Assignment #3
Daniel S Prusinski

Introduction:

This assignment focuses on fitting and assessing a multiple regression model. Using different SAS procedures I learned how to pick optimal variables and an optimal model. In addition to fitting a multiple regression model, I learned how to analyze the goodness-of-fit through analyzing specific diagnostics and metrics. Towards the end of the assignment I dealt with a variable that did not conform to the OLS assumptions, and had to rectify using the variables. This assignment sets the stage for logistic regression seeing that I learned about coding variables and interpreting the results.

Results Part 1:

The fitted simple linear regression model from assignment two is listed below with the parameter estimates.

$$Y = 13.36 + 3.32*X$$

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

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

From the parameter estimates, it can be seen that X1 is statistically significant, and has a strong coefficient of determination. SAS has three variable selection procedures, and each has advantages and disadvantages. It should also be pointed out that Ratner has additional techniques for variable selection, which should be used in addition to the SAS procedures. The Forward Selection model sets a parameter that the P-Value must not be above the .5 significance level. The models starts with just the constant and adds variables that have the greatest simple correlation with the predictor variable. Variables are not added when their p-values are greater than .5, please not this is a high p-value and variables should be scrutinized. The higher the p-Value, greater is the chance that the F-Value is attributed to sampling error. I suspect some collinearity between the variables because the P-values fluctuate as each variable enters the equation.

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

Variable	**Parameter Estimate**	**F Value**	**Pr > F**
Intercept	16.59015	11.57	0.0036
X1	2.21867	7.61	0.0140
X2	6.14082	2.60	0.1261
X4	2.86700	0.54	0.4730
X5	1.85534	2.25	0.1529
X6	-1.31636	1.17	0.2962
X8	-0.04656	0.59	0.4540
X9	2.25175	2.47	0.1355

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

Variable	Parameter Estimate	F Value	Pr > F
Intercept	10.11203	11.39	0.0029
X1	2.71703	30.60	<.0001
X2	6.09851	3.57	0.0727

The Backward Selection model sets a parameter that the P-Value must not be above the .1 significance level. This method is opposite of the forward selection procedure in that it places all the predictor variables in the model and analyzes variables that do not meet the p-value significance level. The backward selection is noted as being the best selection model to use because it does not suppress predictor variables, it also can handle collinear data better than the other selection options.

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

Variable	Parameter Estimate	F Value	Pr > F
Intercept	10.11203	11.39	0.0029
X1	2.71703	30.60	<.0001
X2	6.09851	3.57	0.0727

The Stepwise Selection model sets a parameter that the P-Value must not be above the .15 significance level. This model is similar to the forward selection procedure, but it analyzes to see if redundant predictor variables should be eliminated. Stepwise selection has a downside in that its analysis is based on the variables criterion, the distributions do not follow for the corresponding F and chi-squared tests, and the p-values do not have the correct meaning (Ratner).

Each procedure had different P-value parameters, also known as a test statistic) and thus different models were displayed. It should be noted that the BE procedure and Stepwise procedure produced the same model, and that can be directly related to the fact that the P-value limits had a .05 value difference. To accurately compare these regression models to the simple regression model, utilizing the adjusted r-squared procedure and the Residual Mean Square measure (RMS) should be used. The model with the smaller RMS is preferred given that this assignment is looking to forecast.

The adjusted R-square is given for each equation below:

Table of Adjusted R-Square			
Variable Entered	Number Vars In	Adjusted R-Square	Model
X1	1	0.7530	0
X1 X2 X4 X5 X6 X8 X9	9	0.7830	1
X1 X2	2	0.7788	2

From analyzing the Adjusted R-Square, both regression models are better predictors of Y than the simple linear regression model using X1. The next question then becomes which model is better? Given that model 2 nests model 1, I can conduct an F-test for nested models and determine which is better. I am trying to balance predictive accuracy with parsimony. The hypothesis that I will be testing is: H0: Reduced model is adequate H1: Full model is adequate. The equation is, (([SSE(RM) – SSE(FM)]) / (p+1-k)) / ((SSE(FM) / (n-p-1)) = F. The ANOVA models are listed below have the information needed to compute the equation. The calculation is: ((168-126)/6))/ (126/16) --- 7/7.875 = .888 = F

F-critcal is determined as, 6 for the numerator and 16 for the denominator, and its value at p =.05 is 2.74. Thus, I cannot reject the null hypothesis, which states the reduced model is adequate. Given the results, I choose the model that utilizes X1 and X2 as the optimal model.

