Assignment #2

Introduction:

The purpose of this assignment is to generate the best simple linear regression model for the building_prices data set and assess the adequacy of this model. First, we will create a model utilizing X1 as our single predictor variable. We will use X1 since this is expected to be the best single predictor based on the exploratory data analysis (EDA) that we performed in assignment 1. Second, we will let SAS choose the best single predictor variable based on the R-square values for each possible simple linear regression model. Finally, we will determine the adequacy of this optimal regression model by reviewing diagnostic and residual plots.

Results:

To generate regression parameters for a simple linear regression model utilizing X1 as the predictor variable and Y as the response variable, I used the PROC REG process in SAS. The parameter estimate output for the model is listed below in Figure 1.

Figure 1: Linear regression model parameter estimates

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

Utilizing the SAS output in Figure 1, we estimate the simple linear regression model to be as follows:

$$Y = 13.3553 + (3.3215 * X1)$$

Utilizing the R-square method in SAS to determine the single predictor variable that will result in the simple linear regression model with the highest R-square value we confirm that X1 was the correct choice. The summary of this variable selection procedure is shown in Figure 2 on the following page. In this summary table the variables are ranked from highest to lowest by R-square value showing X1 to be the best predictor and X9 to be the worst. Since X1 was found to be the best predictor variable through the R-square method in addition to our EDA process, the regression model created through the output in Figure 1, and listed on the line below, remains valid.

