Assignment #1

James Gray

Introduction:

The purpose of this assignment is to evaluate the predictive strength of nine variables on the sales price of a home. The predictors (X1 - X9) and response variables are as follows:

Y = Sales price of the house (thousands of US dollars)
X1 = Taxes (thousands of dollars)
X2 = Bathrooms (number)
X3 = Lot size (thousands of feet)
X4 = Living space (thousands of feet)
X5 = Garage stalls (number)
X6 = Rooms (number)
X7 = Bedrooms (number)
X8 = Age of the home (years)
X9 = Fireplaces (number)

Pearson correlation coefficients and a scatterplot matrix will be used to conduct this analysis and identify the predictor variable (X) that is the best single predictor on the sales price of the house (Y). The Pearson correlation coefficient measures both the direction and strength of the linear relationship between Y and X. A correlation greater than zero indicates a positive relationship between the variables, while a correlation less than zero indicates an inverse relationship. Correlations closer to 1 or -1 represent a stronger relationship between Y and X.

A second scatterplot with a Locally Estimated Scatterplot Smoother (LOESS) for Y with each of the nine predictor variables will be used to plot a fitted curve in addition to the linear regression line. The LOESS curve is a locally weighted polynomial regression that gives more weight to a subset of data that is close to where response variable is being estimated. The LOESS curve is useful to mitigating the impact of outliers and creates a robust model.

Results:

Figure 1 below outlines the correlation coefficients between each predictor (X1 – X9) and Y. The top number in column Y for each row represents the relative strength between the predictor and sales price of the house. The relative strength that each predictor has on the sales price of the house from highest to lowest is as follows:

1. Taxes
2. Bathrooms
3. Living space
4. Lot size

5. Rooms
6. Garage stalls
7. Age of the home (negatively correlated)
8. Bedrooms
9. Fireplaces

All of the predictors have a positive correlation with sales prices except for the age of the home (X8) that is negatively correlated.

Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0										
	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**Y**
X1	1.00000	0.65127 0.0006	0.68921 0.0002	0.73427 <.0001	0.45856 0.0242	0.64062 0.0007	0.36711 0.0776	-0.43710 0.0327	0.14668 0.4940	0.87391 <.0001
X2	0.65127 0.0006	1.00000	0.41296 0.0449	0.72859 <.0001	0.22402 0.2926	0.51031 0.0108	0.42640 0.0377	-0.10075 0.6395	0.20412 0.3387	0.70978 0.0001
X3	0.68921 0.0002	0.41296 0.0449	1.00000	0.57155 0.0035	0.20466 0.3374	0.39212 0.0581	0.15161 0.4795	-0.35275 0.0909	0.30599 0.1459	0.64764 0.0006
X4	0.73427 <.0001	0.72859 <.0001	0.57155 0.0035	1.00000	0.35888 0.0850	0.67886 0.0003	0.57434 0.0033	-0.13909 0.5169	0.10656 0.6202	0.70777 0.0001
X5	0.45856 0.0242	0.22402 0.2926	0.20466 0.3374	0.35888 0.0850	1.00000	0.58939 0.0024	0.54130 0.0063	-0.02017 0.9255	0.10162 0.6366	0.46147 0.0232
X6	0.64062 0.0007	0.51031 0.0108	0.39212 0.0581	0.67886 0.0003	0.58939 0.0024	1.00000	0.87039 <.0001	0.12427 0.5629	0.22222 0.2966	0.52844 0.0079
X7	0.36711 0.0776	0.42640 0.0377	0.15161 0.4795	0.57434 0.0033	0.54130 0.0063	0.87039 <.0001	1.00000	0.31351 0.1358	0.00000 1.0000	0.28152 0.1826
X8	-0.43710 0.0327	-0.10075 0.6395	-0.35275 0.0909	-0.13909 0.5169	-0.02017 0.9255	0.12427 0.5629	0.31351 0.1358	1.00000	0.22578 0.2888	-0.39740 0.0545
X9	0.14668 0.4940	0.20412 0.3387	0.30599 0.1459	0.10656 0.6202	0.10162 0.6366	0.22222 0.2966	0.00000 1.0000	0.22578 0.2888	1.00000	0.26688 0.2074
Y	0.87391 <.0001	0.70978 0.0001	0.64764 0.0006	0.70777 0.0001	0.46147 0.0232	0.52844 0.0079	0.28152 0.1826	-0.39740 0.0545	0.26688 0.2074	1.00000

