SPREADSHEET REGRESSION ENHANCEMENTS: PART 11

Abstract. We discuss a fatal error in regression that is often seen in journal articles and textbooks. When using spreadsheet regression, analysts become closer to the data thereby avoiding this mistake.

1. INTRODUCTION

The medical profession is continuously aware of not letting the cure be more harmful than the malady. This awareness should also be felt when using logarithm transformations in regression. The authors highly recommend the employment of logarithm transformations for fitting equations to nonlinear relationships. However, logarithm transformations for correcting violations of the normal and homoskedastic assumptions should be viewed with caution. If a response Y variable is correctly represented by a linear regression equation, out-of-sample predictions from a nonlinear model bring biased, inaccurate, and misleading predictions. Never-the-less, this correction procedure is frequently employed without realizing the serious predictive consequences. In short, logarithm cures for normality and homoskedasticity result in disastrous out-of-sample predictions. Always evaluate the antilog statistics when assessing the worth of a model.

The objective of this note is to promote an awareness that log transformations may be harmful to predictive models. We begin by defining concepts concerning the harmful effects of logarithm transformations on regression models—when used to correct normality and homoskedasticity violations. An example is then presented to fix ideas and clarify the given concepts. We illustrate how correcting normality and homoskedasticity violations with logarithm transformations are often more harmful than doing nothing--especially when the model is employed for prediction purposes. A discussion and concluding remarks follow. A violation of normality and homoskedasticity will weaken statistical power but rarely completely invalidates the study. Employing a curvilinear regression equation when the relationship is linear will indeed nullify the integrity of the study.

2. CONCEPT

Assume the following sample regression equation possesses normality or homoskedasticity violations.

$$\hat{\mathbf{Y}} = \mathbf{Xb}, \qquad (1)$$

where $\mathbf{X}$ is an n x p data matrix of full rank and $\mathbf{b = (X'X)^{-1}X'Y}$. Assume (1) possesses a non-normal distribution: As stated above, an incorrect but routine solution to this problem is to transform the dependent Y values into ln Y values—the natural log of Y. By regressing ln Y with X, linear equation (2) is obtained:

$$\overline{\ln \hat{\mathbf{Y}}} = \mathbf{X} \, \mathbf{ln} \, \mathbf{b}, \qquad (2)$$

However, the anti-log of (2) is exponential equation (3) not linear equation (1):

$$\hat{Y}=b_0b_1^{\mathbf{X}_1} b_2^{\mathbf{X}_2} \dots b_k^{\mathbf{X}_k} \qquad (3)$$

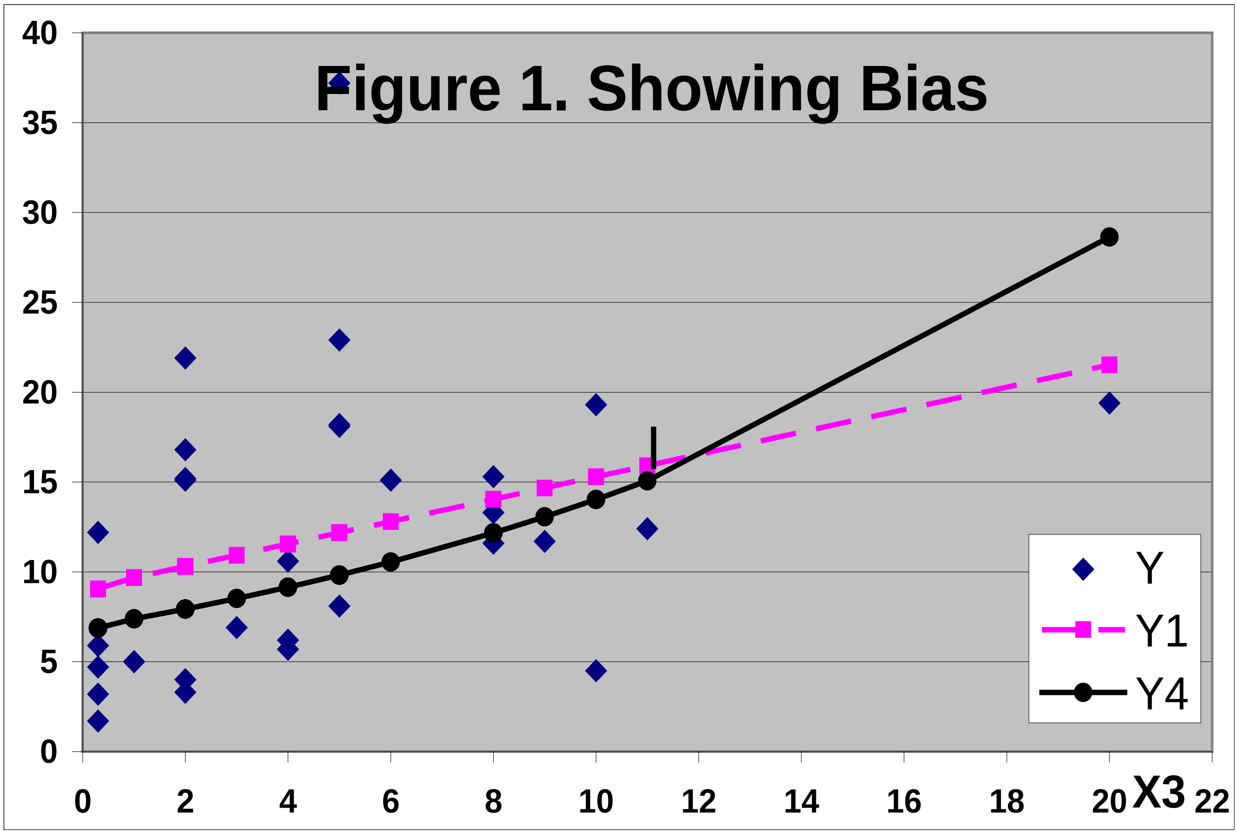
The bivariate exponential equation (4) is obtained when only one independent variable is employed:

