Assignment #2

Daniel S Prusinski

Introduction:

The purpose of assignment two is to become comfortable fitting a linear regression model to data, and analyze the models appropriateness. Specifically assignment two uses the building_prices data. Assignment one focused on steps 1-4 of Regression Analysis, and this assignment works through the last four three for regression analysis (RABE p13). The steps Method of Fitting, Model Fitting, and Model criticism were all followed to precisely fit the data with the best simple linear regression model.

Ordinary Least Squares (OLS) was used to define the best simple linear regression. This is the method we are currently focusing on in 410, and it was assumed in the assignment that this method would be utilized. The building prices data set has nine different Predictor Variables (PV), and I am to discern which PV is the best based on the highest r-squared value. SAS enables users to quickly compute the formulas to find the predictor variable with the greatest r-squared value. Choosing the most accurate predictor variable is imperative to optimally utilizing the data for future predicted values. The best model I chose will be explained in the results section of the homework, and it was chosen based on the strongest coefficient of determination that had strong statistical inference.

In this assignment, the model will consist of fitting one predictor variable. The term used in statistics for computing a regression analysis that has one predictor variable is simple regression. Three parameters exist in this fitted model, and the first is the constant, the second is the coefficient, and the third is the response variable.

Visual graphics and statistical inference is used to critique the model and validate the assumptions for OLS. This assignment enlisted the techniques of four diagnostics to assess the goodness of fit for the best simple regression model. Plotting the fitted regression model over the scatterplot visually verifies that the data is linearly related, and it also visually inspects for outliers. Assessing the normality of the residuals using a Quantile-Quantile plot (QQ plot) verifies the assumption of normality of errors. Ideally, the normality errors fall along a 45 degree line on the graph (http://www.datavis.ca/courses/eda/eda3.html, 2012). Plots that deviate on the graph need further analysis because the residuals might not follow a normal distribution which would invalidate the assumption. Plotting the predictor variable against the residuals verifies the residuals are not correlated to the predictor values, which is one of the assumptions of OLS. Also, the plots should be rather random and have both negative and positive plot. OLS regression is vulnerable to distortion if extreme outliers are in the data. Cook's Distance visually and numerically demonstrates how far all plots fall from the mean, which helps in discerning outliers. The diagnostics from running SAS programs will either validate or invalidate the assumptions of OLS, and can be found in the next section of this report.

I chose to use the predictor variable X1 from assignment 1 to create a simple linear regression model. This predictor variable was chosen based on its high correlation coefficient. The high t-value and low p-value also gives X1 strong statistical significance. The regression model/equation looks like this, Y = 13.36 + 3.32X.

Pearson Correlation Coefficients, N = 24 Prob > |r| under H0: Rho=0		
Y	Y	X1
	1.00000	0.87391
		<.0001

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	13.355	2.595	5.15	<.0001
X1	1	3.321	0.393	8.43	<.0001

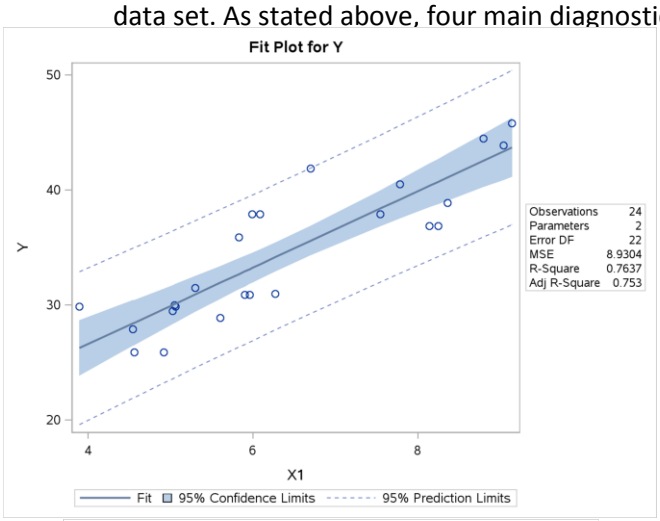
Utilizing the "*selection=rsquare"* option in *"proc reg"* in SAS, I was able to see all the r-squares for the predictor variables. The output below ranks all the variables. The difference between the R-squared and the correlation coefficient of last week's assignment was simply squaring the correlation coefficient to attain the R-squared. It turns out that I selected the optimal predictor variable for this regression model. I attribute this to the fact that I was only looking for one predictor variable, which makes it easier to discern the parameters for the regression model. (http://www.weibull.com/DOEWeb/estimating_regression_models_using_least_squares.htm)

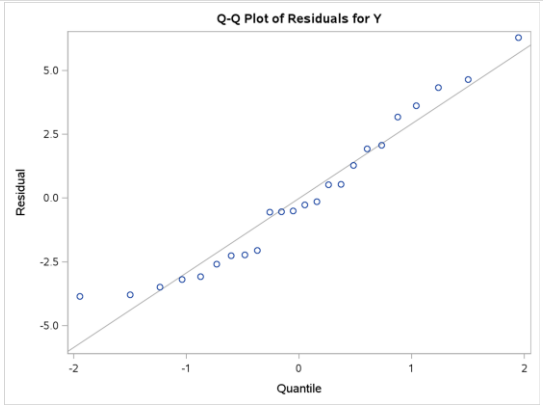
Number in Model	R-Square	Variables in Model
1	0.7637	X1
1	0.5038	X2
1	0.5009	X4
1	0.4194	X3
1	0.2793	X6
1	0.2130	X5
1	0.1579	X8
1	0.0793	X7
1	0.0712	X9

