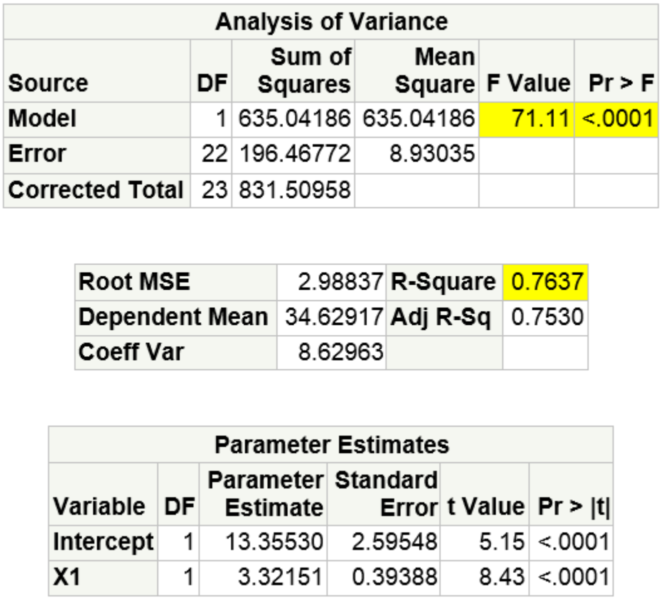
Assignment #2
James Gray

Introduction:

The purpose of this assignment is to fit a simple regression model to a predictor that was identified by an exploratory data analysis as having the highest correlation to the response variable Y in the building_prices data set. The predictor X1, taxes, will be used to fit the simple regression model and predict the price of a house. Multiple regression models are then generated for eight other predictors in the building_prices data set to validate that the selected predictor is indeed the variable with the optimal regression model. An assessment in then performed to evaluate model adequacy using various tests and visualizations.

Results:

The regression results using X1 as the predictor are shown in Figure 1. The model is significant at the p<.0001 level and the R-Square of 0.7637 represents the proportion of variability that is explained by the regression model.



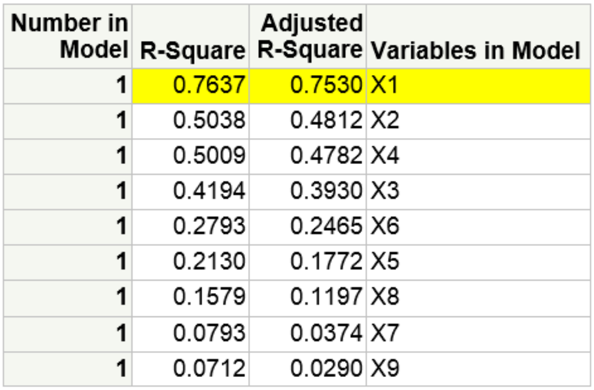
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	635.04186	635.04186	71.11	<.0001
Error	22	196.46772	8.93035		
Corrected Total	23	831.50958			

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

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

Figure 1 - X1 Regression Model

Regressions were run for all nine predictors (X1-X9) using the PROC REG RSQUARE selection method. The stack ranked R-Square results are shown in Figure 2 and X1 represents the optimal regression. This confirms that the variable selected by the exploratory data analysis was the indeed the optimal regression model.



Number in Model	R-Square	Adjusted R-Square	Variables in Model
1	0.7637	0.7530	X1
1	0.5038	0.4812	X2
1	0.5009	0.4782	X4
1	0.4194	0.3930	X3
1	0.2793	0.2465	X6
1	0.2130	0.1772	X5
1	0.1579	0.1197	X8
1	0.0793	0.0374	X7
1	0.0712	0.0290	X9

Figure 2 - Multiple Regressions using PROC REG rsquare option

An assessment of the model adequacy was conducted using multiple techniques and visualizations. The adequacy of regression models are predicated on a set of assumptions and the application of the model may result in errors if these assumptions are violated.

Linearity Assumption

The scatterplot (Figure 3, graph 1) shows the fitted regression line with the mapping of X1 and Y data points from the data set. The data points are evenly dispersed around the regression line and this confirms the linearity assumption that X1 and Y vary linearly.