Analysis of Variance Model 2 (Reduced Model)					
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			

Analysis of Variance Model 1 (Full Model)					
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			

Another method for choosing between two models is the average correlation paired with the correlation coefficient. In an EDA, an analyst is encouraged to use different tools to verify and discern the best model to use. At times, there may be a few optimal models to use. Below I have calculated the average correlation as well as the individual correlations of predictor variables with the dependent variable between the two models. Ratner states that, “the individual correlations indicate the content validity of the model.” (Page 233).

Avg Correlation Model 1

Absolute Values for Pearson Correlation

	X1	X2	X4	X5	X6	X8	X9
X1							
X2	0.65127						
X4	0.73427	0.72859					
X5	0.45856	0.22402	0.35888				
X6	0.64062	0.51031	0.67886	0.58939			
X8	0.4371	0.10075	0.13909	0.02017	0.12427		
X9	0.14668	0.20412	0.10656	0.10162	0.22222	0.22578	
						Avg Corr	0.35253

The average correlation for model has a value of .35, which means the values do not suffer from multi-collinearity. Variables X1, X2, and X4 all have values that are valid for predicting Y, but I would question the validity of using X5-X9. The next step is to compare this model to model two.

Pearson Correlation Coefficients, N = 24 Model 1, Variables PCC with Y							
	X1	X2	X4	X5	X6	X8	X9
Y	0.87391	0.70978	0.70777	0.46147	0.52844	-0.3974	0.26688
	<.0001	0.0001	0.0001	0.0232	0.0079	0.0545	0.2074

Model two has an average correlation of .32, which is comparable with model 1. Through using the average correlation, I can see that both models do not suffer from multi-collinearity. The correlation coefficient's for both variables in model two are valid. The conclusion I have reached from using the F-test for nested models and Ratner's average correlation with correlation coefficient is using model 2 is better than model. The methods used demonstrate that model two is both reliable and valid.

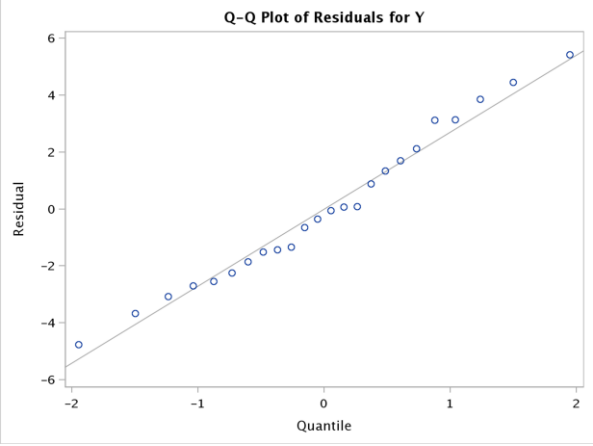
Pearson Correlation Coefficients, N = 24 Average Correlation Model 2		
	X1	X2
X1		
X2	0.65127	

