TECHNOLOGICAL TOOLS FOR CONTRASTING THE ORDINARY AND TOTAL LEAST SQUARES METHODS

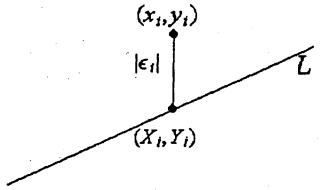
Stephen M. Scariano and Maria E. Calzada
Loyola University New Orleans
New Orleans, LA 70118
scariano@loyno.edu

We offer a Visual Basic program that can be used to compare the Ordinary Least Squares (OLS) and Total Least Squares (TLS) regression lines. It can be obtained at **www.loyno.edu/~scariano**. Here, we introduce the OLS and TLS methods and present a practical application suitable for classroom presentation.

Ordinary Least Squares

Given known data points $\{(x_i, y_i), i = 1,2,\ldots, n\}$, the key idea behind the OLS method to find the "best-fitting" line, $y = mx + b$, passing through the points. This criterion is commonly used to "fit" a line to data, and Gauss was its pioneer. Among a host of references on this topic, Bjorck [1] gives a complete and exhaustive treatment of least squares problems.

Figure 1: Ordinary Least Squares: Minimize the Sum of Squares of Vertical Displacements from line L.



Letting $\hat{y}_i$ denote the linearly "fitted" or "predicted" ordinate corresponding to abscissa x_i, Gauss reasoned that the "best-fitting" line is obtained by minimizing the sum of squared distances between the y-values actually observed and those predicted: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Since the predicted values are assumed to follow some linear trend, Gauss's rationale requires that we minimize

$$\hat{D}(m,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i-1}^{n}(y_i - mx_i - b)^2 \qquad (1)$$

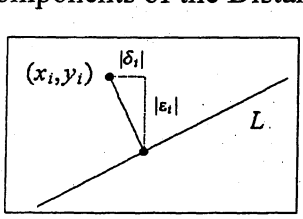
as a function of the variables m and b, denoting the critical values as $\hat{m}$ and $\hat{b}$ when they exist.

Define $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, $s_x^{\,2} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $s_y^{\,2} = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, and $r_{xy} = (ns_x s_y)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. It is a standard problem in bivariate calculus to find the constants $\hat{m}$ and $\hat{b}$ that minimize $\hat{D}(m,b)$. These constants are given by $\hat{m} = r_{xy}\left(\frac{s_y}{s_x}\right)$ and $\hat{b} = \bar{y} - \hat{m}\bar{x} = \bar{y} - r_{xy}\left(\frac{s_y}{s_x}\right)\bar{x}$. Therefore, the OLS "best-fitting" line for a given set of data points $\{(x_i, y_i), i = 1,2,\ldots,n\}$ is $\hat{y} = \hat{m}x + \hat{b}$.

Total Least Squares

Let us now consider the same basic problem from a slightly different perspective. Suppose that a linear relationship $Y = mX + b$ actually exists between the abscissa and ordinate values, yet neither variable can be measured precisely due to instrumentation, human, sampling, and other random errors beyond our control. In an error-free world, we would have $Y_i = mX_i + b$; however, assume that X_i and Y_i are not observed due to the persistent presence of errors in both variables. Instead, $x_i = X_i + \delta_i$ and $y_i = Y_i + \varepsilon_i$ are actually observed, where δ_i and ε_i denote the accumulated aggregate random errors associated with each variable, respectively. Figure 2 depicts this scenario for a single generic data point (x_i, y_i) and suggests the use of the Pythagorean Theorem.

Figure 2: Orthogonal Components of the Distance from (x_i, y_i) to L.



Reasoning as in the OLS case, we should choose m and b so as to minimize $\tilde{D}(m,b) = \sum_{i=1}^{n}(hypotenuse_i)^2 = \sum_{i=1}^{n}\left(\sqrt{\delta_i^2 + \varepsilon_i^2}\right)^2 = \sum_{i=1}^{n}\frac{(y_i - (mx_i + b))^2}{(1+m^2)}$. In that instance, the "best-fitting" line in the "total" sense is the one which minimizes the sum of squared orthogonal distances, rather than just the sum of squared vertical displacements. Of course, the objective here is to minimize $\tilde{D}(m,b)$ with respect to m and b, so we are to find, say, $\tilde{m}$ and $\tilde{b}$ such that $\tilde{D}(\tilde{m},\tilde{b}) \le \tilde{D}(m,b)$ for all $(m,b) \in R^2$. Van Huffel and Vandewalle [2] discuss the concept of total (or orthogonal) least squares extensively. Given the reasoning that lead to equation (1), orthogonal least squares is alternately referred to as the errors in variables model.

