many clusters were found. In that case, three clusters were found. C1 = (x5, x6, x7), C2 = (x1 , x2, x3, x4), C3 = (x8, x9).

What variables were highly correlated?

Cluster Structure			
Cluster	1	2	3
X1	0.553902	0.907428	-.185483
X2	0.446789	0.821593	0.066023
X3	0.284938	0.775674	-.029862
X4	0.617457	0.896918	-.020773
X5	0.782513	0.372104	0.052020
X6	0.939389	0.658831	0.221293
X7	0.922408	0.453143	0.200231
X8	0.166326	-.302207	0.782873
X9	0.122898	0.218187	0.782873

I didn't recognize where correlation coefficients between single variables was shown but when I look at the cluster structure table we can see which variables have the highest correlation to the specified cluster. For example, for cluster 1 we can see that the variables X5, X6, and X7 were correlated at .7825, .9394, and .9224 respectively.

Did the clusters make sense?

Yes, in some ways. For cluster 1, each variable is the amount of some type of room in the home. Thus it seems reasonable that a larger and more expensive house would have more garages, bathrooms, and bedrooms. For cluster 2, the variables taxes, lot size, living area, and number of bathrooms are used. These seem a little more odd as a combination because all except one (taxes) are related to the size of the house. For cluster 3, age of home and number of fireplaces seems reasonable because a new home with lots of fireplaces will probably be more expensive than an old home with a no fireplaces.

Standardized Scoring Coefficients			
Cluster	**1**	**2**	**3**
X1	0.000000	0.312414	0.000000
X2	0.000000	0.282862	0.000000
X3	0.000000	0.267053	0.000000
X4	0.000000	0.308795	0.000000
X5	0.333607	0.000000	0.000000
X6	0.400487	0.000000	0.000000
X7	0.393248	0.000000	0.000000
X8	0.000000	0.000000	0.638673
X9	0.000000	0.000000	0.638673

Inter-Cluster Correlations			
Cluster	**1**	**2**	**3**
1	1.00000	0.56619	0.18472
2	0.56619	1.00000	-0.05366
3	0.18472	-0.05366	1.00000

No cluster meets the criterion for splitting.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	4.214217	0.4682	0.4682	1.706227	0.0180	
2	5.428493	0.6032	0.5138	1.262502	0.0683	0.9460
3	6.475968	0.7196	0.6129	0.774220	0.6017	0.4500