Regression through the Origin

KEYWORDS:
Teaching;
Regression;
Analysis of variance;
Statistical software packages.

Joseph G. Eisenhauer
Canisius College, Buffalo, USA.
e-mail: eisenhauer@canisius.edu

Summary
This article describes situations in which regression through the origin is appropriate, derives the normal equation for such a regression and explains the controversy regarding its evaluative statistics. Differences between three popular software packages that allow regression through the origin are illustrated using examples from previous issues of *Teaching Statistics*.

◆ INTRODUCTION ◆

Although ordinary least-squares (OLS) regression is one of the most familiar statistical tools, far less has been written – especially in the pedagogical literature – on regression through the origin (RTO). Indeed, the subject is surprisingly controversial. The present note highlights situations in which RTO is appropriate, discusses the implementation and evaluation of such models and compares RTO functions among three popular statistical packages. Some examples gleaned from past *Teaching Statistics* articles are used as illustrations. For expository convenience, OLS and RTO refer here to linear regressions obtained by least-squares methods with and without a constant term, respectively.

◆ MODEL SELECTION: ◆ WHEN IS RTO APPROPRIATE?

Textbooks rarely discuss RTO other than to caution against dropping the constant term from a regression, on the grounds that imposing any such restriction can only diminish the model's fit to the data. There are, however, circumstances in which RTO is appropriate or even necessary.

First, RTO may be unavoidable if transformations of the OLS model are needed to correct violations of the Gauss–Markov assumptions. Consider, for example, the simple linear regression of Y on x

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (1)$$

where β_0 is the intercept, β_1 is the slope and e_i denotes the ith residual. Lagging observations and taking first differences (i.e. subtracting each observation from its successor) to correct for serial correlation in the errors requires transforming equation (1) into an RTO equation of the form

$$Y_i - Y_{i-1} = \beta_1(x_i - x_{i-1}) + (e_i - e_{i-1})$$

Alternatively, applying weighted least squares to correct for heteroscedasticity will result in a model with no intercept if the weighting factor (z) is not an independent variable. In that case, β_0 becomes a coefficient and equation (1) is replaced by a multiple linear regression without a constant:

$$Y_i/z_i = \beta_0(1/z_i) + \beta_1(x_i/z_i) + (e_i/z_i)$$

Even without such transformations, however, there are often strong a priori reasons for believing that $Y = 0$ when $x = 0$, and therefore omitting the constant. Indeed, Theil (1971, p. 176) contends 'From an economic point of view, a constant term usually has little or no explanatory virtues'. While that may be a slight exaggeration – it is easy to find examples in which an intercept does matter – there are certainly cases in which economic theory posits the absence of a constant. The widely used Cobb–Douglas production function, for example, relates output (Y) to capital (K) and labour (L) according to $Y = K^{\beta_1}L^{\beta_2}$, and taking logarithms yields $\ln Y = \beta_1 \ln K + \beta_2 \ln L$; imposing a constant

on this model would imply an unrealistic ability to manufacture goods without resources. An agricultural example is provided by Chambers and Dunstan (1986), who regress sugar cane harvests on farmland acreage; clearly, if no land is cultivated, there will be no crop. Casella (1983, p. 150) suggests an engineering example in which gasoline usage is a simple linear function of vehicular weight; he reasons that, in principle, a weightless vehicle would consume no fuel, so 'considering the physical constraints ... it seems most appropriate to fit a line through the origin'. And Adelman and Watkins (1994) apply RTO to the valuation of mineral deposits. Of course, similar instances can be found in almost any discipline; some ornithological and nutritional examples are discussed below.

Even when theory proscribes a constant, however, careful consideration of the observed range of data is needed. As Hocking (1996, p. 177) points out, 'if the data are far from the origin, we have no evidence that the linearity applies over this expanded range. For example, the response may increase exponentially near the origin and then stabilize into a near linear response in the region of typical inputs.' Alternatively, observations at the origin may represent a discontinuity from an otherwise linear function with a positive or negative intercept. Under those circumstances, knowing that $Y = 0$ when $x = 0$ is insufficient justification for RTO.

