Assignment #3
Volunteer

Introduction:
This paper demonstrates a process by which data can be explored in a multiple variable environment, where our objective is to obtain a best-fit model with the best predictive power, that is consistent with Ordinary Least Square assumptions. This paper uses multiple regression models, examines residuals from those models, and uses testing techniques to examine influence and collinearity, to ensure that OLS assumptions are valid.

Using Automated Variable Selection
The three different automated variable selection procedures used to create multiple regression models include the forward, backward, and stepwise methods. In the data set analyzed, the backwards and stepwise approaches produced the same model, while the forward selection approach produced a different model, for two total different models. The forward selection method added predictor variables X1, X2, X9, X8, X5, X6, and X4, in this specific order (see below), until the last p value for the F statistic under 0.50 was added, which is the default setting for this method. The summary of results is presented below:

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 following table contains the Parameter Estimates for this model.

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

On a side note, of particular note when comparing the p values at each step is that the p values for variables already selected in the process change as the next variable is added. Please observe the output between steps 5 and 6, and note the change between the p values for 8:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.86041	3.97178	91.95461	12.18	0.0026
X1	2.00420	0.63603	74.97478	9.93	0.0055
X2	7.16531	3.31751	35.22378	4.66	0.0445
X5	1.35396	1.11383	11.15753	1.48	0.2398
X8	-0.07297	0.05120	15.33427	2.03	0.1712
X9	2.11510	1.38831	17.52590	2.32	0.1450

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	16.23626	4.78752	87.77521	11.50	0.0035
X1	2.37915	0.76330	74.14420	9.72	0.0063
X2	7.40772	3.34611	37.40326	4.90	0.0408
X5	1.77786	1.21490	16.34305	2.14	0.1616
X6	-1.02117	1.13525	6.17498	0.81	0.3809
X8	-0.04554	0.05983	4.42158	0.58	0.4570
X9	2.08987	1.39601	17.10347	2.24	0.1527

We see that the p value for X8 changes from 0.1712 to 0.4570 when the X6 variable is added into the model. This is likely due to collinearity between X8 and X6.

The backward selection process worked to eliminate variables whose p values were above 0.15, with the variables with the highest p value eliminated first. Variables eliminated, in order, were X6, X3, X8, X4, X9, X5, and X7 (see below). Note that it is variable X6 which is eliminated first, with a probability of 0.8200, vs. 0.3809 when added in the forward selection method, further evidence of some collinearity.

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

The remaining variables were X1 and X2, summarized below:

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

The stepwise method worked to add variables with p values < 0.15 and then to eliminate p values > 0.15 afterwards. Variables added, in order, were X1 and X2. No variables were removed in the last step. This process is summarized below:

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

Each procedure produced a different model primarily due to the different default p values in the criteria for adding or subtracting from the model. The stepwise model produced the same model as the forward selection model since they had the same criteria for adding variables to the model. A difference would only appear if the addition of a variable caused an existing variable to have an increased p value in the process at a level above the elimination threshhold.

Comparing Models using Adjusted R-Square, Mallows Cp, AIC, and BIC

If we measure the predictive value simply by the R-square, then all of the models are more predictive than the simple linear regression of X1 on Y, which produced an R-Square of 0.7637. The forward selection model produced an R-Square of 0.8491, which is the best R-Square model; the backward selection model produced an R-Square of 0.7981; and the stepwise selection model also produced an R-Square of 0.7981. The R-Square is a good starting point, but the residuals continue to need attention when evaluating the models, to ensure that the residuals continue to exhibit normality and homogeneity. In addition, we do not want to overfit a model so as to lose its predictive value. Towards this end, when comparing regression models, the Adjusted R-Square is a better measure than R-Square (Cody, 2011, p. 137) as it allows us to compare different sized models, and we can use Mallows Cp, AIC, and BIC to help find the best parsimonious model. According to Cody (2011, p. 142) “Mallows suggested that you choose the first model in which Cp is less than or equal to p.” Also, we should look for the models with the lowest AIC and BIC scores (while we should be relatively indifferent towards those within 2 points of each other). The output for the 10 best R-Square models with Cp, AIC, and BIC scores is as follows:

