March 17, 2002

NORMALITY IS A NECESSARY AND SUFFICIENT CONDITION FOR OLS TO YIELD ROBUST RESULTS

by

Arie Preminger and Haim Shalit

Ben-Gurion University of the Negev

Abstract

Yitzhaki (1996) showed that the OLS estimator is a weighted average of the slopes defined by adjacent observations. The weights depend only on the distribution of the independent variable. In this note, we show that equal weights can only be obtained if, and only if the independent variable is normally distributed. This may serve as the basis for a new test for normality.

Keywords : Ordinary least squares, Robustness, Normality

Address: Haim Shalit
Department of Economics
Ben Gurion University of the Negev
Beer-Sheva, Israel
shalit@bgumail.bgu.ac.il
Phone: +972-8-647-2299
Fax: +972-8-647-2941

NORMALITY IS A NECESSARY AND SUFFICIENT CONDITION FOR OLS TO YIELD ROBUST RESULTS

In a recent paper, Yitzhaki(1996) presented the following interesting results regarding the Ordinary Least Squares (*OLS*) estimator of a simple regression coefficient:[1]

1. The *OLS* estimator of the slope coefficient is a weighted average of the slopes delineated by adjacent observations.
2. The weights used in averaging the slopes depend solely on the distribution of the observations of the independent variable.
3. In particular, if the independent variable is normally distributed, the weights are equal to the normal density. Hence, equal percentiles of the distribution receive equal weights for all the slopes.

The major implication of Yitzhaki's results is that unless the independent variable is normally distributed, the *OLS* estimator is in fact a weighted regression estimator that attributes most of the weight to the more extreme observations. Usually, *OLS* is silent about the distribution of the independent variable that is assumed to be non-random. Indeed, the issue of the distribution of the explanatory variable is not discussed in econometric texts. At most, normality of error terms is required as a prerequisite for statistical inference.

Shalit & Yitzhaki (1998) have shown that for observations characterized by fat tails such as financial data, so called "outliers" are receiving most of the explanatory power of the regression, thus yielding non-robust results. Removal of outliers is not a desirable practice as this eliminates valid information on the behavior of variables. Indeed, in light of the recent high volatility of security prices, the extreme observations are the ones that contribute the most in explaining price behavior.

The purpose of this note is to extend Yitzhaki's results by claiming that equal weights attributed to equal percentiles can only be obtained if, and only if, the independent variable is

---

[1] The properties of the *OLS* estimator carry through to the multiple regression.

normally distributed. Hence a major prerequisite for the practitioner is to test whether or not the data is normally distributed if one is to obtain robust results using *OLS*, independently of whether or not statistical inference is undertaken.

1. The *OLS* Regression Estimator

In the following, we summarize Yitzhaki's results which claim that the *OLS* regression estimator is a weighted average of the slopes of the lines defined by adjacent observations.

Consider a simple regression model where variables are continuously random with a joint density function $f(X, Y)$, where X is the independent variable and Y is the dependent variable. Let

f_X, F_X, μ_X, and σ^2_X denote the marginal density, the marginal cumulative distribution, the expected value and the variance of X. Assume the existence of the first and second moments.

Theorem 1: *Let $E(Y|X) = \alpha + \beta X$ be the best linear predictor of Y, given X. Then β_{OLS} is the weighted average of the slopes of the regression curve $g(x) = E(Y|X=x)$, namely:*

$$\beta_{OLS} = \int_X w(x)\delta(x)dx, \tag{1}$$

where $\delta(x) = g'(x)$ and $w(x) > 0$, $\int w(x)dx = 1$ and the weights are given as:

$$\begin{aligned} w(x) &= (1/\sigma_X^2)[\mu_X F_X(x) - \int_{-\infty}^{x} t f_X(t)dt] \\ &= \int_{-\infty}^{x} (\mu_X - t) f_X(t)dt/\sigma_X^2 \quad . \end{aligned} \tag{2}$$

Proof: See Yitzhaki (1996).

Theorem 1 presents the *OLS* regression coefficient as a weighted average of the dependent variable differences, conditional on the independent variable differences. The weighting scheme depends solely upon the cumulative distribution of the independent variable. In particular, for the normal distribution, the weighting scheme becomes (Yitzhaki,1996):

$$w(x) = -1/\sqrt{2\pi\sigma^2} \int_{-\infty}^{x} t e^{-(t-\mu)^2/2\sigma^2} dt = 1/\sqrt{2\pi\sigma^2}\, e^{-(x-\mu)^2/2\sigma^2} . \qquad (3)$$

The weights are identical to the density of the normal distribution for *X*. Hence, equal percentiles of the distribution receive equal weights and the explanatory power of the regression is distributed evenly among the observations.

2. The Normal Distribution of the Independent Variable

We demonstrate that the only valid case for which equal weights are ensured is when the independent variable is normally distributed. For any other distribution of the explanatory variable, uneven weights will be obtained. Henceforth, robust *OLS* results can be secured solely for a normally distributed independent variable.