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.355	2.595	5.15	<.0001
X1	1	3.321	0.393	8.43	<.0001

Regression Model:
Y = 13.36 + 3.32X +.393E

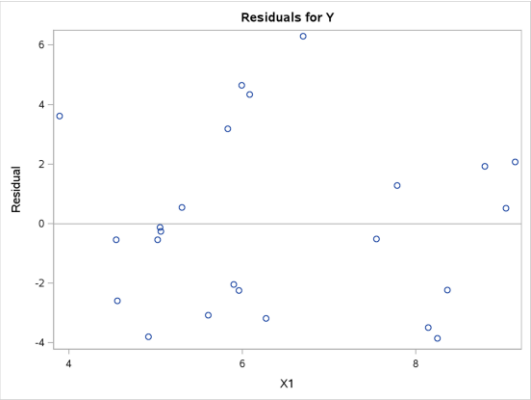
Assessing the model adequacy is the next step in confirming the best regression model for a data set. As stated above, four main diagnostics are utilized to validate the assumptions of OLS.





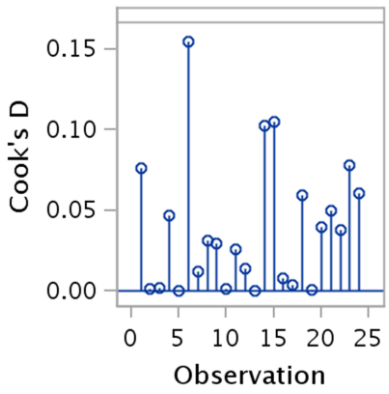
Fitting the regression model over the scatterplot is a graphical technique that validates/violates the OLS assumption of linearity. As seen in the scatterplot to the left, the data enters the graph linearly, which satisfies the linearity assumption of OLS. The solid line represents the fitted regression line plotted by OLS, and it visually lies in between the points. In the graph, the blue fill represents the 95% confidence limit. I understand this to mean if I took repeated random samples from the specific population and calculated the regression line and confidence limits for each sample, the confidence interval for 95% of my sample would include the blue fill (http://udel.edu/~mcdonald/statconf.html). The dotted blue line represents the 95% prediction limit. I understand this to mean if I predicted future occurences from this population, 95% of the observations will fall between the dotted lines (http://en.wikipedia.org/wiki/Prediction_interval) . There are no extreme outliers and the line fits the data quite well.

Looking at the second graph, the normality errors fall along a 45 degree line which proves the assumption that the errors or residuals follow a normal distribution. As the residuals fall closer to the line, the distribution for the errors follows a more normal distribution. This observation validates the Normality Assumption required for OLS. The two graphs above visually validate the Linearity and Normality Assumptions through analyzing the scatter plot of data and the residuals in the Q-Q plot.



Analyzing the residuals against the predictor variable visually allows one to inspect for OLS violations. The residuals should not be correlated with each of the predictor variables, thus the plot should be a random scatter of points with no discernible patter. Given the plot below, one can see that the points are scattered randomly which indicates linearity. Also the random homoscedasticity assumption associated with residuals having the same variance.Cook's distance is used to estimate the influence of data points and outliers. As the value for a point increases, the more inflential the point. Points that have a value greater than one are considered significant and warrant taking a closer look. After analyzing Cook's D for the X values, it can be seen that no point unduly influences the set. This analysis confrims that X2 does not have any major outliers.

Cook's plot



The visual graphics demonstrate that the simple linear regression model using X1 as the predictor variable follows the major assumptions needed for the OLS method to be valid. The diagnostics for SAS are powerful tools that need carful study in order to better understand the story that the data is telling. Through these diagnostics one can better learn about the situation reflected by the data. The diagnostics for simple linear regression prepare one for the diagnostics needed for multiple regression. As the assignments become more complicated, I feel confident in simple regression and the diagnostics that assess model adequacy.

Code:

```
*****Statement to access where the data is stored*****;
libname mydata '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/';
ods graphics on;

*****Program for running a Regression Model, Chapter 9 Cody*****;
Title "Running a Simple Regression Model";
proc reg data=mydata.building_prices;
      model y = x1;
run;

*****Using the RSQUARE Selection Method, 138*****;
Title "Demonstrating the RSQ Selection Method";
proc reg data=mydata.building_prices;
      model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 /
      selection = rsquare cp adjrsq start=1 stop=1;
run;

*****Using Proc Reg for Residual and Diagnostics*****;
title 'Fits of Regression Analysis';
proc reg data=mydata.building_prices plots (only) = (QQplot Fitplot
Diagnostics Residuals);
      model y = x1;
Run;
ODS Graphics off;
```