We choose to minimize $\tilde{D}(m,b)$ with respect to m and b using the bivariate Test for Extrema. The unique solutions minimizing $\tilde{D}(m,b)$ are

$\tilde{m} = \dfrac{(s_y^2 - s_x^2) + \sqrt{(s_y^2 - s_x^2)^2 + 4s_x^2 s_y^2 r_{xy}^2}}{2s_x s_y r_{xy}}$ and $\tilde{b} = \bar{y} - \tilde{m}\bar{x}$. Moreover, explicit expressions for the predicted values $(\tilde{x}_i, \tilde{y}_i)$, the closest point to (x_i, y_i) on the TLS line, can be obtained. These predicted values are given by $\tilde{x}_i = \left[\dfrac{\tilde{m}(y_i - \bar{y}) + (x_i - \bar{x})}{\tilde{m}^2 + 1}\right] + \bar{x}$ and $\tilde{y}_i = \tilde{m}(\tilde{x}_i - \bar{x}) + \bar{y}$.

The propositions that follow next show interesting relationships between the OLS and TLS methods. Proposition 1 demonstrates that the minimal sum of squares under the OLS criterion is always at least as large as the minimal sum of squares under the TLS criterion. This does not, however, mean that the TLS method is "better" than the OLS method: the defining criteria are simply different. Proposition 2 shows that in practice the TLS line is always steeper than the OLS line, while Proposition 3 gives upper and lower bounds for the TLS slope estimate $\tilde{m}$ in terms of the OLS slope estimate $\hat{m}$. Propositions 4 and 5 are more technical results.

***Proposition* 1**: $\hat{D}(\hat{m}, \hat{b}) \geq \tilde{D}(\tilde{m}, \tilde{b})$ with strict inequality when $\hat{m} \neq 0$.

***Proposition* 2**: $|\tilde{m}| \geq |\hat{m}|$. If $r_{xy}^2 = 1$, then $\tilde{m} = \hat{m}$.

***Proposition* 3**: If $\hat{m} \neq 0$ and $r_{xy}^2 \neq 1$, $\left(\hat{m}^{-1} + \sqrt{r_{xy}^{-2} - 1}\right)^{-1} \leq \tilde{m} \leq \left(\hat{m}^{-1} - \sqrt{r_{xy}^{-2} - 1}\right)^{-1}$.

***Proposition* 4**: Under the TLS criterion, "inverse" regression (of "y on x") and regression of "x on y" are identical whenever r_{xy} is well defined and nonzero.

***Proposition* 5**: If r_{xy} is well-defined and nonzero, then

$$\tilde{m} = \frac{1}{2}(\hat{m}_x^{-1} - \hat{m}_y^{-1}) + \sqrt{\left\{\frac{1}{2}(\hat{m}_x^{-1} - \hat{m}_y^{-1})\right\}^2 + 1} \quad \text{,when } \hat{m}_y > 0$$

$$\tilde{m} = \frac{1}{2}(\hat{m}_x^{-1} - \hat{m}_y^{-1}) - \sqrt{\left\{\frac{1}{2}(\hat{m}_x^{-1} - \hat{m}_y^{-1})\right\}^2 + 1} \quad \text{,when } \hat{m}_y < 0,$$

where $\hat{m}_x$ and $\hat{m}_y$ are the "x on y" and "y on x" OLS regression slope estimators, respectively.

SAT Data

As reported by the College Board [3], the data in Table 1 represent Mean SAT/SAT I mathematics and verbal scores by gender for College-Bound Seniors over the time period 1972-2000. Although this data set can be analyzed from a variety of perspectives, let us concentrate on the verbal scores for males and females. Clearly, there is no *a priori* justification for labeling either of these variables as "explanatory" and the other as "response", and there may well be random error in measuring both of these variables. Nonetheless, the scatterplot in Figure 3, with the outliers (1972-74) removed, shows positive association between the male (abscissa) and female (ordinate) average verbal scores, and the point cloud is roughly linear with a positive slope. The regression equations are $OLS: \hat{y} = 0.6426x + 173.9226$ and $TLS: \tilde{y} = 1.0605x - 38.5827$, with

$r_{xy} = 0.6196$. The OLS and TLS residual plots for these models are shown in Figure 4, which shows no nonrandom patterns. Table 2 predicts female scores from male scores.

Table 1: Mean SAT/SAT I Scores for College-Bound Seniors (1972-2000).

	Male Verbal	Female Verbal	Male Math	Female Math		Male Verbal	Female Verbal	Male Math	Female Math
1972	531	529	527	489	**1987**	512	502	523	481
1973	523	521	525	489	**1988**	512	499	521	483
1974	524	520	524	488	**1989**	510	498	523	482
1975	515	509	518	479	**1990**	505	496	521	483
1976	511	508	520	475	**1991**	503	495	520	482
1977	509	505	520	474	**1992**	504	496	521	484
1978	511	503	517	474	**1993**	504	497	524	484
1979	509	501	516	473	**1994**	501	497	523	487
1980	506	498	515	473	**1995**	505	502	525	490
1981	508	496	516	473	**1996**	507	503	527	492
1982	509	499	516	473	**1997**	507	503	530	494
1983	508	498	516	474	**1998**	509	502	531	496
1984	511	498	518	478	**1999**	509	502	531	495
1985	514	503	522	480	**2000**	507	504	533	498
1986	515	504	523	479					

Figure 3: Scatterplot of SAT/SAT I average verbal scores (outliers (1972-74) removed).