Linearity is also confirmed by plotting the residual error against the response variable Y (Figure 3, graph 2). This graph does not indicate a non-linear relationship.

The scatterplot of residuals by predictor X1 in Figure 4 shows a random scatter thereby confirming that these variables are uncorrelated. This confirms the linearity assumption is valid.

The fit plot in Figure 5 also confirms the linear relationship between X1 and Y. The blue inner band represents there is 95% confidence that given X1 the mean response variable Y is within this band. The outer blue dashed lines represent the 95% confidence interval for the data set observations.

Normality Assumption

The Quantile-Quantile plot (Figure 3, graph 3) is a graphical method that confirms that data come from a normal distribution when the data points cluster closely to the straight line. The residual errors not explained by the regression model are tightly dispersed around the line thereby confirming the errors have a normal distribution. This confirms the normality assumption.

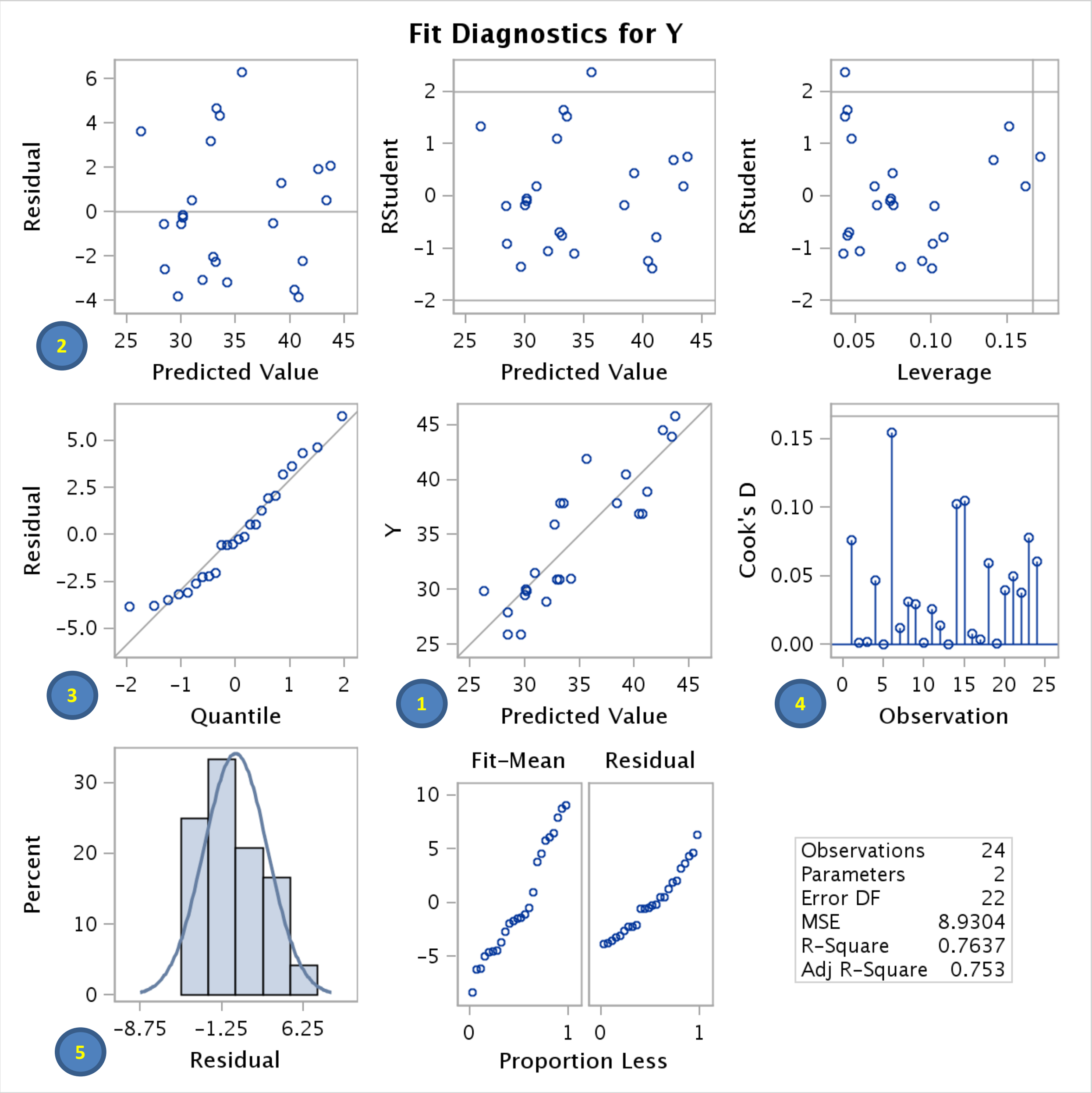
Normality is also confirmed by the frequency histogram is Figure 3, graph 5 that shows a normal distribution of the residuals.

