Assignment #2

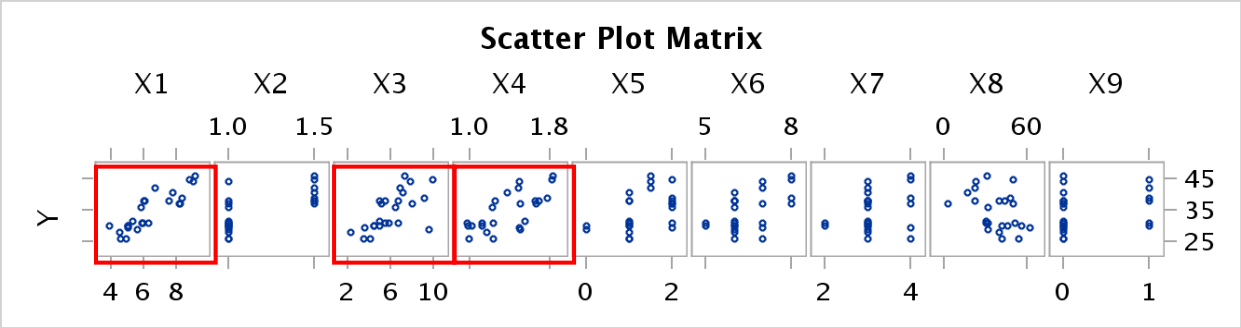
Introduction:

This exercise seeks to fit a simple regression model based on the exploratory data analysis (EDA) conducted in Assignment #1 on the **building_prices** dataset. Following that, the R-Square measure will be used to select an optimal simple regression model. Finally, the optimal model selected based on R-Square value will be evaluated for model adequacy.

Results:

Simple Linear Regression Model using EDA

The key finding from the EDA done on the building_prices dataset is that only three predictor variables have a linear relationship with Y – namely X1, X3 and X4 of which X1 has the highest Pearson's correlation coefficient (and is also statistically significant). These two data points are reproduced below for ease of reference:



Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0									
Y	X1	X2	X4	X3	X6	X5	X8	X7	X9
	0.87391	0.70978	0.70777	0.64764	0.52844	0.46147	-0.39740	0.28152	0.26688
	<.0001	0.0001	0.0001	0.0006	0.0079	0.0232	0.0545	0.1826	0.2074

Based on the above analysis, the best predictor variable for building a simple regression model is X1.

Thus, fitting a simple regression model of the form **Y-hat = beta0-hat + beta1-hat*X1** we get the following estimates for the parameters (regression coefficients):

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

The above data indicates a value of beta0-hat = 13.355 and beta1-hat = 3.321. Both estimates are statistically significant based on t-test as shown above ($p < .0001$). This yields a simple regression model of

$$\text{Y-hat} = 13.355 + 3.321*X1$$

Assessing how well the above model explains the observations, an F-test of significance (F is ratio of MSR (Mean Square due to regression)/MSE (Mean Square due to Error) is run yielding the results below:

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			

The F value is statistically significant as shown above ($p < .0001$), indicating that X1 does have predictive power in explaining variability in Y. R-square, a statistical measure of goodness-of-fit is then calculated and yields the results shown below:

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

R-Square value of .76 indicates that 76% of variation in response variable can be explained by predictor variable X1, which is fairly high.

This completes the analysis of the simple regression model indicated through EDA and identification of the best predictor variable.

Selecting the Optimal Simple Linear Regression Model using R-Square

In this step, we look at the R-Square value as generated for all models using each of the predictor variables (X1-X9) in the data set.

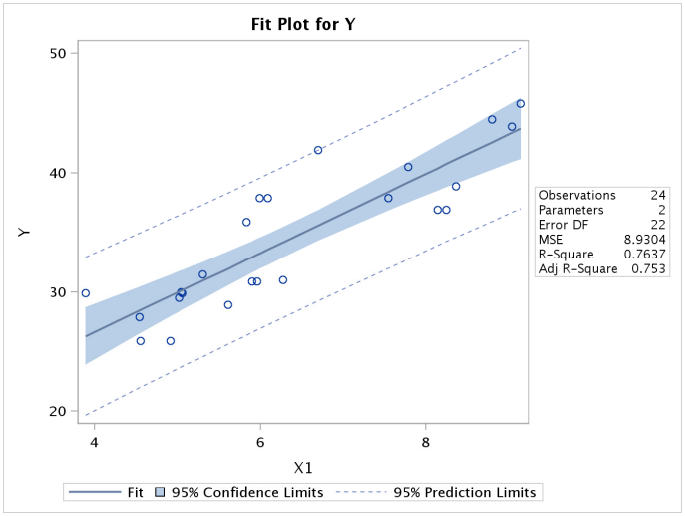
The following is a table that summarizes the R-Square value for a simple regression model developed between Y and each of the predictor variables (X1-X9):

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

Based on the results above, the model with the highest R-square value is the one between Y and X1, and this is the same predictor variable selection that the EDA led us towards. Thus the EDA pointed us towards the optimal regression model.