Avg Corr 0.325635

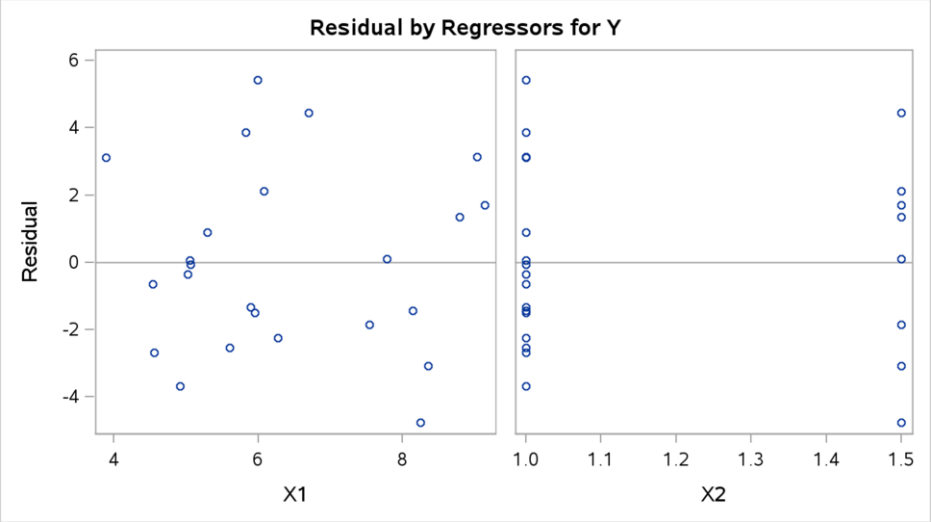
Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0		
	X1	X2
Y	0.87391	0.70978
	<.0001	0.0001

Part 2:

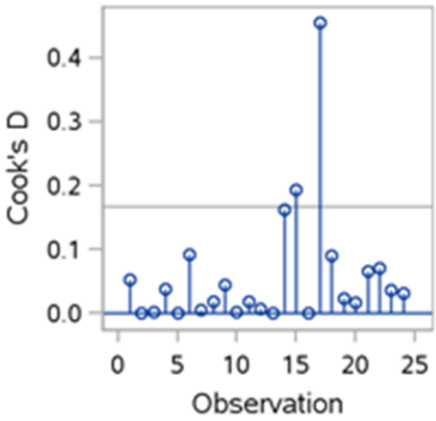
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.11203	2.99614	3.38	0.0029
X1	1	2.71703	0.49115	5.53	<.0001
X2	1	6.09851	3.22705	1.89	0.0727



The fit model utilizing X1 and X2 is, Y = 10.1 + 2.7X1 + 6.1X2. The parameter estimates are shown to the left. X2 has a low t-value which statistically means that the coefficient is not that accurate and the high p-value is suspect for sampling error. At this point analyzing the diagnostics is necessary for further assessment of the model adequacy. Analyzing the Q-Q plot in SAS visually inspects to verify the assumption that the residuals follow a normal distribution. The normality errors fall along a 45 degree line which proves the assumption that the errors or residuals follow a normal distribution. It should be noted that fitting the regression model over the scatter plot is not relevant in multiple regression because more than one variable is used in the model, and the scatter plot with multiple variables does not validate linearity.



The plots demonstrate a random pattern for X1 around 0, which validates the assumption that the X1 predictor variable and residuals are not related. X2 also has a distribution about 0, but notice that the residuals are in a straight line. From this plot, work needs to be done to create dummy variables.



The plots demonstrate there are not any major outliers. One point to note is the plot that is over .4. But given that this is nowhere near one standard deviation, it can be assumed that it is not an extreme outlier.

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

For, both predictor variables listed, there is not an issue of collinearity given that the Variance Inflation (VI) is well below ten. Please refer to the average correlation for this model that is posted above. Bruce Ratner makes the case that a model that has an average correlation less than .35 does not suffer from multicollinearity. This models average correlation is .32, which is another metric to confirm that this model does not suffer from multicollinearity.

Part 3:

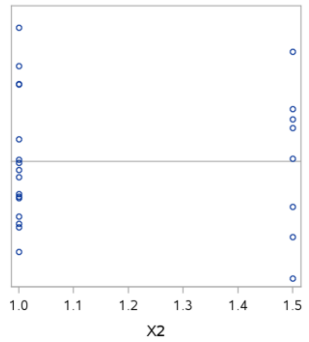
For part two of this assignment I fit the model using the variables X1 and X2. From the inferential statistics, X2 has a low t-score and high p-value. This statistically suggests that X2 does not significantly reduce the prediction error at the 5% level. If I want to use this variable as a predictor variable in my equation I need to modify the model such that X2 does not wield such variance. Please refer to the diagnostics from part 2 to interpret this model.

