Assignment #3

Introduction:

Automated variable selection is typically done using one of three methods: forward selection, backward elimination and stepwise selection. Forward selection adds one variable at a time based on each variable's correlation with Y until the next variable's regression coefficient doesn't meet the threshold of significance. Backward selection, on the other hand, starts with all variables in the model and deletes them one at a time based on how each variable contributes to the reduction in error sum of squares. Stepwise selection is basically the same as forward selection but can also choose to delete variables at each step. For the first part of this analysis, I am going to run all three methods of automated variable selection and discuss the results.

After identifying a regression model, it is important to assess the model adequacy to make sure that the model does not violate ordinary least squares regression assumptions. This includes verifying that the model represents a linear relationship, that the errors in the regression equation are normally distributed and have the same variance, and that there are not individual observations that unduly influence the regression model. For the second part of this analysis, I am going to analyze the regression model chosen from Part 1 graphically to check for model adequacy.

Finally, I will investigate whether a variable with just two values should be treated as a continuous predictor variable or a categorical predictor variable. I will also look at whether treating a variable with just two values as a continuous variable violates any OLS regression assumption.

Results:

Part 1:

To find the optimal multiple regression model, I have used three types of automated variable selection: forward, backward and stepwise. Forward selection chose the following model: Y = 16.59 + 2.22*X1 + 6.14*X2 +2.87*X4 + 1.86*X5 – 1.32*X6 - .05*X8 +2.25*X9

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

Since the threshold for entry in forward selection is set to a default slentry=0.5 in SAS, the forward selection method resulted in 7 variables being included in the regression 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

Backward elimination chose the following model: Y = 10.11 + 2.72*X1 + 6.10*X2

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

Since the threshold for staying in backward elimination is set to a default slstay=0.1 in SAS, the backward elimination method resulted in just two variables being included in the 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

Stepwise selection chose the following model: Y = 10.11 + 2.72*X1 + 6.10*X2

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

Since the threshold for entry and staying is set to a default slentry=0.15 and slstay=0.15 in SAS, the stepwise selection method resulted in just two variables being included in the regression 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

Forward, backward and selection methods of variable selection did not all choose the same model. Forward selection chose a 7 variable model, while backward elimination and stepwise selection chose the same 2 variable model. The default value for slentry in SAS for forward selection is set much higher than that in stepwise selection, while the default value for slstay in backward elimination is close to that in stepwise selection. It is likely that these methods selected different variables based on the variety in their slentry and slstay default values. A slentry value of 0.5 means that the probability associated with the F test must be less than 50% to enter the regression model. This is a very low threshold since we're usually more interested in a 5% cutoff. Since stepwise selection increases this threshold to 15%, fewer variables enter the model than in forward selection.

To analyze the best predictive power of each regression model, along with the simple linear regression model from Assignment 2, we can look at a number of goodness-of-fit metrics for each model. The results are summarized in the following table.

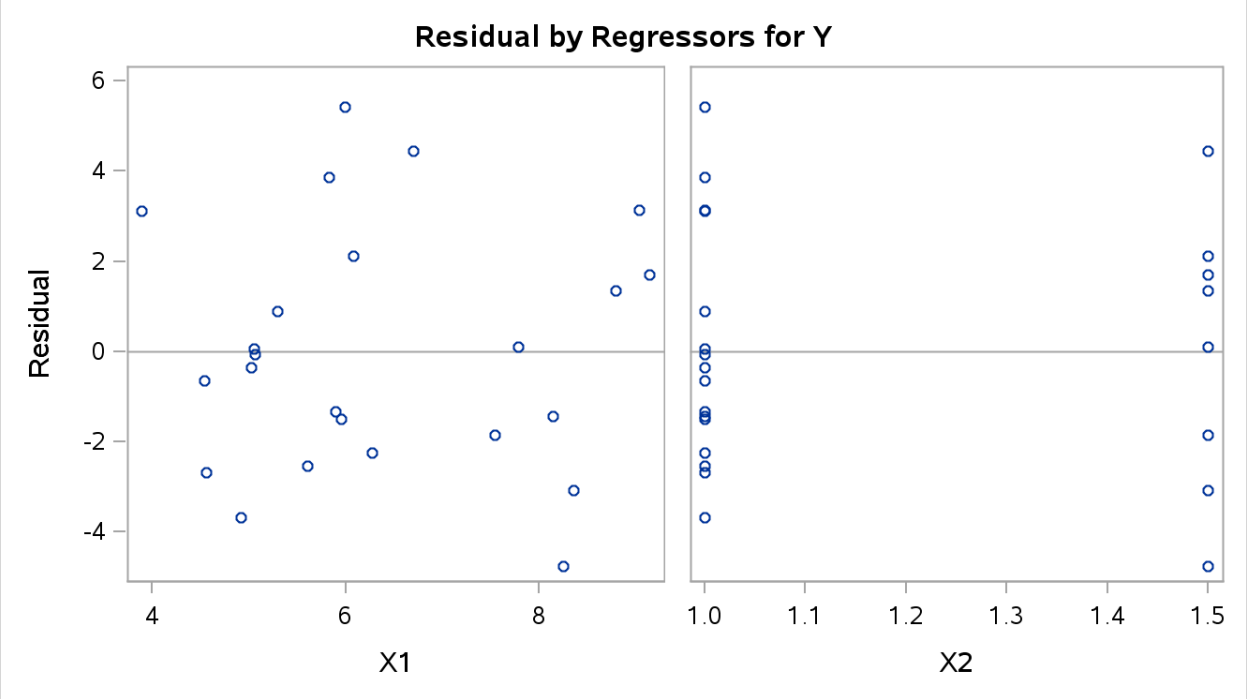
Metric	Forward model	Backward/Stepwise model	Simple linear regression model
Parameters	8	3	2
R^2	0.8491	0.7981	0.7637
Adjusted-R^2	0.783	0.7788	0.753
AIC	55.703	52.689	54.459
BIC	65.203	55.387	56.365
CP	8	3.4066	4.5714