Figure 1 - Pearson Correlation Matrix

While the Pearson correlation coefficients in Figure 1 do provide a useful measure of the strength and direction of the relationship between Y and X, an analysis of the scatterplot is needed to indeed confirm the linear relationship exists. Column Y in Figure 2 below displays the scatterplot of each predictor on the sales price of the home. The X1 data points (taxes) highlighted in the red box do show a relatively straight line relationship with Y (sales price). Therefore "taxes" is the best single predictor of sales prices due to its highest correlation coefficient.

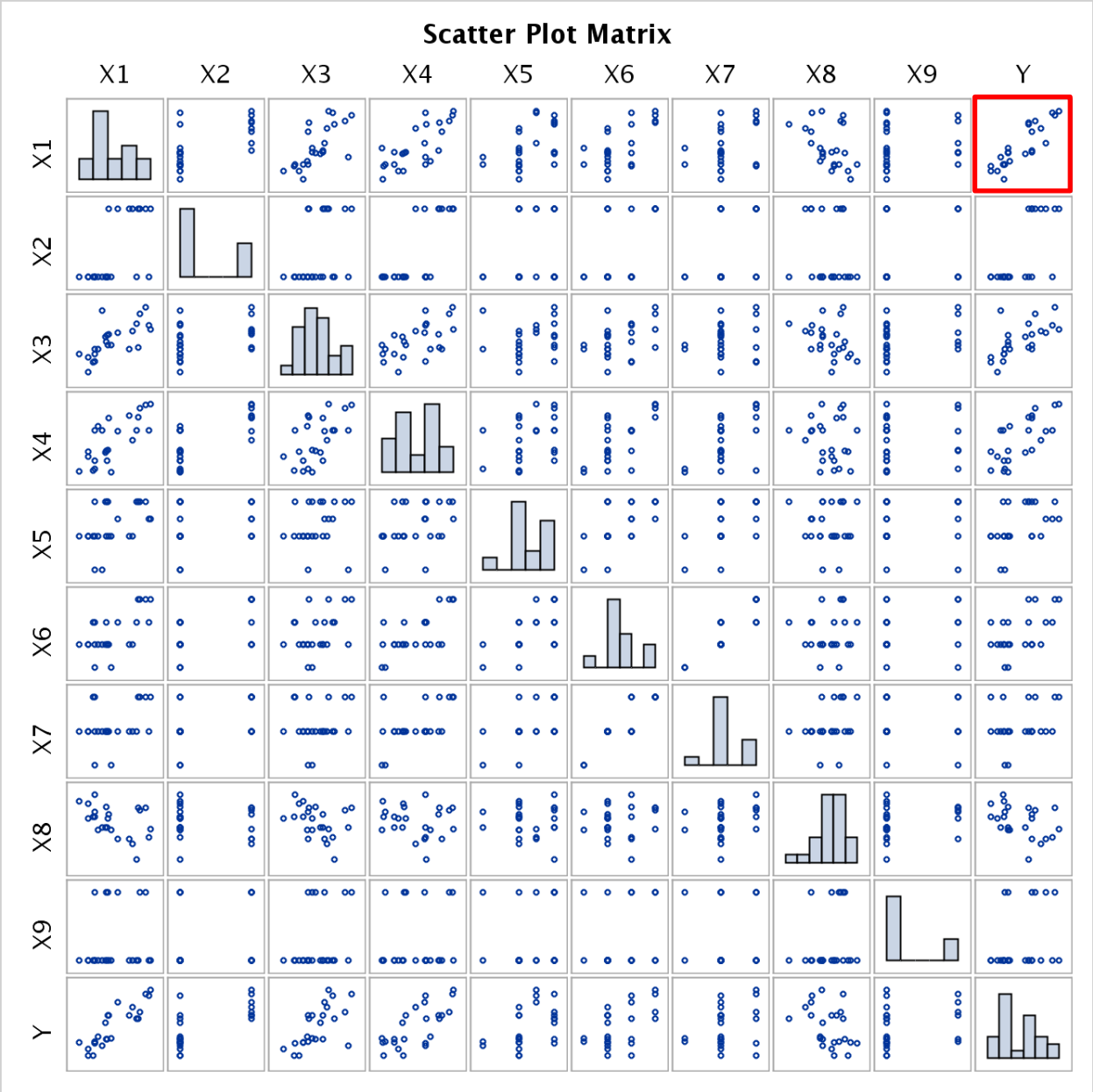

Figure 2 - Scatterplot Matrix

Figure 3 depicts the regression line and LOESS curve for each of the predictors and Y. The regression line and LOESS curve traces are virtually identical for most of the predictors except for living space (X3), rooms (X5), garage stalls (X6) and bedrooms (X7). Figure 4 shows the nonlinear relationships in more detail such as how an increase from 1.0 to 1.5 rooms has a large effect on the sales price.