$$\hat{Y}=b_0b_1^{\mathbf{X}_1}. \qquad (4)$$

In short, a cure for non-normality has been traded for an improperly structured regression model. The latter violation is considerably more serious. An exponential equation fitted by least squares operations may show a linear relationship for in-sample fitted values but becomes curvilinear obvious for out-of-sample predictions. Hence, this frequent mistake of using logs for correcting non-normality and/or heteroskedasticity may not be recognized until cross validation statistics are applied to the anti-logs.

3. EXAMPLE

Consider the following example taken from Lattin, Carroll, and Green (2003), p48, ANALYZING MULTIVARIATE DATA. The example concerns regressing the price of land (Y) with seven explanatory variables. The dependent variable showed signs of non-normality and logarithms are taken. Figure 1 shows in a two dimensional manner how logs will bias a linear related data set. We regress both Y and ln Y with the seven explanatory variables obtaining $\hat{Y}$or Y1 from (1) and $\overline{\ln \hat{Y}}$ or Y2 from (2) above. Taking anti-logs of (2) we obtain Y3. This is equivalent to employing (3). Again, log transformations work well when the data are related in a curvilinear fashion. Unfortunately, this data is related in a linear fashion.



Y is the actual values Y1 or $\hat{Y}_1$, is from (1) Y4 or $\hat{Y}_4$, is from (4)

Figure 1 confirms that values at the end of the data set are subject to the greatest errors. Using (3), Figure 1 reveals that the predicted value for a parcel of land from (3) is predicted to sale for $28,600 an acre; hence, the error (Y- $\hat{Y}_4$) =(19,400-28,600). Using (1), the land is predicted to sale for $21,500 an acre with a smaller error of (Y- $\hat{Y}_1$) = (19,400-21,500). Actual sales price is $19, 400. Again, the large error difference in predicted sales is from the employment of a curvilinear model when the relationship is linear.

4. DISCUSSION AND CONCLUDING REMARKS

The object of the paper is to promote an awareness that log transformations may be harmful to predictive models. This cure for non-normality and heteroskedasticity is often more harmful than the malady. We content that unless severely non-normal or heteroskedasticity, these violations do not completely invalidate the model's effectiveness. These violations will reduce the power of statistical measures but will not invalidate the complete study. The F-test is robust against these violations. Kutner, et. al (2004, page 793), " ...the lack of normality is not an important matter, provided the departure from normality is not extreme."   Scheffé (1959) also concurs.

Assuming linearity when the relationship is non-linear or the reverse will negate the study. Hence, employing log transformations when the relationship between Y and the X variables are linear will lead to inaccurate, misleading, and often disastrous results. Again, use caution when correcting heteroskedasticity and non-normality with log transformations. Do not let the cure be more harmful than the malady—do no harm.

REFERENCE

Kutner, M., C. Nachtsheim, J. Neter (2004) *Applied Linear Regression Models*, 4th ed., Chicago, Irwin

Scheffé, H., "The Analysis of Variance," New York, Wiley