Assignment #3: Multiple Regression Model (100 points)

Introduction:

The objective of **Scientific Mass Appraisal®** is to develop automated valuation models designed to predict the sale price of a home based on selected physical attributes of the property. To facilitate this process, a multiple regression model will be built (Assignment #3) by analyzing historical data recorded for 24 homes using ten (10) variables including the sales price (i.e., response variable) and other features/characteristics of each home as illustrated in *Table 1, Data Overview*. The optimal regression model will be the model that most accurately predicts sales price in thousands of dollars (Y) for given values of the following predictor variables characterized below:

Table 1 | Data Overview

Model Variables	Label	Description	Units
Response Variable	Y	Sales price	$1000 (USD)
Predictor Variables	X1	Taxes in thousands of dollars	$1000 (USD)
	X2	Number of bathrooms	Count
	X3	Lot size in thousands of square feet	1000'(SQFT)
	X4	Living space in thousands of square feet	1000'(SQFT)
	X5	Number of garage stalls	Count
	X6	Number of rooms	Count
	X7	Number of bedrooms	Count
	X8	Age of home in years	Years
	X9	Number of fireplaces	Count

Source: data temp**;** set mydata.building_prices**;**

Part 1: For this assignment we will fit a multiple regression (a regression model with one or more predictor variables) to the building_prices data set. For reference, our multiple regression models will be compared to a simple linear model from Assignment #2 where the predictor variable, *Taxes in thousands of dollars (X_1)* was prioritized as the variable with the strongest linear relationship with the response variable, *Sales Price (Y).* This model has the form:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Parameter estimates and ANOVA metrics for this model are characterized in Tables 2 & 3.

Table 2 | Parameter Estimates for Simple Linear Regression Model

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

Table 3 | 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			

Additional predictor variables will also be assessed to determine if the predictive accuracy the simple linear regression model can be improved. Automated variable selection procedures including forward, backward, and stepwise variable selection were implemented and characterized below:

Forward Selection Method:

Table 4 | Analysis of Variance – Forward Selection

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			

Table 5 | Parameter Estimates – Forward Selection

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

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

Table 6 | 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

Backward Selection Method:

Table 7 | Analysis of Variance – Backward Selection

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			

Table 8 | Parameter Estimates – Backward Elimination					
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

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

Table 9 | 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

Stepwise Selection Method:

Table 10 | Analysis of Variance - Stepwise					
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			

Table 11 | Parameter Estimates - Stepwise					
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

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.

Table 12 | 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

The automated variable selection methods recommended the following models:

Forward selection: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon$
Backward selection: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
Stepwise selection: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

The backward and stepwise selection processed identified the same model and included fewer predictor variables. However, each variable selection method could have possibly selected a different model because the selection criteria and/or sequence of application of selection criteria are different for each model.

The forward selection method starts with the single best variable (the one that yields the largest F-statistic) and adds variables one at a time until the p-value for the variables being entered is larger than 0.50 (default setting in SAS). (Cody , *SAS® Statistics by Example,* 2011)

The backward selection method starts with all the 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 0.10 (default setting in SAS). (Cody , *SAS® Statistics by Example,* 2011)

The stepwise selection method is almost the same as the forward method except that a variable that has already been added to the model at a previous step might be removed later. (Cody , *SAS® Statistics by Example,* 2011)

The automated variable selection methods applied to develop a multiple regression model have identified models with more predictive accuracy as characterized by comparison of the mean square error (MSE) of each model.

Variable Selection	DF	Sum of Squares	Mean Square	F Value	Pr > F
Forward Selection	7	706.00703	**100.85815**	12.86	<.0001
Backward Selection	2	663.59779	331.79890	41.50	<.0001
Stepwise Selection Method	2	663.59779	331.79890	41.50	<.0001
Simple Linear Regression	1	635.04186	635.04186	71.11	<.0001

The model that resulted in the lowest MSE was the generated using the forward selection process.

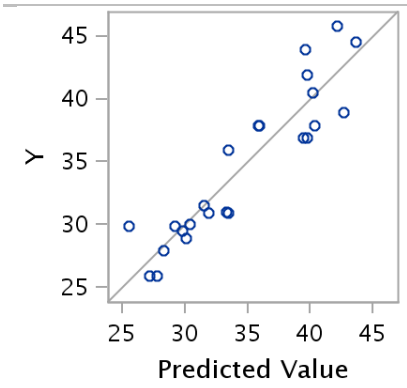
Part 2:

To ensure this model is adequate for predictive purposes, diagnostic and residual plots were generated and assess to ensure model adequacy. Goodness of fit assessments were implemented to validate model assumptions and detect outliers by generating several diagnostic and residual plots and characterized below.

