Pitfalls in Linear Regression Analysis

Due to the widespread availability of spreadsheet and statistical software for disposal, many of us do not really have a good understanding of how to use regression analysis properly. Indeed, before one can use regression analysis in a proper manner, one should realize the various difficulties associated in using it, namely:

1. Lacking an awareness of the assumptions of least-squares regression
2. Not knowing how to evaluate the assumptions of least-squares regression
3. Not knowing what the alternatives to least-squares regression are if a particular assumptions is violated
4. Worst still, using a regression model without knowledge of the subject matter

How can a user be expected to know what the alternatives to least-squares regression are if a particular assumption is violated, when he or she in many instances is not even aware of the assumptions of regression, let alone how the assumptions can be evaluated? Hence, it is necessary to go beyond the basic number crunching exercise such as the computation of the Y-intercept, the gradient (slope) and r^2.

Let us recall what the three major assumptions of regression and correlation are:

1. Assumption of normality:
   This assumption requires that errors around the line of regression be normally distributed at each value of X. Like the t test and the ANOVA F test, regression analysis is fairly robust against departures from the normality assumption. As long as the distribution of the errors around the line of regression at each level of X is not extremely different from a normal distribution, inferences about the line of regression and the regression coefficient will not be seriously affected.

2. Assumption of homoscedasticity:
   This requires that the variation around the line of regression be constant for all values of X. This means that the errors in the Y-responses vary by the same amount when X is a low value as when X is a high value. This homoscedasticity assumption is important for using the least-squares method of determining the regression coefficients. If there are serious departures from this assumption, either data transformations or weighted least-squares methods can be applied.

3. Assuming independence of errors:
   This requires that the errors should be independent for each value of X. This assumption is particularly important when data are collected over a period of time. In such situation, the errors for a particular time period are often correlated with those of the previous time period. If this occurs, alternatives to least-squares regression analysis need to be considered.

This discussion of pitfalls in regression can be best illustrated by referring to Table 1 on three sets of artificial data that demonstrate the importance of observations through scatter plots and residual analysis.

Table 1: Three sets of artificial data

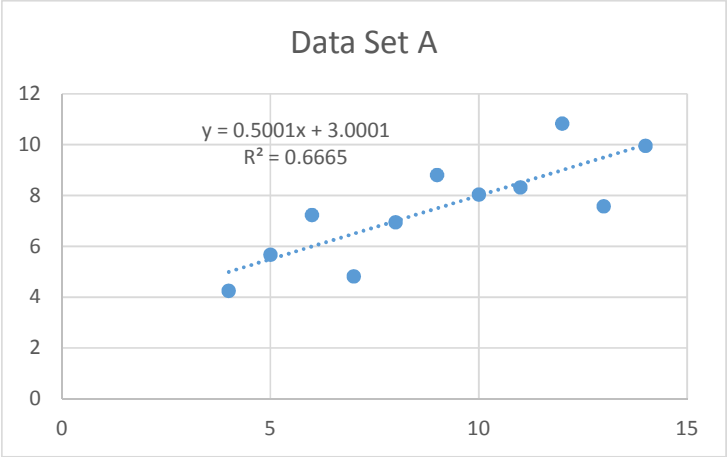
Data Set A		Data Set B		Data Set C	
X	Y	X	Y	X	Y
4	4.26	4	3.10	4	5.39
5	5.68	5	4.74	5	5.73
6	7.24	6	6.13	6	6.08
7	4.82	7	7.26	7	6.42
8	6.95	8	8.14	8	6.77
9	8.81	9	8.77	9	7.11
10	8.04	10	9.14	10	7.46
11	8.33	11	9.26	11	7.81
12	10.84	12	9.13	12	8.15
13	7.58	13	8.74	13	12.74
14	9.96	14	8.10	14	8.84

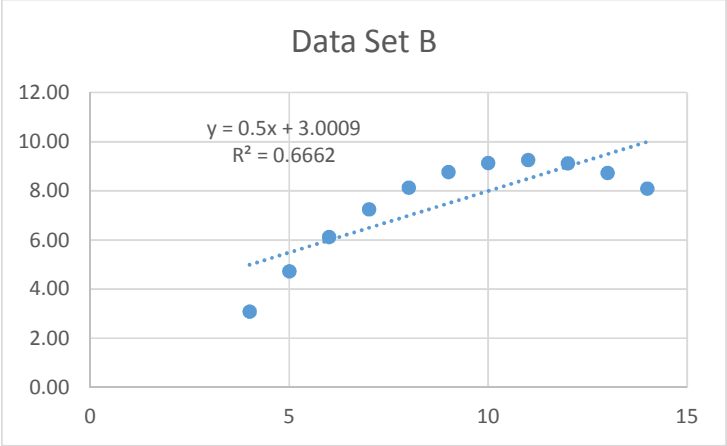
It is interesting to note that these three sets of data yield almost the same statistical parameter values:

$$Y = 0.50X + 3.00$$
$$r^2 = 0.667$$
$$SSR = 27.51$$
$$SSE = 13.76$$
$$SST = 41.27$$

Had we stopped our analysis at this point, we would lose valuable information in the data collected and might be led to wrong inferences and conclusions. However, if we were to examine the scatter diagrams of these three sets of data as shown in Figure 1 and the residual plots in Figure 2, we would see how different the data sets were.

Figure 1: Scatter Plots of Data Sets A, B and C


Data Set A
y = 0.5001x + 3.0001
R² = 0.6665
12
10
8
6
4
2
0
0
5
10
15


Data Set B
y = 0.5x + 3.0009
R² = 0.6662
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0
5
10
15

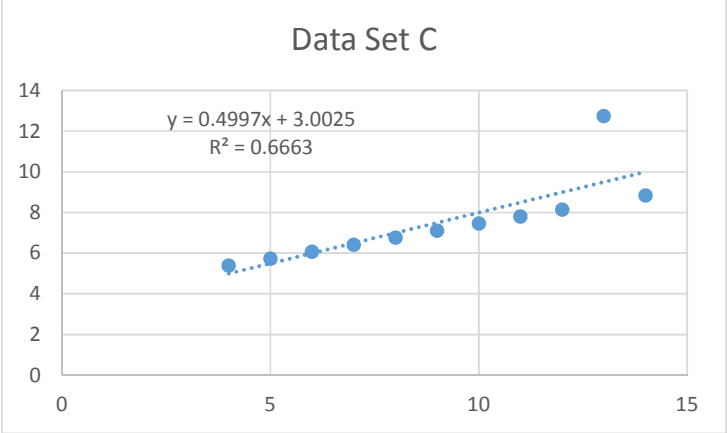
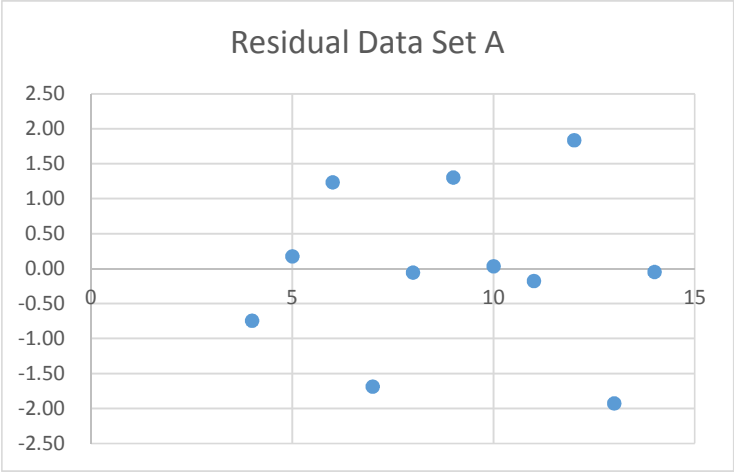
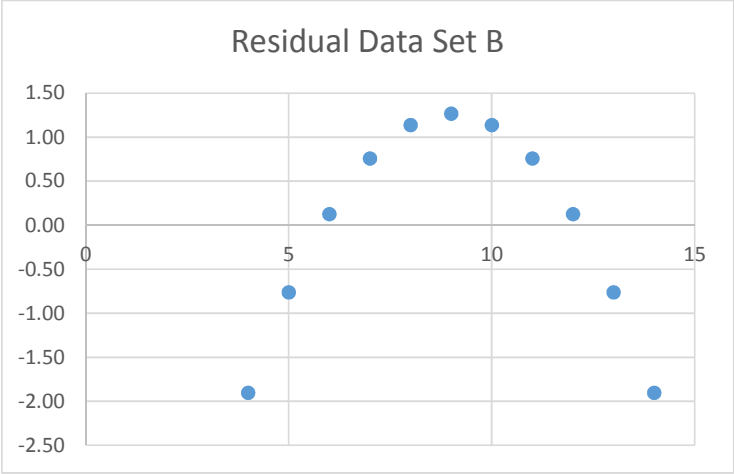
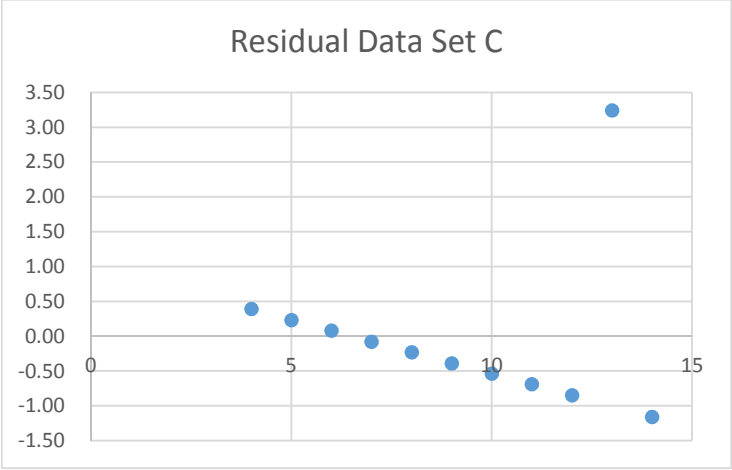