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	Variables in Model
4	0.7968	0.8321	1.7954	52.2572	58.9118	X1 X2 X5 X7
5	0.7950	0.8396	3.0912	53.1627	61.5496	X1 X2 X5 X7 X9
5	0.7944	0.8391	3.1353	53.2326	61.5749	X1 X2 X3 X5 X7
5	0.7938	0.8387	3.1801	53.3036	61.6006	X1 X2 X5 X6 X9
5	0.7927	0.8378	3.2645	53.4366	61.6489	X1 X2 X4 X5 X7
5	0.7911	0.8365	3.3785	53.6152	61.7141	X1 X2 X5 X8 X9
6	0.7911	0.8456	4.5242	54.2435	64.5260	X1 X2 X4 X5 X7 X9
6	0.7889	0.8440	4.6798	54.4993	64.5816	X1 X2 X5 X6 X8 X9
6	0.7888	0.8439	4.6880	54.5126	64.5846	X1 X2 X5 X7 X8 X9
6	0.7883	0.8435	4.7232	54.5700	64.5972	X1 X2 X4 X5 X6 X9

According to this approach, the first model would be the best, since the Cp score of 1.7854 is less than the number of predictors (including the intercept), which would equal 5. The model would include predictors X1, X2, X5, and X7, and would have an Adjusted R-Square of 0.7968. However, according to Cody, (2011, p.142) Hocking recommended that you choose the first model where Cp is less than or equal to 2p-pfull+1" if you are using the model to explain the relationships among the variables, vs. using the model for predictive purposes. If we were using the model for descriptive purposes, then the best model would be using X1, X2, X5, X7, and X9 (because 2*6-10+1=3 and the Cp score rounds down to 3), with an Adjusted R-Square of .7950. Since this class is focused on Predictive Analytics, we will assume the model that includes predictors X1 (taxes), X2 (# bathrooms), X5 (# garage stalls), and X7 (# bedrooms). It is important to evaluate the usefulness of the Cp statistic, however, as its value is diminished if a good estimate of σ^2 is not available. To evaluate whether it is appropriate to use Cp, we refer back to the Cp results disaplayed in the summary results from the forward selection approach. For models with increasing p, the Cp score continues to increase. That pattern suggests that Cp is a good tool as each of the variables under consideration has some "explanatory power". (Chatterjee and Hadi, 2012, p.313) In any case, the AIC and BIC scores confirm that the model containing X1, X2, X5, and X7 contain the lowest scores in each case, and the BIC metric in particular indicates that this model is better than the next best since it is more than 2 points different. The ANOVA table for this model is below:

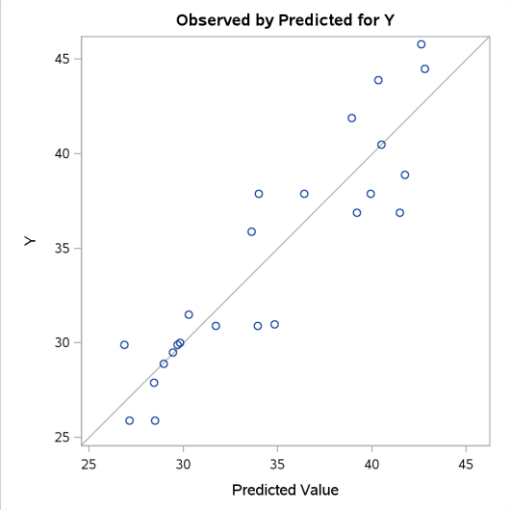
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	691.91122	172.97781	23.54	<.0001
Error	19	139.59836	7.34728		
Corrected Total	23	831.50958			

Root MSE	2.71059	**R-Square**	0.8321
Dependent Mean	34.62917	**Adj R-Sq**	0.7968
Coeff Var	7.82747		

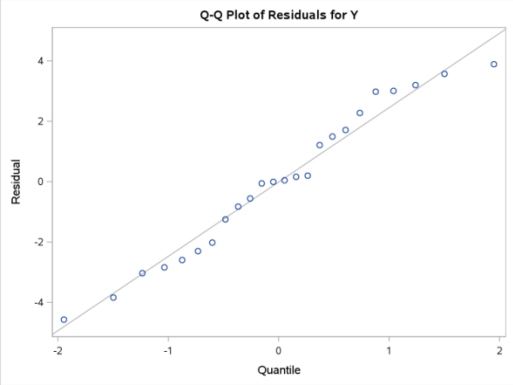
Determining Goodness-Of-Fit for the Selected Model