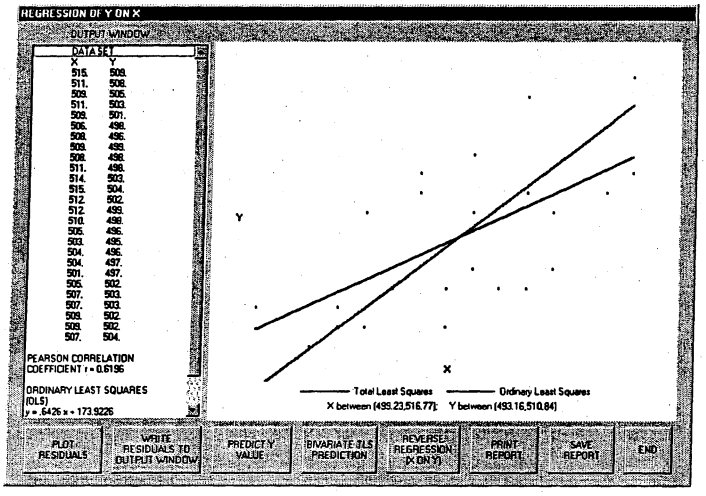
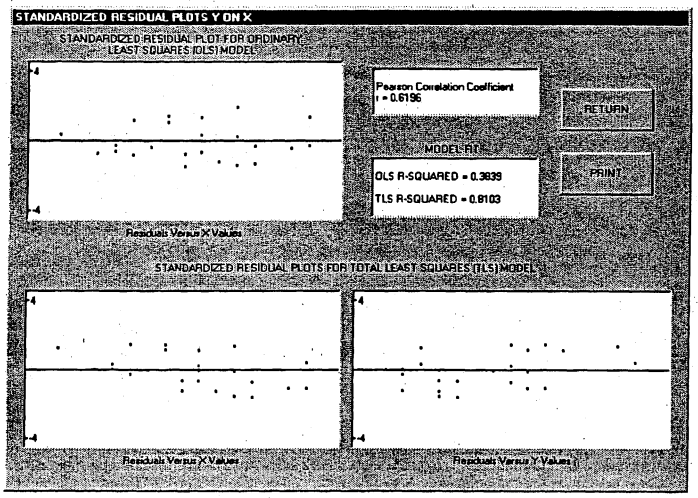


Table 2: Predicted Female average verbal score from Male average verbal score.

Male Average Score (X)	**500**	**505**	**510**	**515**
OLS Predicted Female Average Score (Y)	495.2	498.4	501.7	504.9
TLS Predicted Female Average Score (Y)	491.7	497.0	502.3	507.6

Figure 4: Residual plots of SAT/SAT I average verbal scores



Recommendations:

The SAT data set illustrates that there is no clear-cut answer as to whether the OLS or TLS technique should be preferred in a given application. The method chosen by the data analyst must be based on a keen understanding of the context of the data as well as the nature and sources of the errors present in the observations. If, in a given regression setting, the roles of "explanatory variable" and "response variable" are meaningful and the explanatory variable can be measured relatively precisely, OLS is the appropriate method to use. But, when both variables under study are subject to measurement, random, and other types of errors, the TLS method should be preferred. As with all statistical techniques, deciding which of these methods to use in practice must be done sensibly, not in a vacuum.

Although the total least squares (or errors-in-variables) model has been widely studied in the literature, far too little has been done at the elementary level to interest students in this technique. This is quite regrettable for three reasons: (i) the elementary TLS results are quite accessible to students with a calculus background, (ii) the derivation of the TLS parameter estimates is an uncontrived use of the bivariate Test for Extrema, and, most importantly, (iii) as mathematics educators, we miss the opportunity to strengthen student understanding of the least squares concept via the exercise of comparing and contrasting the OLS and TLS techniques. The visual basic program we offer here will facilitate the analyses, enhance the graphical presentation, and ease the computations.

[1] Bjorck, A. (1996), *Numerical Methods for Least Squares Problems*, Philadelphia: Society for Industrial and Applied Mathematics.

[2] Van Huffel, S. and Vandewalle, J. (1991), The Total Least Squares Problem: Computational Aspects and Analysis, Frontiers in Applied Mathematics series, 9, Philadelphia, PA:Society for Industrial and Applied Mathematics.

[3] The College Board, http://www.collegeboard.org/sat/cbsenior/yr2000/nat/72-00.html, (accessed December 2001).