If there is uncertainty regarding the appropriateness of including an intercept, several diagnostic devices can provide guidance. Most obviously, one can run the OLS regression and test the null hypothesis $H_0 : \beta_0 = 0$ using the Student's t statistic to determine whether the intercept is significant. Alternatively, Hahn (1977) suggests running the regression with and without an intercept, and comparing the standard errors to decide whether OLS or RTO provides a superior fit. And Casella (1983) suggests artificially creating an extra observation – a leverage point – that pulls the OLS regression line naturally through the origin. Unless the data set is small and the observations cluster near the origin, any such leverage point is likely to be an outlier but, if it appears to be a plausible extrapolation of the actual data, one may conclude that RTO is an acceptable model. Unfortunately, there are infinitely many such leverage points that could be chosen for that exercise, and the reasonableness of RTO will depend on which point is used.

◆ IMPLEMENTATION AND ◆ EVALUATION OF RTO

In one respect, RTO is merely a special case of OLS, and the absence of the constant is actually a simplification. Indeed, minimizing the sum of squared errors for the simple linear RTO model

$$Y_i = \beta x_i + e_i$$

involves far less calculation than it does for the OLS model of equation (1). The problem

$$\underset{\beta}{\text{Min}} \sum (Y_i - \beta x_i)^2 = \sum Y_i^2 - 2\beta \sum x_i Y_i + \beta^2 \sum x_i^2$$

has only one normal equation or first-order condition

$$-2\sum x_i Y_i + 2\hat{\beta} \sum x_i^2 = 0$$

and the easily derived second-order condition, $2\sum x_i^2 > 0$, clearly guarantees a minimum. From the normal equation, the estimated slope of the regression line is

$$\hat{\beta} = \sum x_i Y_i \Big/ \sum x_i^2$$

as noted by, for example, Pettit and Peers (1991). (For weighted versions, see Turner, 1960.)

Unfortunately, the RTO residuals will usually have a nonzero mean, because forcing the regression line through the origin is generally inconsistent with the best fit. The proper method for evaluating RTO has long been disputed (see, for example, Marquardt and Snee 1974; Maddala 1977; Gordon 1981). To appreciate the controversy, note the familiar identity

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \qquad (2)$$

where $\bar{Y}$ denotes the mean of the dependent variable and $\hat{Y}_i$ is the ith fitted value. Squaring both sides and summing across all observations gives

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

but, as is well known, the cross-product term is equal to zero in the case of OLS. The remaining terms therefore constitute the usual analysis of variance decomposition

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \qquad (3)$$

where the left-hand side is the sum of squares total (SST), the first term on the right is the sum of squares due to error (SSE) and the final term is the sum of squares due to regression (SSR). The coefficient of determination for OLS is then defined by the ratio of SSR to SST

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

or equivalently

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \qquad (4)$$

Some authors maintain that because this diagnostic measure is based on an identity, it should not depend on the inclusion or exclusion of a constant term in the regression. From that perspective, equation (4) is equally valid for RTO and OLS.

However, when there is no constant in the regression, $\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$ will generally take a nonzero value, so equation (3) is not a valid basis for analysis of variance in RTO. And if the RTO model provides a sufficiently poor fit, the data may exhibit more variation around the regression line than around $\bar{Y}$, in which case $\sum(Y_i - \hat{Y}_i)^2 > \sum(Y_i - \bar{Y})^2$. Heedlessly applying equation (4) would then result in an implausibly *negative* (and thus uninterpretable) coefficient of determination as well as a negative F ratio. Moreover, it is often argued that defining SST as the sum of squared deviations from the mean is inappropriate when the regression line is forced through the origin but does not necessarily pass through $(\bar{x}, \bar{Y})$; when so viewed, equation (2) is replaced by the identity

$$(Y_i - 0) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - 0) \qquad (2')$$

Squaring and summing yields

$$\sum Y_i^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum\hat{Y}_i^2 + 2\sum\hat{Y}_i(Y_i - \hat{Y}_i)$$

but the final (cross-product) term in this equation equals zero under RTO, because

$$\begin{aligned}\sum\hat{Y}_i(Y_i - \hat{Y}_i) &= \sum\hat{\beta}x_i(Y_i - \hat{\beta}x_i) = \hat{\beta}[\sum x_iY_i - \hat{\beta}\sum x_i^2] \\ &= \hat{\beta}[\sum x_iY_i - (\sum x_iY_i/\sum x_i^2)\sum x_i^2] = 0\end{aligned}$$

Thus, equation (3) is replaced by

$$\sum Y_i^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum\hat{Y}_i^2 \qquad (3')$$

Applying equation (3′) rather than equation (3) to RTO, one finds that SSE is unchanged, but SST = $\sum Y_i^2$ and SSR = $\sum\hat{Y}_i^2$. Redefining SST and SSR in this manner results in

$$R^2 = \frac{\sum\hat{Y}_i^2}{\sum Y_i^2} \qquad (4')$$

a strictly non-negative coefficient of determination that equals or exceeds the measure in equation (4). Of course, these definitions also affect the adjusted R^2 and F statistics, but do not alter the standard error of the regression (S_e). Note that, without a constant, the degrees of freedom for SST, SSR and SSE are n, k and $n - k$, respectively, where n is the sample size and k is the number of independent variables; thus, $S_e = \sqrt{SSE/(n-k)}$ regardless of how SST is defined.