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

This is the same model that I chose for my optimal model in Part 2. What I would like to point out is the small t-value and p-value.

$$Y = 10.112 + 2.717X + 6.099X2$$

Plot of the Residuals Vs. X2



The diagnostic used for multiple regression to validate the linear assumption is to plot the residuals against each predictor variable. It is clear from this plot that the data does not show signs of heteroscedasticity.

X2 as a Dummy Variable

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

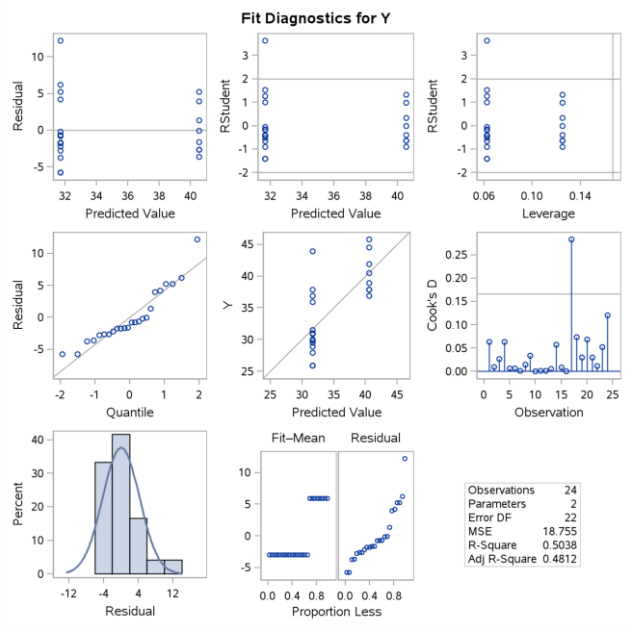
Using X2 as a dummy variable fits the model better. The intercept is larger and has greater statistical significance. Also, X2 (bath_dummy) has less of a parameter influence and less standard error. But, the low t-score and p-value is of concern.

Fitted Model of X2

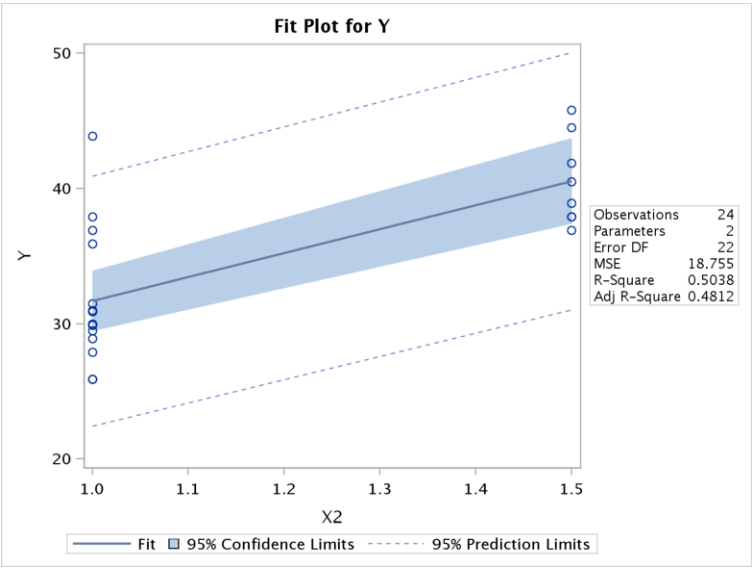
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.95000	4.46398	3.13	0.0049
X2	1	17.72500	3.75049	4.73	0.0001

The t-value is too low to reject the null hypothesis at the 95% level. Inferentially, it should be noted that this parameter does not enter the equation linearly. The p-value is strong, signifying that this is not by chance.

Diagnostics for X2



The diagnostics for X2 appear fine for the SAS diagnostic output to the left. The residuals are scattered about 0, but there seems to be quite a variance. This might invalidate the OLS assumption that the errors have a constant variance. The observation at the top left is violating this assumption.The Q-Q plot hovers around the 45 degree line except for the one point in the upper right. Cooks D appears to have only one minor outlier.