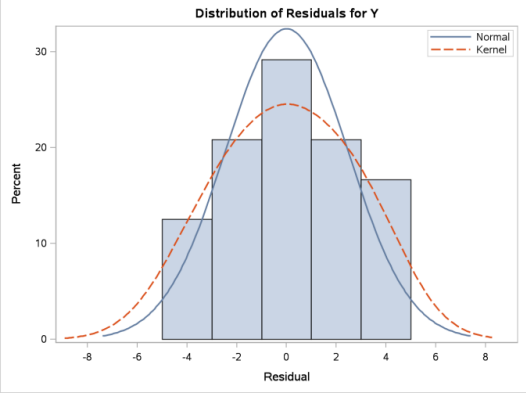
Examining the Residuals

As part of checking the model for adequacy, we plot the fitted regression model over the scatterplot. As evidenced in the scatterplot below, the residuals appear to be homogenous and normal.

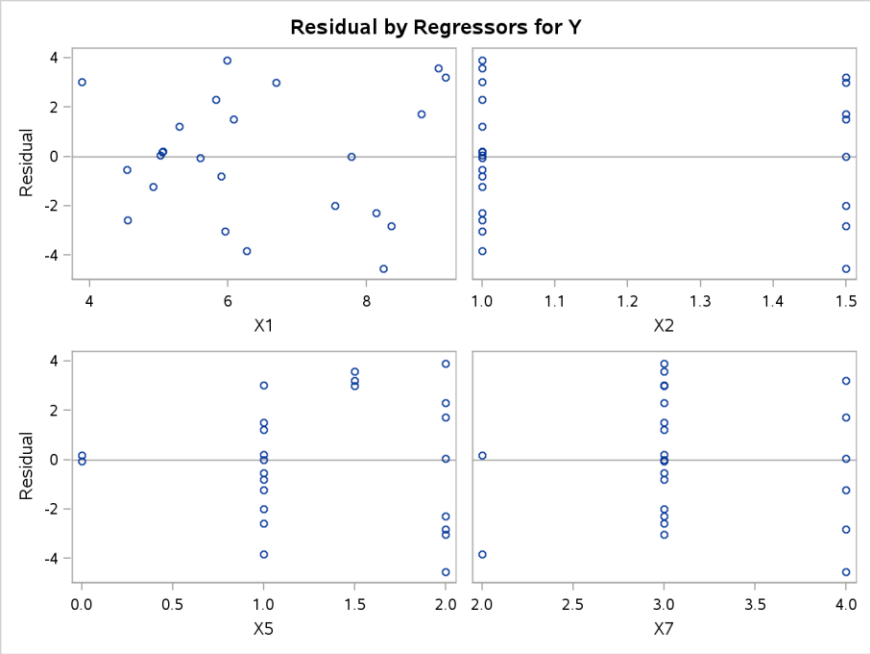


Further evidence of the normality of the residuals is seen below in the Quantile-Quantile plot (QQ plot), as well as in the distribution of residuals.



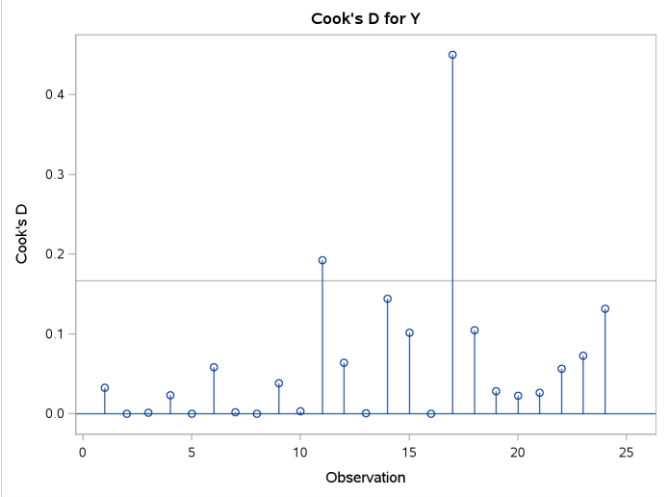


We can also see in the plots below that on an individual predictor variable level, that residuals are homogenous and have constant variance.