Validation of Model Assumptions

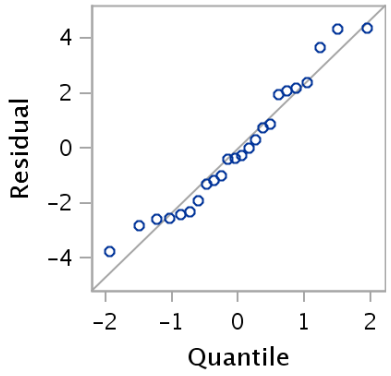
a) **Linearity** – Visual analysis of the scatterplot was applied to determine if a linear relationship between the *Home Sales Price (Y)* and *Taxes in thousands of dollars* (X1) and Number of bathrooms (X2), Living space in thousands of square feet (X4), Number of garage stalls (X5), number of rooms (X6), age of the home in years (X8), and number of fireplaces (X9) can be reasonably assumed to exist. As illustrated in Table 3, the scatterplot suggests a linear positive relationship does is plausible.

Table 3 |Scatter Plot for Home Sales Price by Taxes & Number of Bathrooms



b) **Normality** – The QQ Plot for Sales Price by the modeled predictor variables was analyzed to evaluate the normality assumption by determining if the frequency distribution of the errors between the predicted vs. actual home sales prices can be reasonably assumed to follow a normal distribution. As illustrated in Table 6, the residuals generally fall onto the 45° degree axis without any substantial delineation implying our residuals are normally distributed. Consequently, the regression line proposed by our model minimizes the error between the observed and predicted values and represents and adequate model for prediction.

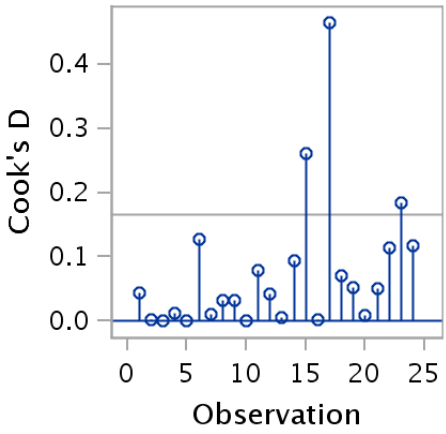
Table 6 | QQ Plot for Home Sales Price by Predictor Variables



Detecting Outliers

As illustrated in Table X, the plot of Cook's D statistic indicates that observations #15, #17, and #24, exceed the threshold value indicating that both observations significantly influence the model's accuracy and warrant further examination and/or removal from the model. Although, observation #14, did not exceed the threshold value, further examination may also warrant in consideration of the proximity to the threshold line.

Table X | Cook's Distance Plot



Detecting Collinearity

To detect collinerity in our model, the variance inflation factors (VIF) for each predictor variable was generated and analyzed. Values for VIF greater are considered large and worrisome. Although predictors with VIF values between 5-10 are tolerable, they do warrant further examination . However, as illustrated in Table X, the VIF for all predictor variables is < 5. Consequently, there is no meaningful linear dependence between the predictor variables that may influence the model's predictive accuracy.

Table X | 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

Part 3.

Analysis of Regression Model – Number of Bathrooms (X2)

The second regression model obtained from the *selection rsquare option* included the predictor variable, *Number of Bathrooms (X2).* This model was selected based on generation of a variable selection list prioritized by highest R^2 value. To ensure this model is adequate for predictive purposes, diagnostic and residual plots were generated and assess to ensure model adequacy.