Theorem 2: *The weights equal the density distribution if, and only if, the independent variable is normally distributed.*

Proof: We propose two different proofs to show the robustness of the Theorem. One proof, which uses central moments, is provided in the Appendix. The second, which follows, consists of solving the differential equation obtained from Equation (2) once one substitutes $w(x) = f(x)$ for all x. Let us integrate by parts Equation (2) to yield:

$$\sigma^2 f(x) = (\mu - x)F(x) + \int_{-\infty}^{x} F(t)\,dt \tag{4}$$

where $dF(x) = f(x) \equiv F'(x)$. We differentiate Equation (4) with respect to x and obtain:

$$\sigma^2 F''(x) - (\mu - x)F'(x) = 0\,. \tag{5}$$

The general solution to the differential Equation (5) is given as:

$$F(x) = C \int_{-\infty}^{x} e^{-(t^2 - 2\mu t)/2\sigma^2} dt \tag{6}$$

The constant C is obtained for $F(\infty) = 1$ as:

$$C = 1/\sqrt{2\pi\sigma^2}\, e^{-\mu^2/2\sigma^2}$$

After substituting for C we obtain the solution for Equation (5) which the normal probability distribution function:

$$F(x) = 1/\sqrt{2\pi\sigma^2} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt \tag{7}$$

■

4. Conclusion and Implications

We have shown that if, and only if, the probability distribution of the independent variable is normal the weights are equal to the density of the distribution. The main implication of this result is that normality is a necessary and sufficient condition in order for each percentile of the population to receive an equal share of the weights used by the regression. Only then, will all observations contribute evenly to the *OLS* estimation and yield robust estimators. As shown by Shalit & Yitzhaki (1998), if observations are not normally distributed, other regression techniques should be used to ensure robustness.

An additional implication of the result of this paper is the derivation of a new normality test using a standard goodness-of-fit test (Madansky, 1988). The main advantage of the proposed procedure testing normality is that it does not require to specify the mean and the variance of the normal distribution for the null hypothesis. The procedure mainly consists of testing whether or not the weights used by *OLS* are equal.

Appendix

Proof: Let us first show that:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(\mu - t)f(t)dt\cdot(x - \mu)\,dx = \frac{1}{2}\int_{-\infty}^{+\infty}(x - \mu)^3 f(x)\,dx \qquad (8)$$

This is obtained by integrating Equation (8) by parts and recalling that:

$$(x^2 - 2\mu x)/2\int_{-\infty}^{x}(\mu - t)f(t)dt\Big|_{-\infty}^{+\infty} = 0 \quad .$$

We start with symmetric distributions. We know that for these distributions the odd central moments vanish. Thus:

$$\int_{-\infty}^{\infty}(x - \mu)^3 f(x)\,dx = 0 \qquad (9)$$

Dividing Equation (8) by σ_X^2 and using (9), we obtain:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{x}[(\mu - t)f(t)/\sigma^2]dt\,(x - \mu)\,dx = 0 \quad . \qquad (10)$$

Inserting the weights expressed by (2) we obtain:

$$\int_{-\infty}^{\infty}w(x)(x - \mu)\,dx = \int_{-\infty}^{\infty}[w(x) - f(x)]\,x\,dx = 0 \quad . \qquad (11)$$

Hence for some symmetric distributions $w(x) = f(x)$ for all x ≠ 0. However, for asymmetric distributions, the weights *cannot* equal the density function because Equation (8) does not vanish. Yitzhaki (1996) proved that if the distribution is normal the weights are equal to the density.

We now show that if the distribution is symmetric but not normal Equation (11) cannot hold. With the same technique used to obtain Equation (8) we demonstrate that:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(\mu - t)f(t)\,dt\cdot(x-\mu)^2\,dx = \frac{1}{3}\int_{-\infty}^{+\infty}(x-\mu)^4 f(x)\,dx \quad . \tag{12}$$

The right- hand side of Equation(8) is related to the kurtosis of a distribution which is defined as:

$$\gamma_4 = \int_{-\infty}^{\infty}(x-\mu_X)^4 f_X(x)dx/\sigma_X^4 - 3 \tag{13}$$

For all non-normal distributions, the kurtosis is non zero. In Equation (12) we replace the right-hand-side integral with γ_4 and divide it by σ_X^2 to obtain:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{x}[(\mu - t)f(t)/\sigma^2]\,dt\cdot(x-\mu)^2\,dx - \sigma^2 = \sigma^2\gamma_4/3 \tag{14}$$

We now insert the weights defined by Equation (2) to obtain:

$$\int_{-\infty}^{+\infty}[w(x) - f(x)](x-\mu)^2\,dx = \sigma^2\gamma_4/3 \tag{15}$$

For symmetric distributions with a non-null kurtosis, Equation (15) cannot be zero. Hence for these distributions, the weights cannot equal the density function. ■

References

Madansky, Albert, (1988), *Prescriptions for Working Statisticians*, Springer-Verlag, New York.

Shalit, Haim and Shlomo Yitzhaki, (2002), "Estimating Beta", *Review of Quantitative Finance and Accounting*, 18, 95-118.

Yitzhaki, Shlomo, (1996), "On Using Linear Regressions in Welfare Economics", *Journal of Business and Economic Statistics,* 14, 478-486