Examining Influence

We also need to check for potential outliers using Cook's Distance. We observe one points that as a candidate for having influence as its distance is further on a relative basis, but even the highest figure is below the standard benchmark of 1 and the suggested benchmark of 50% for the F test statistic (Chatterjee and Hadi, 2012, p. 112). Thus, it is not deemed to have undue influence. The SAS output is presented as follows:



Testing for Collinearity

According to Cody (2011, p. 156) Variance Inflation Factor values "that are greater than 10 are considered large," and "you should also pay attention to VIF values between 5 and 10." None of the VIF values in the model below are above 10, so we do not see much collinearity among these variables.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13.62125	3.67252	3.71	0.0015	0
X1	1	2.41230	0.52251	4.62	0.0002	2.13888
X2	1	8.45895	3.33004	2.54	0.0200	2.01238
X5	1	2.06036	1.22350	1.68	0.1085	1.71271
X7	1	-2.21545	1.29006	-1.72	0.1022	1.66108

Using Dummy Variables

Let's consider a model that contains only taxes in thousands of dollars (X1) and number of bathrooms (X2). We can fit a model that treats X2 as a continuous predictor variable, or a model that treats X2 through the use of a dummy variable. Let's compare the models for each and examine the differences. These two models look quite similar. In fact, the ANOVA, R-Square, and Adjusted R-Square all show the same results, as follows:

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			

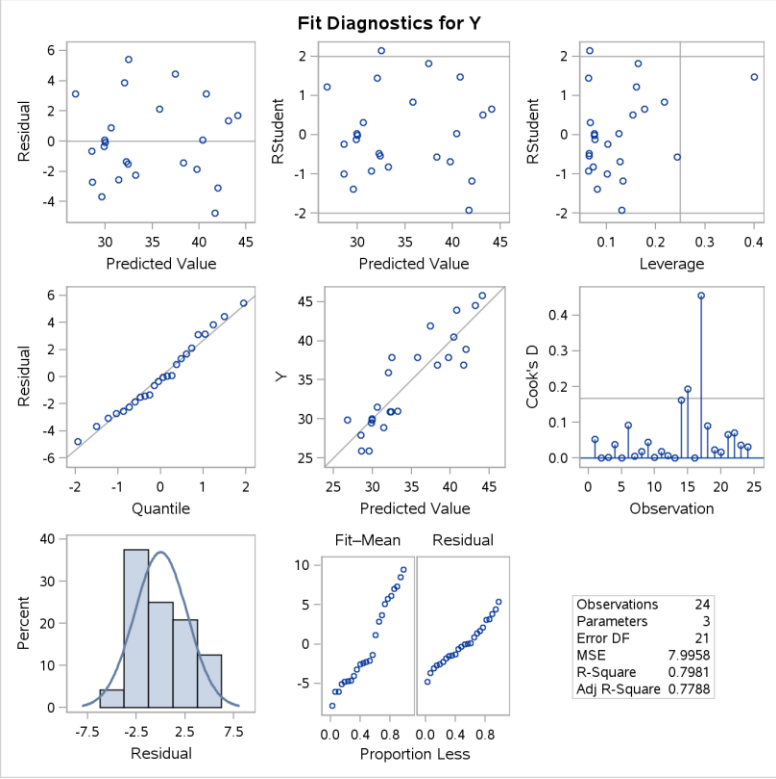
Root MSE	2.82768	R-Square	0.7981
Dependent Mean	34.62917	Adj R-Sq	0.7788
Coeff Var	8.16562		

The Parameter Estimates and standard errors, however, appear different (see below). This means that the intercept for the actual model has changed from 10.1120 to 16.2105 and that a change in each unit of X2 predicting a 6.0985 value change in Y as changed to a dummy variable change from 0 to 1 will predict a 3.0493 value change in Y. Interestingly enough, from a significance standpoint, both versions have identical t values and probabilities.

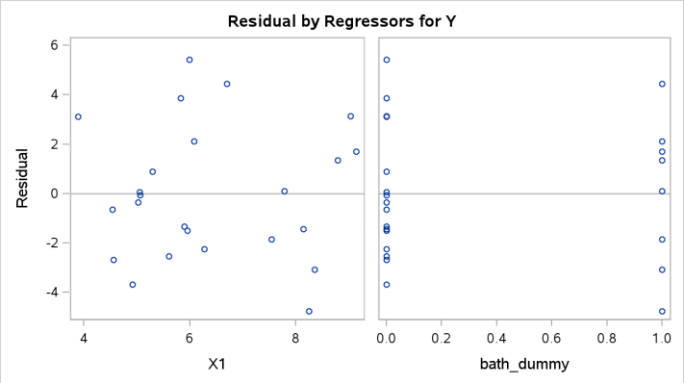
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.11203	2.99614	3.38	0.0029
X1	1	2.71703	0.49115	5.53	<.0001
X2	1	6.09851	3.22705	1.89	0.0727