Data Set C
y = 0.4997x + 3.0025
R² = 0.6663
14
12
10
8
6
4
2
0
0
5
10
15

Figure 2: Residual plots for Data Sets A, B and C



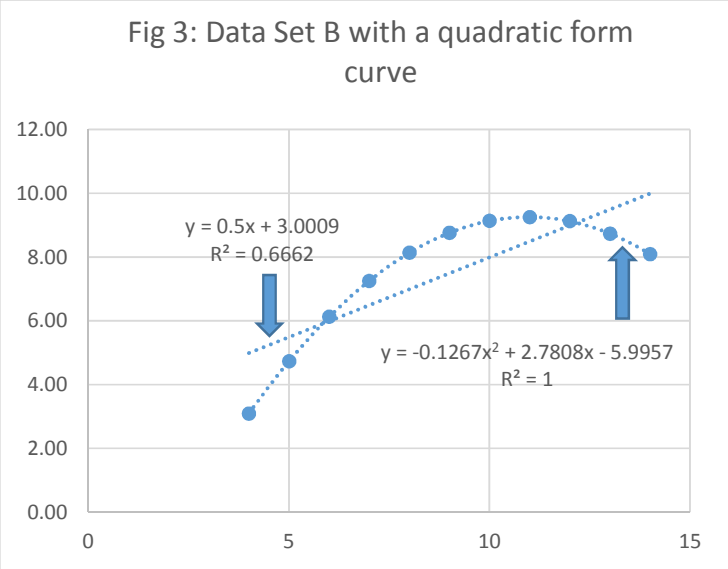




From the Figures 1 and 2, we see that the only data set that seems to follow an approximate straight line is data set A. The residual plot for data set A does show random patterns around the zero and does not appear to have

outlying residuals.

This is certainly not the case for data sets B and C. The scatter plot of data set B seems to indicate that a quadratic regression model should be more suitable. This observation is reinforced by the clear parabolic form of the residual plot for data set B. Indeed, the quadratic equation $Y = -0.1267X^2 + 2.7808X - 5.9957$ fits the points very well with $r^2 = 1$ as shown in Figure 3 below:



The scatter diagram and the residual plot for data set C clearly depict what may very well be an outlying observation. If this is the case, we may want to remove the obvious outlier and re-estimate the basic model. It should be noted that upon removal of the outlier, the result of re-estimating the model might lead to a relationship that is very much different from the one originally conjectured.

We see here that residual plots are of vital importance to a complete regression analysis. They do provide good basic information to a credible analysis. Therefore, these plots should always be included as part of a regression analysis.

We summarize a strategy for avoiding the pitfalls of regression as follows:

1. Always start with a scatter plot to observe the possible relationship between X and Y
2. Check the assumptions of regression after the regression model has been fitted, before moving on to using the results of the model
3. Plot the residuals versus the independent variable. This chart will enable

us to determine whether the model fitted to the data is an appropriate one and will allow us to check visually for violation of the homoscedasticity assumption;

4. We can use a histogram, box-and-whisker plot, or normal probability plot of the residuals to evaluate graphically whether the normality assumption has been seriously violated;
5. If the evaluation done in (3) and (4) indicates violations in the assumptions, we then use methods alternative to least-squares regression or alternative least-squares models (quadratic or multiple regression), depending on what the evaluation has indicated any of this direction;
6. If the evaluation done in (3) and (4) does not indicate violation of the assumptions, then the inferential aspects of the regression analysis can then be proceeded. Tests for the significance of the regression coefficients can be done and confidence and prediction intervals can then be developed.