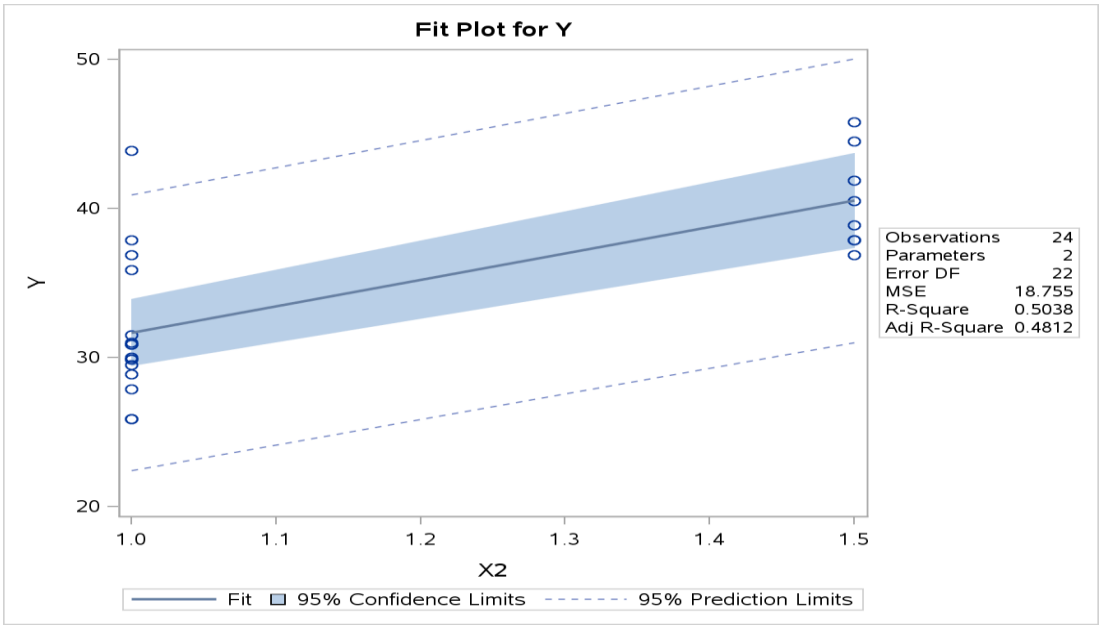
Model Diagnostics

Goodness of fit assessments were implemented to 1) validate model assumptions and 2) detect outliers by generating several diagnostic and residual plots and characterized below.

Validate Model Assumptions

a. **Linearity** - –visual analysis of the ***Fit Plot for Y*** was applied to assess if a linear relationship between the *Home Sales Price* (response variable) and *Number of Bathrooms* (predictor variable) can be reasonably assumed to exist. As illustrated in Table 8, the plot suggests that a **non-linear relationship** between *Home Sales Price (Y)* and *Number of Bathrooms (X2)* is a reasonable assumption.

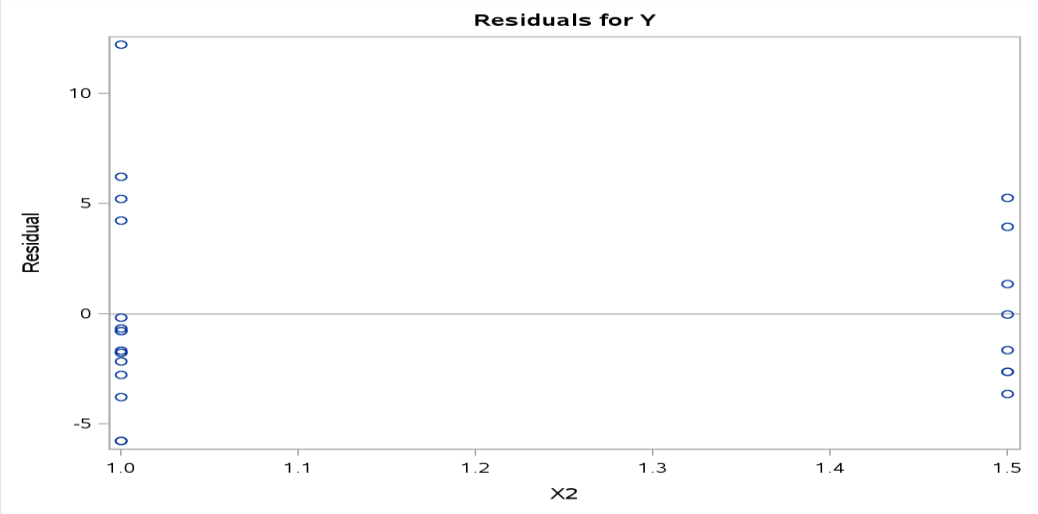
Table 8 | Fit Plot for Home Sales Price by Number of Bathrooms



Homoscendasticity (constant variance) – the variation of error terms should be constant with respect to the predictor variable *Number of Bathrooms*. Assessment of the diagnostic residual plots can confirm whether or not the variation in error for a predictor variable is consistent. The residuals plotted against the predictor variable can be evaluated to validate required model assumptions related to constant variance.

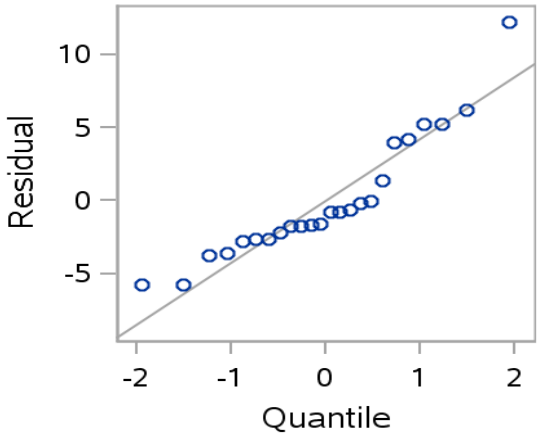
As illustrated in Table 9, the *Residuals for Y* plot is **not random** and the distribution of the residuals is **not constant**. As the number of bathrooms increases, we do not observe observable changes in the differences between the observed vs. actual values of the sales price of homes.