R^2 shows the percentage of the variable Y that can be predicted from the variables in the regression model, but it does not adjust for adding useless variables to the model but continuously goes up with each additional variable. Thus it is biased in favor of choosing the largest variable model and is not a good indicator for multiple linear regression. In this comparison, we see that R^2 is larger for models with larger parameters. Adjusted-R^2, on the other hand, does adjust for number of variables, but it does not have the same meaning as R^2 – it does not represent how much variation in Y is due to the regression variables in the model. Adjusted-R^2 still shows bias for the larger parameter model. AIC and BIC are alternative variables that show goodness-of-fit by balancing accuracy with parsimony. BIC differs from AIC in that it penalizes a model with more variables more severely. The AIC score for the backward/stepwise model is the lowest, almost lower than the simple linear regression model by 2, which is considered substantial. Thus AIC and BIC scores indicate that the backward/stepwise model is the best of the three and has more predictive power than the simple linear regression model from Assignment 2. Mallows C_p is a final goodness-of-fit metric that can be used to show whether a model is too sparse or overfitted. The best C_p score should be close to the number of parameters. The backward/stepwise model is the only one that fulfills this criteria for goodness of fit.

Part 2:

If a regression model violates any of the OLS regression assumptions, then its validity is questionable. Thus we want to examine the adequacy of the regression model chosen from Part 1 and make sure that it adheres to OLS regression assumptions, namely linearity of the model, normality and homogeneity of the errors, linear independence of the predictor variables and equal reliability and roles of observations.

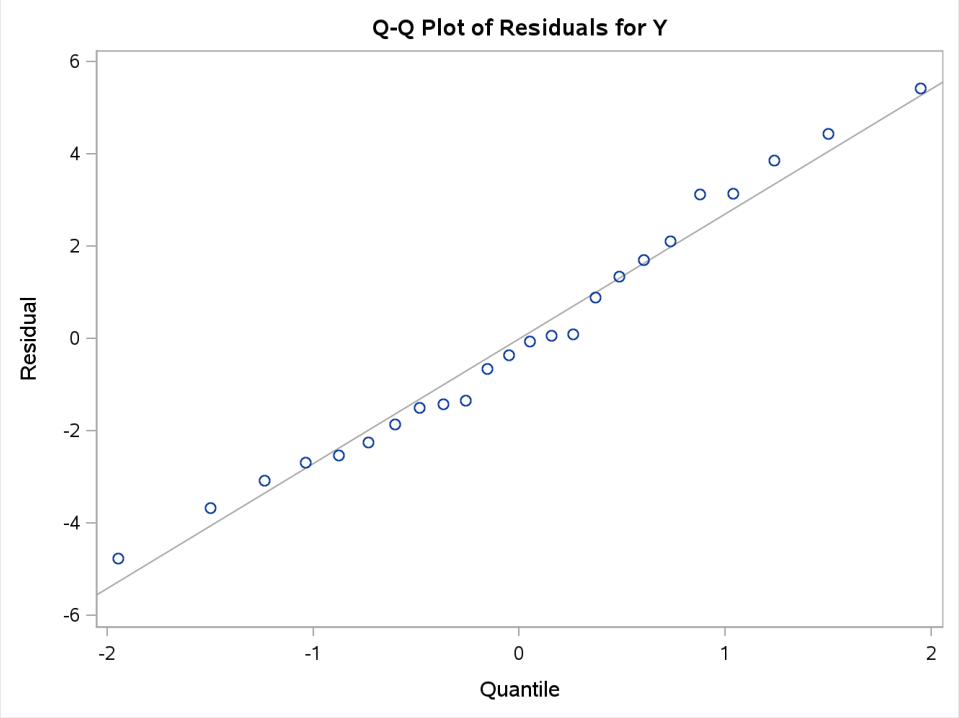
The first step in examining adequacy is to plot the regression equation over the scatterplot of observations. Since this is multiple regression, though, we don't have the tools to create this plot since there are two predictor variables and we would need a three dimensional plot. Instead, I will look at plots of each predictor variable against the residuals for Y. If the model represents a linear relationship, then there should be no discernable pattern in the plot. The residuals against variables X1 and X2 plots are below:



In the plot of residuals against X1, the data points look randomly scattered, which is what we are looking for to show normality. In the plot of residuals against X2, we see that the data points fall along two values on the x-axis – 1.0 and 1.5. Since there are only two possible values for X2, this is to be expected. What is important to glean from the plot, however, is the scatter of the data points within each x-axis value – there is no clustering along x2=1.5 and there is light clustering but good dispersion along x2=1.0. We can interpret the dispersion of residuals along the two values as a lack of pattern and feel confident that the residual plots show model linearity.

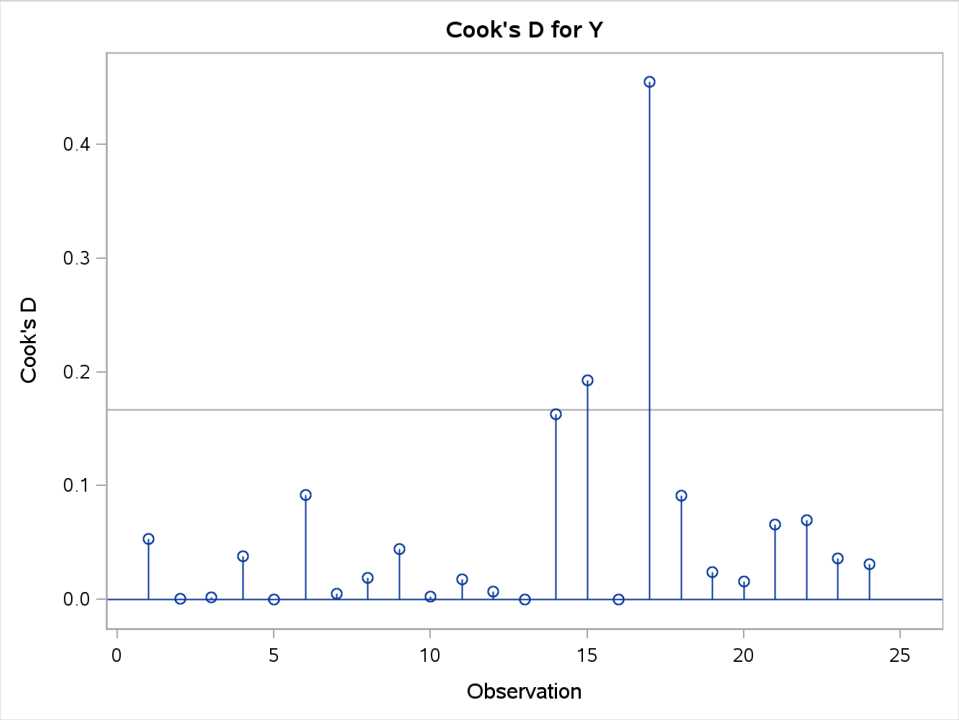
In addition, we can use the plot of the predictor variables against the residuals to examine the homogeneity assumption. The homogeneity assumption specifies that the errors have equal variance. If this assumption is true, the residual plots will have no discernible pattern. Since we have shown that the residual plots above show randomly scattered data points, we can confirm that this regression model meets the homogeneity assumption.

We also want to see if the normality assumption holds true – that the errors are normally distributed. A Quantile-Quantile (QQ) plot shows the standardized residuals by quantile, where a normal distribution can be seen if the plot resembles a straight line with equation y=x. The results of the QQ plot for this regression model are below:



In this plot, the residuals hug the straight line quite nicely and confirm the normality assumption.