Constant Variance Assumption

Figure 4 is a plot of the residual error by predicted value (X1). The errors do not appear to show a pattern as X1 varies and the scatter is fairly even around the zero line. This confirms the constant variance assumption that errors should be constant over the range of the predictor variable.

Outliers

Analysis is required to ensure that model fitting is not heavily influenced by one or a few data points. The Cooke's distance metric measures the influence of each observation and identifies any highly influential observations above a critical line (Figure 3, graph 4). The plot does not appear to show any observations above the threshold. Additional residual analysis was conducted by plotting studentized residuals by leverage (Figure 5). Observation #15 was identified as an outlier and observation #24 was identified as highly influential. The regression model may improve fit if these two observations are removed from the model.



Observations	24
Parameters	2
Error DF	22
MSE	8.9304
R-Square	0.7637
Adj R-Square	0.753

Figure 3 – Model Fit Diagnostics

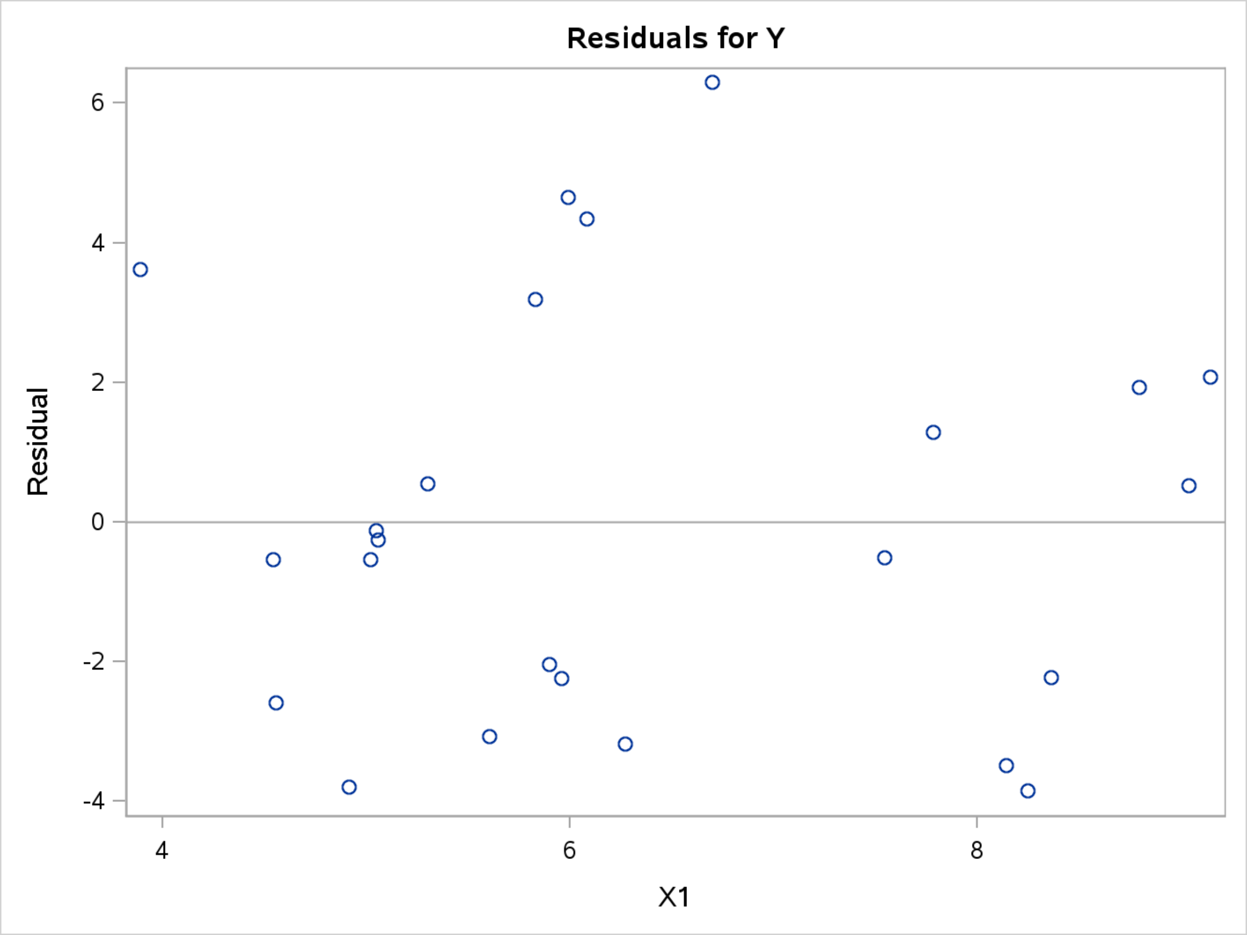


Figure 4 - Residual by Predictor X1 plot to test Normality

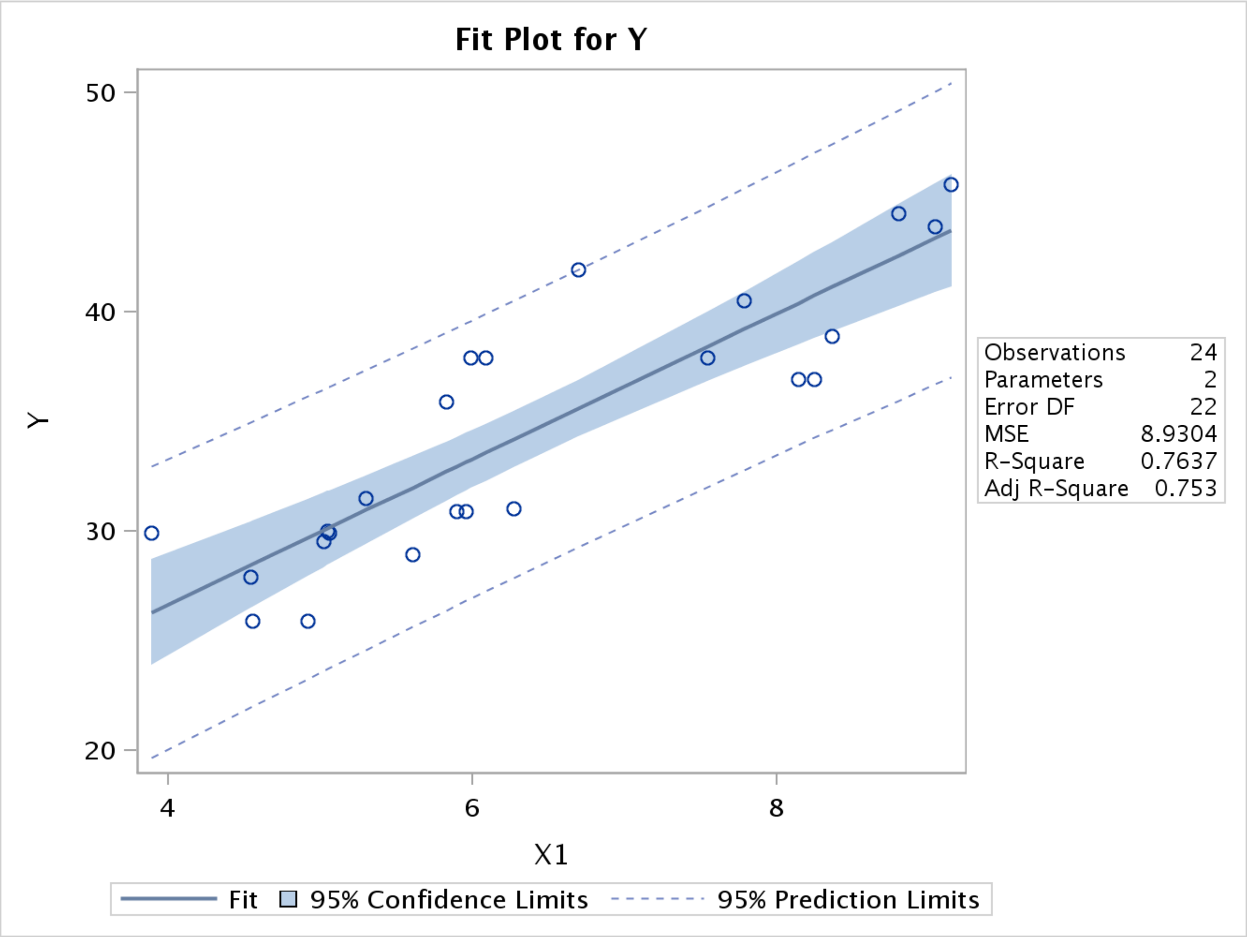