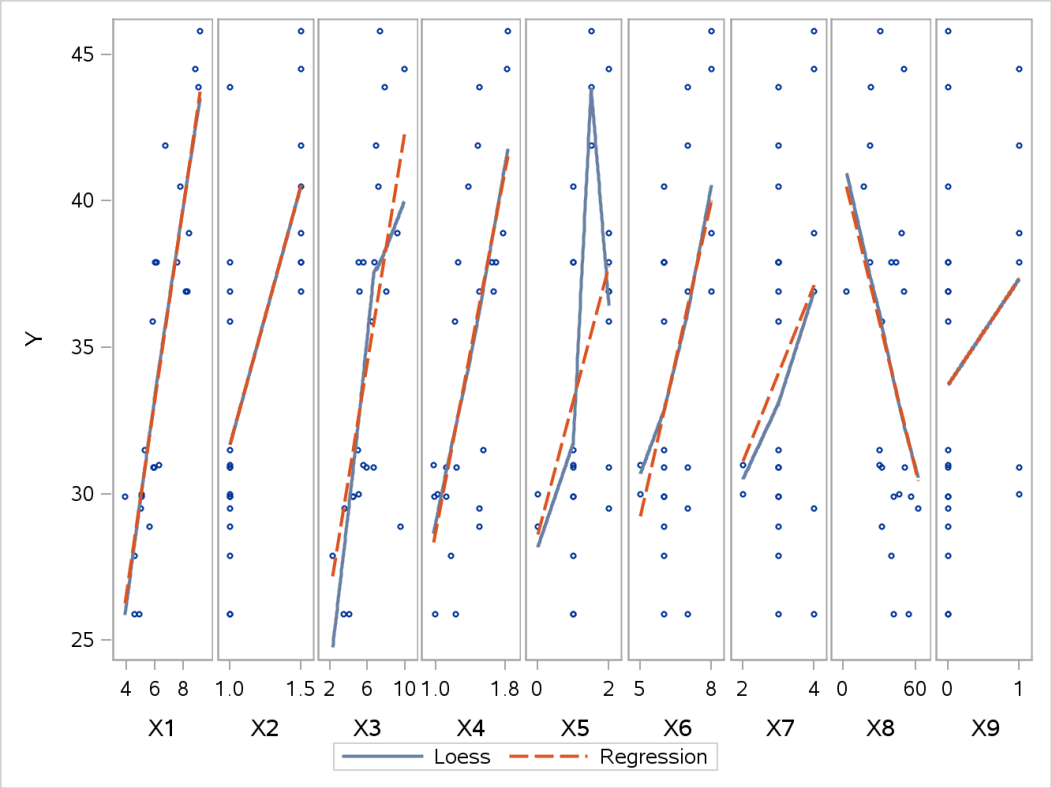

Figure 3 - LOESS and Regression Scatterplot

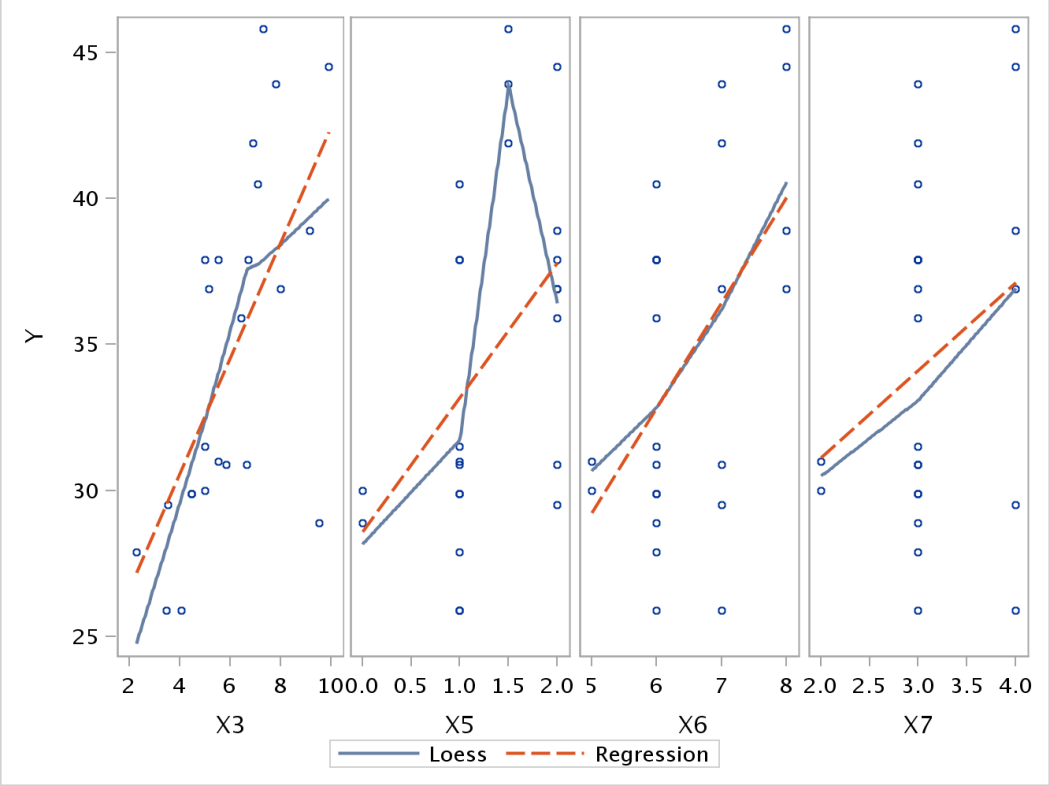

Figure 4 - LOESS and Regression Scatterplot for Non-Linear Relationships

Conclusions:

This data analysis evaluated nine predictors of sales price of a home and determined that taxes was the single best predictor. The age of the home was negatively correlated as newer homes would have higher prices that older homes.

Living space, rooms, garage stalls and bedrooms have nonlinear relationships with sales price. Taxes, bathrooms, living space, age of the home and fireplaces all have positive relationships with the sales price of the home.

Code:

```
/*      James Gray
        2013.06.29
        graymatter@u.northwestern.edu
        Assignment1_JG.sas
*/

/*      This code is for PREDICT 410 Assignment #1 - Exploratory Data Analysis for Regression. The code will
        process a predefined dataset on the SAS OnDemand server to calculate the Pearson R coefficient and a
        scatterplot matrix for 9 predictors and the response variable Y (sales prices of a house in thousands
        of $USD).
*/

********************************************************************************************;
* 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;

********************************************************************************************;
* create correlation matrix and scatterplots
********************************************************************************************;
ods graphics on;
proc corr data=mydata.building_prices plots=matrix(histogram nvar=all);
run;
ods graphics off;

********************************************************************************************;
* create LOcally Estimated Scatterplot Smoother (LOESS)
********************************************************************************************;
```

```
ods graphics on;
proc sgscatter data=mydata.building_prices; * produce scatterplot;
compare x=(x1 x2 x3 x4 x5 x6 x7 x8 x9)
                    y=Y / loess reg; * plot comparison of each predictor against Y with loess and regression line;
proc sgscatter data=mydata.building_prices; * produce scatterplot of nonlinear in more detail
compare x=(x3 x5 x6 x7)
                    y=Y / loess reg;
compare x=(x7 x8 x9)
                    y=Y / loess reg;
run; quit;
ods graphics off;

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