The controversy over SST is not merely academic: practitioners (and students) running RTO will obtain various outputs depending on which computer packages they use. Indeed, as Prvan et al. (2002, p. 74) observed in a recent comparison of Minitab, SPSS and Excel, 'Obtaining a simple linear regression is easy in all three packages, and all three give the standard output options (such as regression through the origin)'. But in fact the three packages all give *different* outputs for RTO! Two illustrative examples are provided below.

◆ EXAMPLES ◆

In Kimber's essay on the shape of birds' eggs, egg height is regressed on width, both with and without an intercept (Kimber 1995). Her study of 281 species can be approximately replicated using the 96 observations provided in the Data Bank section of the Summer 1990 issue of *Teaching Statistics* (Data Bank 1990). Regardless of which computer package is used, OLS yields the following output:

Height = –1.774 + 1.444 Width
[0.001] [0.000]
$S_e = 2.123$ $F = 5720.076$ $R^2 = 0.984$ $\bar{R}^2 = 0.984$

where two-tailed p-values are shown in brackets below the estimates. Notice that the intercept is statistically significant; although it is, of course, impossible for an egg to have a zero width, the intercept may nevertheless be important, as it represents the extrapolation of the regression line back to the vertical axis. The effect of removing the intercept can be seen by running RTO. If Excel is used, as it was by Kimber, RTO yields

$$\begin{array}{c}\text{Height} = 1.382\ \text{Width} \\ \qquad [0.000] \\ S_e = 2.251 \quad F = 5073.616 \quad R^2 = 0.9816 \quad \bar{R}^2 = 0.9711\end{array}$$

which indicates a poorer fit by all diagnostic measures: the standard error, F and R^2 (adjusted and unadjusted). However, the SPSS linear regression procedure without an intercept yields

$$\begin{array}{c}\textit{Height} = 1.382\ \textit{Width} \\ \qquad [0.000] \\ S_e = 2.251 \quad F = 24{,}283.995 \quad R^2 = 0.996 \quad \bar{R}^2 = 0.996\end{array}$$

Notice that the regression equation and standard error are the same in the two programs, but the F and R^2 statistics are different. Indeed, in SPSS these statistics seem to indicate a better fit without the intercept than with it. The discrepancy between software packages arises because Excel is based on equations (3) and (4), while the RTO function in SPSS uses equations (3′) and (4′). The SPSS output, however, is accompanied by the disclaimer 'For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept' [emphasis in original].

To make matters more confusing still, SPSS offers a nonlinear regression option, which requires a model statement and initial parameter values. If one uses the nonlinear option but specifies the linear model and a reasonable initial value for the slope, this option yields results identical to those for Excel – that is, it applies equation (4) to compute R^2! Meanwhile, the Minitab option for RTO gives the same regression equation and standard error as Excel and SPSS, but reports neither the F nor the R^2 statistic. However, Minitab's ANOVA table, from which F and R^2 would be derived, is based on equation (3′).

Because Excel and the nonlinear option in SPSS apply equation (4) regardless of whether an intercept is present, it is easy (and perhaps instructive for students) to construct examples that generate negative R^2 and F statistics for regressions through the origin using these packages. (One need only construct a line with a large intercept and then estimate it without the intercept.) Extreme cases of that sort can provide a springboard for discussion, and make a compelling argument for using equation (4′) rather than equation (4) to evaluate RTO.