Figure 5 - Fit Plot for Y

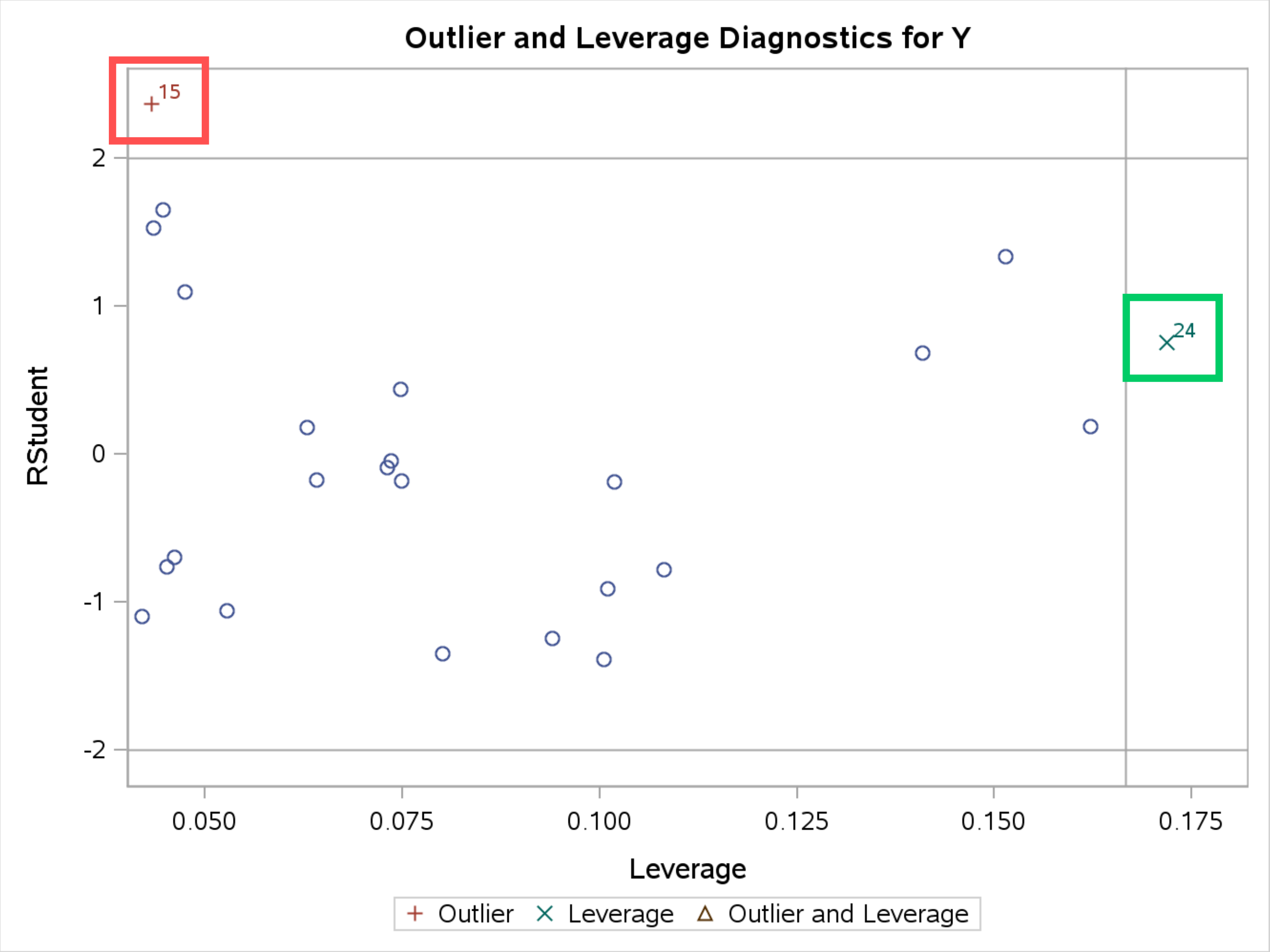


Figure 6 - Outlier and Leverage Diagnostics for Y

Conclusions:

This assignment confirmed that the predictor X1 identified by Exploratory Data Analysis as likely the best simple regression model was indeed the optimal regression model. The multiple regression variable selection procedure provided a stack ranking of R-square coefficients that is the proportion of the response variable Y that is explained by each predictor. The regression model's predictive capability is only valid if it does not violate a set of assumptions. The linearity, normality and constant variance assumptions were all confirmed as valid using graphical techniques. The analysis also identified that observations 15 and 24 were highly influential on the model and a better fit may exist if these two observations are removed. Additional analysis should be conducted to understand why these outlier and influential data points exist before discarding the observations.

Code:

```
/*          James Gray
            2013.07.05
            graymatter@u.northwestern.edu
            Assignment2_JG.sas
*/

/*          This code is for PREDICT 410 Assignment #2 - Single Regression Model. The code will