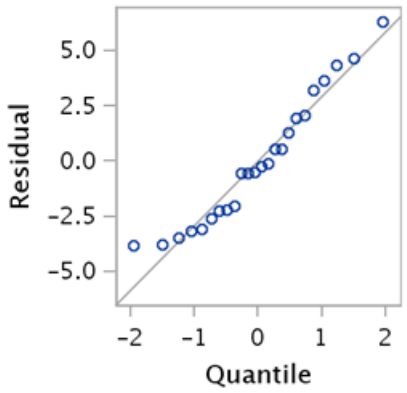
Evaluating Adequacy of the Optimal Simple Linear Regression Model chosen using R-Square

To test the adequacy of the chosen simple regression model, we walk through a few steps:
1. **Visual inspection** – plotting the fitted regression model over the scatterplot of observations. This is shown below for the optimal model (Y with X1)



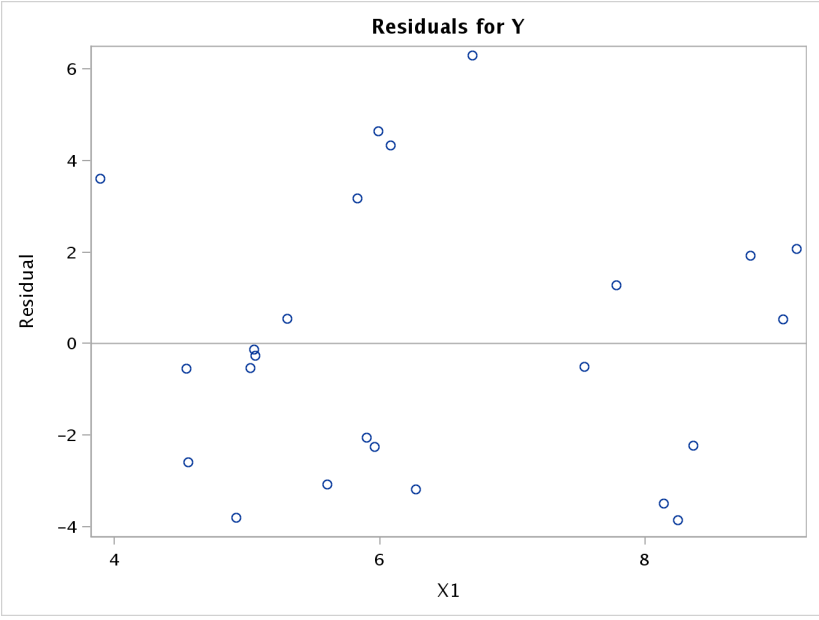
As can be visually examined, the scatterplot data is clustering around the fitted regression model, and thus provides a visual confirmation of the model's adequacy in explaining the observations.

2. **Normal distribution of residuals:** An assumption in ordinary least squares regression is that the residuals are normally distributed. For the selected model, this is best observed through a QQ plot of standardized residuals against a normal curve – both shown using quartiles. If the residuals are normally distributed, this should be a straight line.



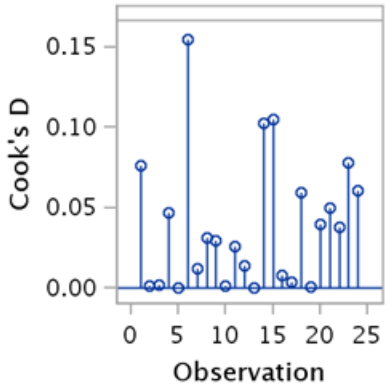
The picture shows that the residuals are on a straight line plotted against a normal curve, indicating that the residuals for the chosen simple linear regression are normally distributed.

3. **Residuals uncorrelated to predictor variables:** Another standard assumption is that the residuals are uncorrelated to the predictor variables in the model. This is best checked through a plot of standardized residuals against the predictor variable (in this case X1). This should not show any particular pattern if the residuals are uncorrelated with the predictor variable. This can be observed in the plot below:



The data above do not show any particular pattern and thus indicate that the residuals are uncorrelated to the predictor variable.

4. **Influential Observations:** In validating a model for adequacy, it is important to check for observations that are highly influential and may perhaps be skewing the model in a particular direction. A measure called Cook's distance is used to evaluate a model for any influential observations, and value > 1 in Cook's distance is frequently seen as an influential observation. The figure below shows Cook's D across the observations for this dataset.



As can be seen from the figure above, the highest value of Cook's D is around .15 and thus cannot be classified as influential. There are thus no influential observations in this dataset by this measure.

The optimal model selected by the R-Square process (which is also the same as selected by the EDA process) also shows itself as adequate per the model adequacy tests undertaken.

Conclusions:

This assignment demonstrated that the EDA process identified the same predictor variable for use in a simple linear regression as the R-Square selection process that compared the various models using their R-Square value. Further, examination of the chosen model for adequacy revealed that key assumptions such as normality of residuals and absence of any correlation between residuals and predictor variables held true. Applying visual inspection, the model showed a good fit over the set of observations. And importantly, no single observation is shown to have an outsized influence on the model. Overall, the model presents itself as a good fit with an R-Square value of .76 indicating that 76% of variability in the response variable Y can be explained by the predictor variable X1.

It is worth examining in future whether any of the other variables in the data set can help improve upon this by explaining the unexplained part of variability in Y.

Code:

```
/* Assignment 2 - 410-57
*/

/* Library at the SAS Servers
```

```
*/

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

/* Loading the data into 'temp' dataset
*/
data temp;
  set mydata.building_prices;
run;

/* Enabling plotting and other graphics
*/
ods graphics on;

/*
  Generating correlation coefficient and panel of scatterplot
  for x1-x9 predictor variables and response variable
*/

proc corr data=temp plots(only)=matrix(nvar=all);
  var x1-x9;
  with y;
run;


/*
  Simple linear regression model using only X1 as the predictor
  variable, selected via EDA
*/
proc reg data=temp;
  model y = x1;
run;

/*
  Selecting a model using R-square, the procedure below generates
  the R-square value for each simple linear regression model created
  using each of the predictor variables x1-x9, and selects the fitplot,
  diagnostics and residual plots for the model with the highest R-Square
  value
*/
proc reg data=temp plots(only) = (fitplot diagnostics residuals);
  model y = x1-x9 / selection=rsquare start=1 stop=1;
run;


ods graphics off;
quit;
```