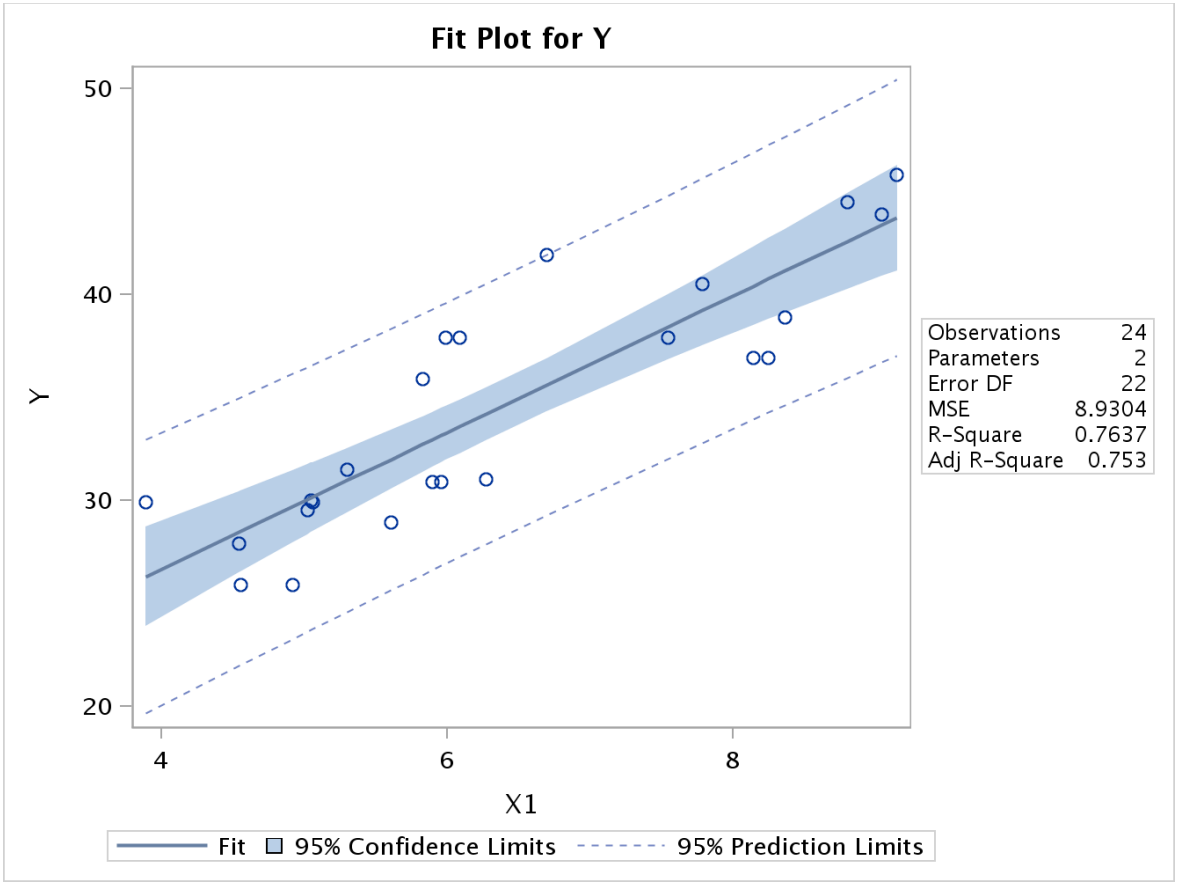
$$Y = 13.3553 + (3.3215 * X1)$$

Figure 2: Variable selection summary table

Number in Model	R-Square	Variables in Model
1	0.7637	X1
1	0.5038	X2
1	0.5009	X4
1	0.4194	X3
1	0.2793	X6
1	0.2130	X5
1	0.1579	X8
1	0.0793	X7
1	0.0712	X9

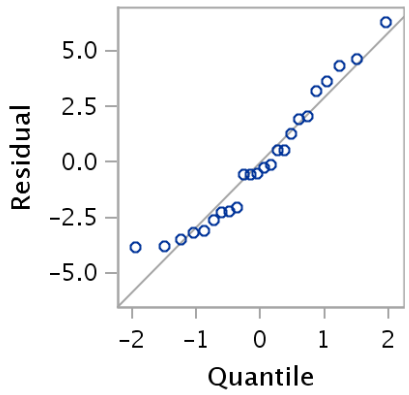
To assess the adequacy of the model we will first look at the scatter plot overlaid with the line generated by the regression model which is shown in Figure 3. The points on the scatter plot are grouped fairly close around the regression line which is an indicator that the model does a good job of explaining the relationship between the variables. Also, all of the points lie on or within the dashed line representing the 95% confidence range for individual observations.

Figure 3: Scatter plot with regression line



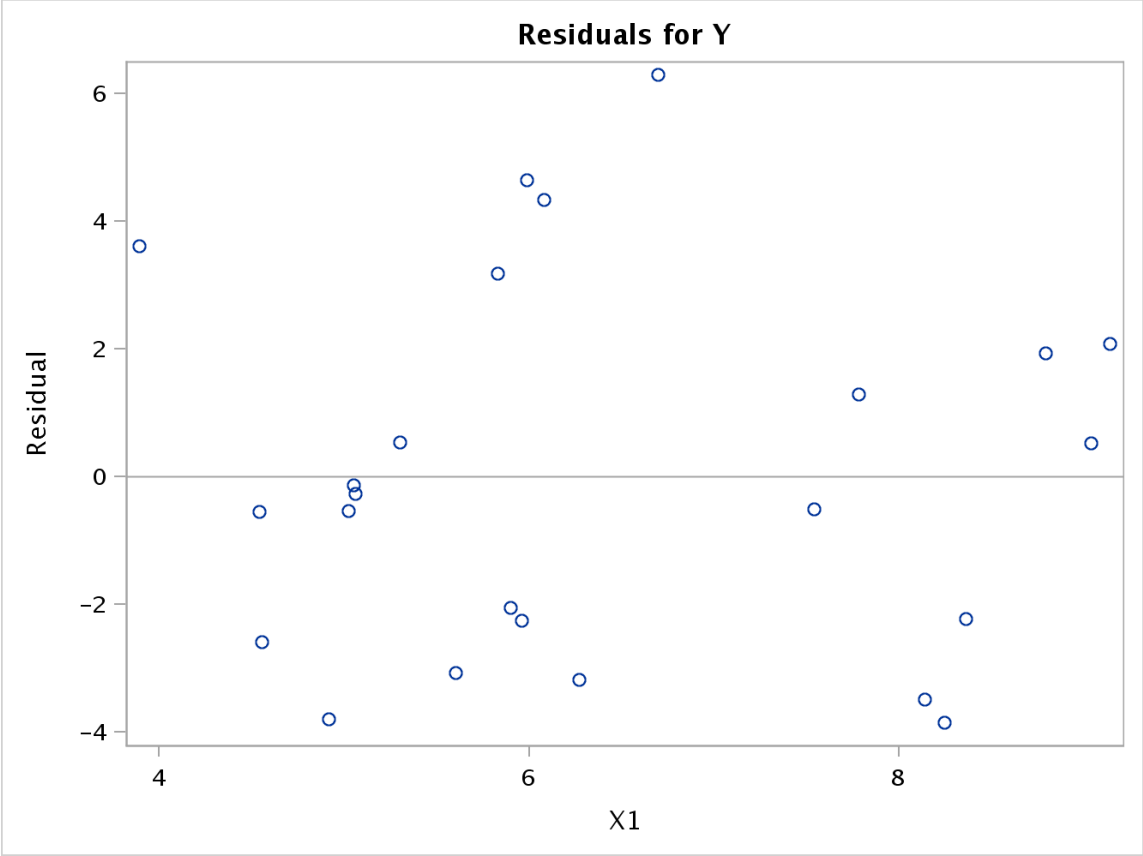
Our next check of model adequacy will be to look at the quantile-quantile plot (QQ plot) shown below in Figure 4. The better the fit of the model is, the closer the points on the QQ plot will be to the diagonal line shown on the graph. In this case, the points are fairly close to the diagonal line, but the fit shows room for improvement.