            build a simple regression model based on the EDA from Assignment #1. Other predictors
            will be evaluated by using the RSQUARE option of PROC REG to determine if the predictor
            selected by EDA is indeed the optimal regression model. A series of model accuracy processes
            will be executed to evaluate adequacy. The same dataset from Assignment #1 will be used.
*/

*******************************************************************************************************;
* Get the data on the SAS server - mydata.building_prices - Regression by Example pg. 328-9
* Y = Sales price of the house (thousands of dollars)
* X1 = Taxes (thousands of dollars)
* X2 = Number of bathrooms
* X3 = Lot size (thousands of feet)
* X4 = Living space (thousands of feet)
* X5 = Garage stalls (#)
* X6 = Rooms (#)
* X7 = Bedrooms (#)
* X8 = Age of the home (years)
* X9 = Fireplaces (#)
*******************************************************************************************************;
libname mydata '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
run;

*******************************************************************************************************;
* Fit a regression model using X1(Taxes) as the best predictor based on the EDA
*******************************************************************************************************;
proc reg data=mydata.building_prices;
            model Y = X1; * generate the model using X1 as the predictor;
run;

*******************************************************************************************************;
* Run all possible regressions with the predictors (X1-X9)
*******************************************************************************************************;
ods graphics on;
* fit a regression and generate plots;
proc reg data=mydata.building_prices plots = (fitplot diagnostics residuals cooksd(label)
            RStudentLeverage(label));
            id Obs;
            * evaluate all predictors, limit model to a simple model with one predictor;
            model Y = X1 X2 X3 X4 X5 X6 X7 X8 X9 /
            selection = rsquare adjrsq start=1 stop=1; * run regressions for each predictor only;
run;
quit;
ods graphics off;
```

```
**************************************************************************************************;
* END
**************************************************************************************************;
```