The fit plot for X2 is where I have a minor issue. Treating X2 as a continuous variable violates continuity as seen in the scatter plot. X2 only has two possible outcomes rather than continuous outcomes. The line still falls in between the variables, and it is linear.

After assessing the diagnostic plots, I did not find any major infractions of the OLS assumptions. There are a few fringe violations, but nothing jumped out at me as a major violation. This really disappointed me, because I spent a lot of time trying to find a violation. But, it is very clear that creating a discrete variable for X2 creates a better model than using it as a continuous variable. By creating X2 as a discrete variable the variable coefficient is smaller and its values are not arbitrarily viewed as continuous variables.

Conclusion:

Through this assignment, I learned how to fit a multiple regression model and methods for discerning which variables and models are accurate and valid. I enjoyed referencing Bruce Ratner's average correlation with individual correlation coefficients to prove validity when comparing models. Proving the OLS assumptions were difficult and I need more practice. While I understand coding variables that are categorical, I am looking forward to learning more about binary response variables.

Code:

```
*Daniel Prusinski Assignment 3 Version 1*******************
************************************************************
************************************************************;

*****Statement to access where the data is stored*****;
libname mydata '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/';
ods graphics on;

*****Fitting X1*****;
title "Checking which Model is Best";
proc reg data=mydata.building_prices;
      model Y = x1 /
run;

*****Demonstrating Forward, Backward, and Stepwise Selection Mothods Chapter
9*****;
title "Forward, Backward, and Stepwise Selection Methods";
title 2 "Using Default Values for SLENTRY and SLSTAY";
proc reg data=mydata.building_prices;
      Forward: model Y = x1 x2 x3 x4 x5 x6 x7 x8 x9 /
      selection = forward;
      Backword: 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;

*****Checking Adjusted RSquare*****;
title "Checking which Model is Best";
proc reg data=mydata.building_prices;
      model Y = x1 x2 x4 x5 x6 x8 x9 /
run;

*****Fitting X1 and X2*****;
title "Checking which Model is Best";
proc reg data=mydata.building_prices;
      model Y = x1 x2 /
run;

*****Using Proc Reg for Residual and Diagnostics*****;
title 'Fits of Regression Analysis';
proc reg data=mydata.building_prices plots (only) = (QQplot Diagnostics
Residuals);
      model y = x1 x2;
Run;
```

```
*****Using the VIF to Detect Collinearity, 156*****;
proc reg data=mydata.building_prices;
      model y = x1 x2 / VIF;
run;
ODS Graphics off;

*****Creating Dummy Variables for Regression, 153*****;
data temp;
      set mydata.building_prices;
      if (X2=1.5) then bath_dummy=1;
      else bath_dummy=0;
run;

*****Running regression analysis with diagnostics for X1 and dummy
variable****;
proc reg data=temp;
      model y = x1 bath_dummy;
      plots (only) = (QQplot Fitplot Diagnostics Residuals);
run;
ODS Graphics off;

*****Running regression analysis and diagnostics for X2*****;
proc reg data=mydata.building_prices;
      model y = x2;
      plots = (Fitplot Diagnostics Residuals);
run;
ODS Graphics off;

Title "Ratner's Avg Correlation";
proc corr data=mydata.building_prices;
      var x1 x2 x4 x5 x6 x8 x9;
      with x1 x2 x4 x5 x6 x8 x9;
run;

Title "Pearson CC with Y";
proc corr data=mydata.building_prices;
      var x1 x2 x4 x5 x6 x8 x9;
      with y;
run;

Title "Ratner's Avg Correlation";
Title 2 "Model 2 Avg Correlation";
proc corr data=mydata.building_prices;
      var x1 x2;
      with x1 x2;
run;

Title "Pearson CC with Y";
Title 2 "Model 2 Pearson CC";
proc corr data=mydata.building_prices;
      var x1 x2;
      with y;
run;
```