We also want to look at the observations themselves and make sure that there are not any significant outliers or highly influential points that may skew the regression equation. For this purpose, we can use a plot of Cook's distance, which measures the difference between the regression coefficients when the full set of observations is used and when a specific observation is deleted. The results are below:



Cook's distance measures the influence of each observation. From the graph, observation 17 stands out from the other observations by a significant margin. Although the rule of thumb is to classify an observation with a Cook's distance of 1 or greater as influential, it is also reasonable to classify an observation as influential if its Cook's distance varies significantly from the other observations. With observation 17, the Cook's distance score is twice the size of the next largest score. I will conclude that observation 17 is influential and should be examined for error. If there is no error, we should consider whether data transformation is necessary or possible.

Finally, we can look at the Variance Inflation Factor (VIF) scores for each predictor variable in the regression model to see if there is any multicollinearity between variables. Collinear variables are those that are correlated to each other and may be redundant in a regression model. The VIF is the eigenvalue of the correlation matrix of the predictor variables. If the score is 10 or greater, the VIF indicates that there is collinearity between the variables. The results are below:

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

The VIF scores for both X1 and X2 are significantly below 10 and thus we can conclude that there is no multicollinearity between the variables.

Part 3:

Looking at the regression model that includes the predictor variables X1 and X2, there are some peculiarities about X2. First of all, it is not a qualitative or categorical variable, but it only contains two values 1.0 or 1.5. Typically discrete versus continuous

variables are defined as the whole integers versus values that can take any value, but a more accurate definition is that continuous variables can take any value and discrete variables have finite possibilities. By the first definition, X2 would be continuous since the values are 1 and 1.5, but I think it is more accurate to say that X2 is discrete since there are a finite number of possibilities for a variable such as number of bathrooms. The possibility of .5 bathrooms does not mean that number of bathrooms can be represented on a continuum because there will still be space between the values, even if that space is not a whole integer.

Since there are only two values for X2, we can treat it as a categorical variable and code a dummy variable for X2 for a new regression model. This changes the parameters of the regression model significantly. The regression equation for the model of X1 and X2 (Model 1) is: Y = 10.11 + 2.72*X1 + 6.10*X2

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

The regression equation for the model of X1 and X2 as a dummy variable (Model 2) is Y = 16.21 + 2.72*X1 + 3.05*dummyX2

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

While the regression equations vary significantly between Model 1 and Model 2, the diagnostic plots are exactly the same. This indicates not which model is a better fit, but instead that both models are adequate and meet the OLS assumptions discussed in Part 2 for Model 1.

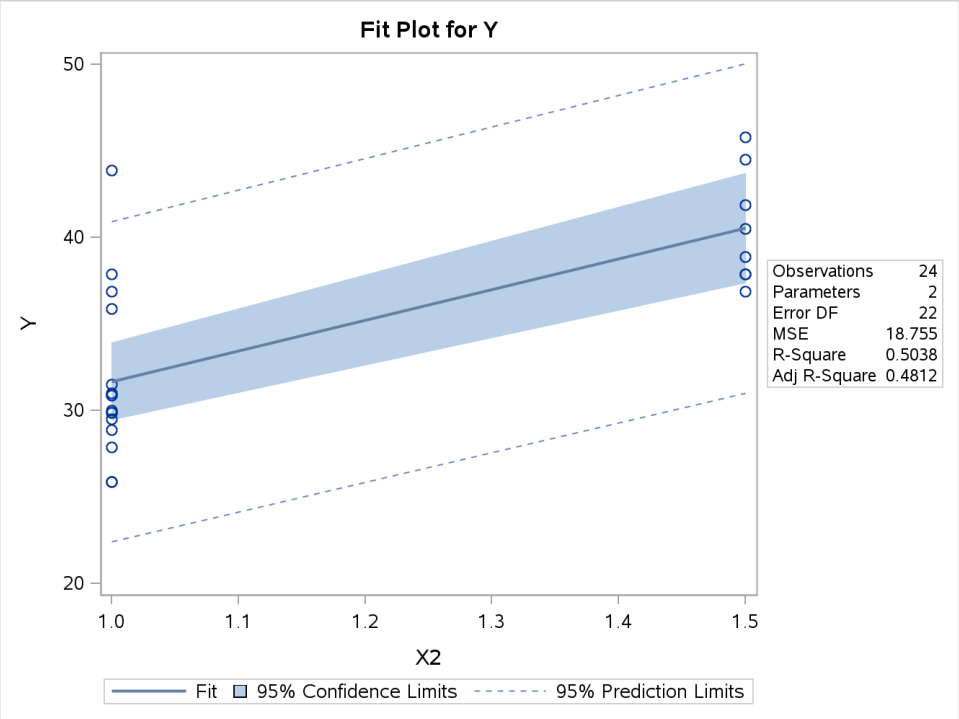
To determine which model is a better fit, we need to look at the indicators of goodness-of-fit from Part 1. This is difficult again, as the results for R^2 and adjusted-R^2 are exactly the same. My interpretation of these results is that neither model fits better and both models are adequate.