Figure 4: QQ plot



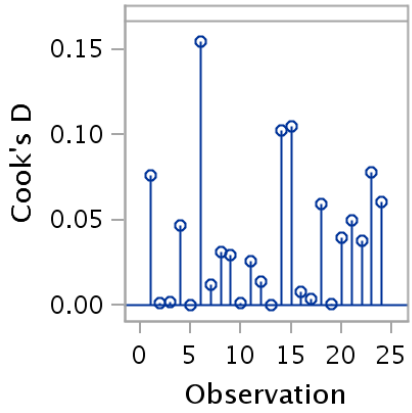
The scatter plot of the predictor variable X1 plotted against the residuals from our regression model, which is shown in Figure 5, constitutes our next check for model adequacy. If the key assumptions for linear regression hold true, then the points on this scatter plot should be randomly distributed with no discernable pattern. The points on the graph in Figure 5 do appear to be random which helps confirm the assumptions of normality and linearity.

Figure 5: Scatter plot of X1 and residuals



Our final check for model adequacy will be to look at the plot of Cook's distances for our model. Looking at the Cook's distance figures will help us to identify potential outliers or influential observations. The larger the Cook's difference is, the larger the difference between the regression parameter values when that particular observation is included in the calculations and when that observation is excluded. The plot of Cook's distance for each observation is shown below in Figure 6. The Cook's distance for observation number 6 stands out from the rest of the values. To a lesser degree, observations 14 and 15 also stand out. These observations can be examined further in an attempt to understand the cause of these influential points which may lead to a better understanding of the data and ultimately a better model. However, we must keep in mind that these are still small values for Cook's distance and are not cause for alarm.

Figure 6: Plot of Cook's distance for each observation



Conclusions:

Using the R-square variable selection process, we were able to confirm our conclusions from the EDA that we conducted in assignment 1 stating that X1 would be the best single predictor variable that we could use to build a simple linear regression model. We utilized SAS to calculate the regression parameter for our regression model and produce several graphical tools to evaluate the adequacy of our model. Both our numerical and graphical measures of model adequacy showed that the model we created does a fairly good job of explaining the variability in our response 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;
run;
quit;

proc reg data=temp plots = (fitplot diagnostics residualplot);
      model y = x1-x9 / selection=rsquare start=1 stop=1;
run;
quit;
ods graphics off;
```