The same issues arise, of course, in multiple linear regressions. Consider the nutritional study conducted by Johnson (1995): the caloric contents of various foods are regressed on their fat, protein and carbohydrate contents. For the 13 foods in his sample, OLS yields

$$\begin{array}{c}\text{Calories} = 4.446 + 8.715\ \textit{Fat} + 4.044\ \text{Protein} \\ \qquad [0.395] \quad [0.000] \qquad [0.000] \\ + 3.841\ \text{Carbohydrates} \\ \qquad [0.000] \\ S_e = 6.97 \quad F = 232 \quad R^2 = 0.987 \quad \bar{R}^2 = 0.983\end{array}$$

regardless of which statistical software is used. But here the constant is insignificant and, as Johnson observes, nutritional theory indicates that a constant is inappropriate for this regression. In SPSS, removing the constant gives

$$\begin{array}{c}\text{Calories} = 8.888\ \text{Fat} + 4.266\ \text{Protein} \\ \qquad [0.000] \qquad [0.000] \\ + 3.978\ \text{Carbohydrates} \\ \qquad [0.000] \\ S_e = 6.90 \quad F = 1459.66 \quad R^2 = 0.998 \quad \bar{R}^2 = 0.997\end{array}$$

with all diagnostics indicating an improved fit. Minitab and Excel produce the same equation but different diagnostics. Minitab again reports only S_e, while Excel generates

$$\begin{array}{c}\text{Calories} = 8.888\ \text{Fat} + 4.266\ \text{Protein} \\ \qquad [0.000] \qquad [0.000] \\ + 3.978\ \text{Carbohydrates} \\ \qquad [0.000] \\ S_e = 6.895 \quad F = 236.5 \quad R^2 = 0.986 \quad \bar{R}^2 = 0.883\end{array}$$

In contrast to the previous example, the Excel output now seems more confusing than the SPSS output. Notice that Excel's R^2 and adjusted R^2 statistics for RTO indicate a worse fit, while its S_e and F statistics indicate a better fit, compared to the OLS model.

Given these inconsistencies, Hocking (1996, p. 178) notes: 'It is natural to ask if there is a measure analogous to R^2 for the no-intercept model. We suggest the square of the sample correlation between observed and predicted values'. It can easily be shown that this measure is equal to the unadjusted coefficient of determination for the OLS model. It therefore gives an interpretable measure of the quality of an RTO model, but does not help in comparing RTO with OLS. For that purpose, the best measures appear to be the p-value of the OLS constant and the standard errors of the OLS and RTO regressions. Using these measures, the constant should be retained in the eggs example given above, but not in the nutrition example.

◆ CONCLUSION ◆

Regression through the origin is an important and useful tool in applied statistics, but it remains a subject of pedagogical neglect, controversy and confusion. Hopefully, this synthesis provides some clarity. However, in the light of the unresolved debate, perhaps the strongest conclusion to be drawn from this review is that the practice of statistics remains as much an art as it is a science, and the development of statistical judgment is therefore as important as computational skill.

Acknowledgements

The author would like to thank Donald Dale, Scott Trees and Luigi Ventura, whose discussions of results in another paper prompted him to write this one. He also thanks the editor and anonymous referees for providing helpful comments. Any errors are his own.

References

Adelman, M.A. and Watkins, G.C. (1994). Reserve asset values and the Hotelling valuation principle: further evidence. *Southern Economic Journal*, **61**(1), 664–73.

Casella, G. (1983). Leverage and regression through the origin. *American Statistician*, **37**(2), 147–52.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**(3), 597–604.

Data Bank (1990). Birds, eggs, and databases. *Teaching Statistics*, **12**(2), 62–3.

Gordon, H.A. (1981). Errors in computer packages: least squares regression through the origin. *The Statistician*, **30**(1), 23–9.

Hahn, G.J. (1977). Fitting regression models with no intercept term. *Journal of Quality Technology*, **9**(2), 56–61.

Hocking, R.R. (1996). *Methods and Applications of Linear Models: Regression and Analysis of Variance*. New York: John Wiley.

Johnson, R. (1995). A multiple regression project. *Teaching Statistics*, **17**(2), 64–6.

Kimber, H. (1995). The 'golden egg'. *Teaching Statistics*, **17**(2), 34–7.

Maddala, G.S. (1977). *Econometrics*. New York: McGraw-Hill.

Marquardt, D.W. and Snee, R.D. (1974). Test statistics for mixture models. *Technometrics*, **16**(4), 533–7.

Pettit, L.I. and Peers, H.W. (1991). An example not to be followed? *Teaching Statistics*, (**13**)1, 8.

Prvan, T., Reid, A. and Petocz, P. (2002). Statistical laboratories using Minitab, SPSS, and Excel: a practical comparison. *Teaching Statistics*, **24**(2), 68–75.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley.

Turner, M.E. (1960). Straight line regression through the origin. *Biometrics*, **16**(3), 483–5.