There is still the question, though, of whether it is better to fit a regression model with a binary variable as a continuous or dummy variable. It is possible that treating X2 as a continuous predictor variable could violate OLS assumptions. Assumptions about the predictors fall into three categories: the predictor variables are nonrandom, the values are measured without error, and the predictor variables are linearly independent of each other. I have already determined that X1 and X2 as a continuous predictor variable are not collinear in Part 2. The first two predictor variable assumptions are difficult to validate, though.

We can, however, look at whether the variable X2 violates any OLS assumptions by itself in a simple linear regression. The regression equation for simple linear regression with X2 is: Y = 13.95 + 17.73*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 first assumption to check is the linearity assumption that states that the regression model must be linear. To examine this, we can look at a fit plot that shows the fitted regression equation over a scatterplot of the observations. The results are below:



The scatterplot points do not correspond to the regression equation at all. Instead, the scatterplot points fall at two values of X2, 1.0 and 1.5. Thus we can conclude that a simple linear regression model with X2 does not even meet the basic linearity assumption and should be discounted. This does not, however, mean that X2 cannot be used as a predictor variable in multiple linear regression, just that X2 does not meet the most basic requirement for simple linear regression. No further examination into the assumptions is necessary since the linearity assumption is the most important assumption and X2 can't even meet that.

Conclusions:

In Part 1, I used three automated variable selection methods to find the optimal multiple linear regression model from the predictor variables X1-X9. Forward selection returned a 7 variable model while backward and stepwise selection returned a 2 variable model. Parsimony aside, the backward and stepwise model had better indicators of goodness-of-fit, namely AIC, BIC and Mallows' C_p and thus I chose that model as the optimal model. In Part 2, I analyzed the adequacy of the optimal multiple regression model and concluded that it met the major OLS regression assumptions. In Part 3, I compared regression models with variables X1 and X2, one where X2 was treated as a continuous

predictor variable and one where X2 was coded into a dummy variable and treated as a categorical variable. I concluded that both models fit adequately and had the same goodness-of-fit indicators. However, it is dangerous to use X2 as a continuous predictor variable since a simple linear regression model of X2 alone violates the most basic OLS assumption of linearity. Although OLS regression assumptions did not appear to be violated in the regression model with X1 and X2 as a continuous predictor variable, X2 should still be treated as a discrete variable and coded as a dummy variable

Code:

```
libname mydata
'/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;

data temp;
set mydata.building_prices;
run;

ods graphics on;

proc reg data=temp;
model y=x1;
title "Simple Linear Regression from Assignment 2";
run;


proc reg data=temp;
model y=x1-x9 / selection=forward slentry=0.5;
title "Variable Selection using Forward Selection";
run;


proc reg data=temp;
model y=x1-x9 / selection=backward slstay=0.1;
title "Variable Selection using Backward Selection";
run;


proc reg data=temp;
model y=x1-x9 / selection=stepwise slentry=0.15 slstay=0.15;
title "Variable Selection using Stepwise Selection";
run;


proc reg data=temp plots=diagnostics(stats=(default aic bic
cp));
model y=x1 x2 x4 x5 x6 x8 x9;
```

```
model y=x1 x2;
model y=x1;
title "Comparing Variable Selection Models";
run;


proc reg data=temp plots(only)=(diagnostics(unpack)
residualplot);
model y=x1 x2 / VIF;
title "Multiple Linear Regression";
run;


proc reg data=temp plots(only)=(diagnostics(unpack) residualplot
fitplot);
model y=x2;
title "Simple Linear Regression with X2";
run;


quit;

ods graphics off;


data temp;
set mydata.building_prices;
if (x2=1.5) then bath_dummy=1;
else bath_dummy=0;
run;

ods graphics on;

proc reg data=temp plots(only)=(diagnostics(unpack)
residualplot);
model y=x1 bath_dummy;
title "Changing X2 to a Dummy Variable";
run;


quit;

ods graphics off;
```