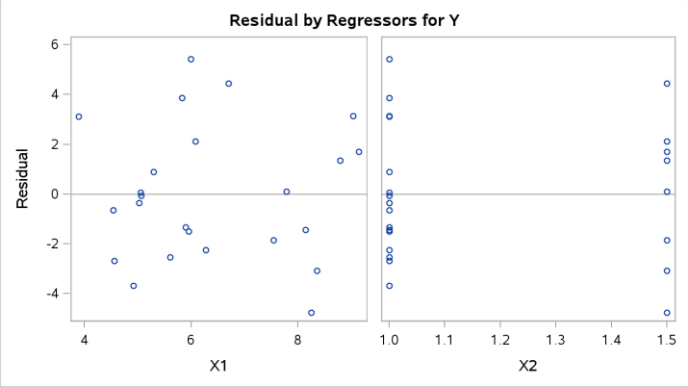
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16.21054	2.88345	5.62	<.0001
X1	1	2.71703	0.49115	5.53	<.0001
bath_dummy	1	3.04925	1.61353	1.89	0.0727

In addition, the residuals and patterns therein are identical.



While the values of the second variable, whether dummy or X2, represented in the residual by regressor output, are different, the patterns of the residuals remain unchanged (see below).





The homes in the population have either 1 or 1.5 baths. The dummy variable assigns a value of 1 or 0. One could argue that the number of baths is a continuous variable since a house could have any number of baths, and a half bath is not as valuable as a full bath, so the values are continuous. One could also argue that a half bath is not exactly worth half a bath and that the value added of a half bath vs. a full bath is somewhat unknown, so you might observe problems in the residuals if this variable is not optimally continuous. Either way, it does not appear that any of the OLS assumptions are violated. Both sets of residuals show that the errors are independently and identically distributed, and that there is normality and homoscedasicity.

Conclusions:

Selecting the best model is both an exhausting and iterative process. There are several approaches to the problem and many different 'right' answers. I believe the best results are achieved by using and analyzing results from all of the approaches highlighted in this paper. In doing so, we evaluate several 'candidate' models, and can compare their Cp, AIC, and BIC metrics to determine which models are optimal and parsimonious, we measure Cook's D for influence, and we have evaluated plots of residuals to ensure that none of the OLS assumptions are violated for the model selected. I believe that the model of X1, X2, X5, and X7 is a good fit based on the measures discussed in this paper, and we can see that treating X2 as a dummy variable can influence what the parameters look like, but that it does not influence the OLS assumptions.

Code:

In order to explore the data to find the best fit single variable, we can use this code:

```
libname mydata '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
data temp;
      set mydata.building_prices;
run;
proc reg data=temp;
model y = x1-x9 / selection=rsquare start=1 stop=1;
run;
```

The code to find the optimal regression model using the automated variable selection process (for multiple regression) for forward, backward, and stepwise methods (respectively), with default values for SLENTRY and SLSTAY is:

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

proc reg data=temp;
model y = x1-x9 / selection=forward slentry=0.5;
run;

proc reg data=temp;
model y = x1-x9 / selection=backward slstay=0.1;
run;

proc reg data=temp;
model y = x1-x9 / selection=stepwise slentry=0.15 slstay=0.15;
run;
```

The code to find the best parsimonious model using Mallows' Cp, AIC, and BIC is:

```
libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
data temp;
      set mydata.building_prices;
proc reg data=temp;
model y=X1-X9 / selection=adjrsq cp aic bic best=10;
run;
```

To check the selected model for adequacy (goodness-of-fit), and to obtain the Variance Inflation Factors:

```
libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
data temp;
      set mydata.building_prices;
ods graphics on;
proc reg data=temp plots=(fitplot residuals diagnostics);
model y=x1 X2 X5 X7 / VIF;
run;
ods graphics off;
```

To obtain same, using unpack option:

```
libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
data temp;
      set mydata.building_prices;
ods graphics on;
proc reg data=temp plots (unpack)=(fitplot residuals diagnostics);
model y=x1 X2 X5 X7 / VIF;
run;
ods graphics off;
```

To test the model using predictor variables X1 and X2, so as to compare the model to the version with a dummy variable for X2:

```
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 X2;
run;
quit;
ods graphics off;
```

To use a dummy variable for X2:

```
libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/'
access=readonly;
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;
model y=X1 bath_dummy;
run;
quit;
ods graphics off;
```