Table 9 | Residuals for Home Sales Price by Number of Bathrooms



b. **Normality** – The *Normal Probability Q-Q Plot* was analyzed to evaluate the normality assumption by determining if the frequency distribution of the errors between the predicted vs. actual home sales prices can be reasonably assumed to follow a normal distribution. As illustrated in Table 10, the residuals generally fall onto the 45° degree axis without any substantial delineation implying our residuals are normally distributed with the exception of two (2) observations. The regression line proposed by our model minimizes the error between the observed and predicted values and may represent and adequate model for prediction depending on further examination of observation #1 and observation #24.

Table 10 | Normal Probability Q-Q Plot Home House Price by Number of Bathrooms

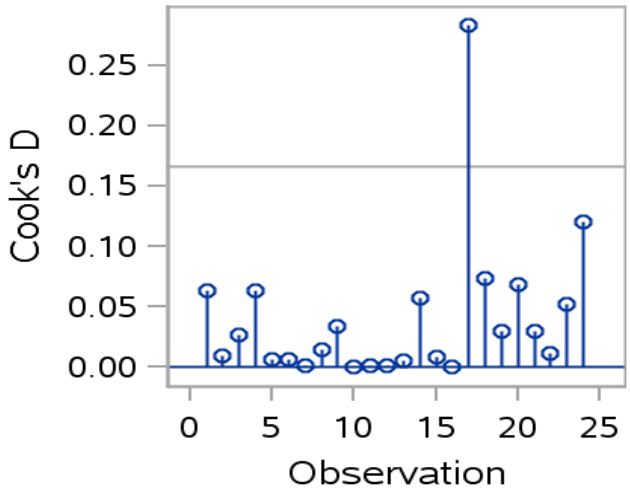


Detecting Outliers

Detecting and understanding the influence of outliers is a critical element of evaluating the adequacy of the model. **Cook's distance** "measures the difference between the regression coefficients obtained from the full data" (Chatterjee & Hadi) and the regression coefficients obtained by deleting each observation of the predictor to determine if this observation significantly influences the adequacy of the model.

As illustrated in Table 11, the plot of Cook's D statistic indicates that one value, **observation #17**, exceeds the threshold value indicating that this observation significantly influences the model's accuracy.

Table 11 | Cook's Distance Plot Home House Price by Number of Bathrooms



Although evaluation of R^2 suggested that the number of bathrooms provided the second most influential linear relationship associated with home selling price, model diagnostics were evaluated and concluded that a linear relationship does not exist and that non-linear models or variable transformations should be considered.

Number of bathrooms is a categorical variable and is recommended to be transformed to be evaluated as a continuous variable.

Note: Unfortunately, I was not able to get my SAS program to work for this final analysis; I will assume that the transformation of the number of bathrooms to a continuous variable satisfied the OLS assumptions.

Conclusions:

The forward selection process identified a multiple regression model superior to the simple linear regression model. However, in consideration of the significant number of predictor variables, a more "efficient" and more interpretable model is recommended. By transforming the X2 variable, we were compliance with parsimony facilitated the development of a more accurate model.

Code:

```
libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
proc datasets lib=mydata; run; quit;
proc contents data=mydata.building_prices; run;
data temp;
set mydata.building_prices; run;
proc print data=temp(obs=10); run; quit;

ods graphics on;
title "ASSIGNMENT #3 - PART 1 - FIND OPTIMAL REGRESSION MODELS USING
AUTOMATED VARIABLE SELECTION PROCEDURES"
;
title2 "FORWARD, BACKWARD, AND STEPWISE SELECTION METHODS"
;
proc reg data=temp;
FORWARD: model Y=X1 X2 X3 X4 X5 X6 X7 X8 X9 / selection=forward;
BACKWARD: model Y=X1 X2 X3 X4 X5 X6 X7 X8 X9 / selection=backward;
STEPWISE: model Y=X1 X2 X3 X4 X5 X6 X7 X8 X9 / selection=stepwise;
run
;
quit
;
ods graphics off;

ods graphics on;
title "ASSIGNMENT #3 - PART 2 - FIND OPTIMAL REGRESSION MODELS USING
AUTOMATED VARIABLE SELECTION PROCEDURES"
;
proc reg data=temp;
model Y=X1 X2 X4 X5 X6  X8 X9  / VIF;
proc reg plots =(diagnostics residualplot qqplot);
model Y=X1 X2 X4 X5 X6 X8 X9;
run;
quit
;
ods graphics off;
```