Notes on regression through the origin

by Antal Kozak¹ and Robert A. Kozak²

The calculation of R^2 and the test of significance for the least squares regression solution through the origin are not well documented in textbooks. Many of the statistical packages compute these statistics in such a way that the researcher may easily, though inadvertently, be misled. This paper contains a detailed discussion, supplemented by an example, of the formulae required to properly calculate these statistics. Also, recommendations for correcting misleading computer results are presented.

Keywords: regression through the origin, coefficient of determination, test of significance

Le calcul de R^2 et le test de validité des équations de régression selon les moindres carrés passant par l'origine ne sont pas documentés adéquatement dans les manuels de référence. Plusieurs des programmes d'analyse statistique compile ces statistiques d'une façon telle que le chercheur peut facilement, par inadvertance, être induit en erreur. Cet article comprend une discussion détaillée, appuyée par un exemple, de la formule requise pour calculer adéquatement ces statistiques. De plus, l'auteur avance des recommandations visant à corriger les erreurs entachant les résultats des calculs informatiques.

Mots clés: régression passant par l'origine, coefficient de détermination, test de validité

Introduction

In many instances, regression models which pass through the origin are well suited to describe relationships between dependent and independent variables. In the field of forestry, certain forms of volume equations, tree height and diameter at breast height (DBH) relationships, and some of the biomass equations are but a few of the examples that can be found in the literature. Examples can also be found in the field of wood science where relationships between stress and strain, force and deformation, specific gravity (or density) and strength properties, moisture content and swelling, and relative humidity and moisture content (and many others) can also be seen to pass through the origin.

Having prior knowledge that a regression model passes through the origin modifies the estimation procedure somewhat. This is due to the fact that the estimation of the intercept is no longer required since its value is known. As such, the estimation of regression coefficients (the slope in the case of a simple linear regression) is carried out by setting the intercept to zero.

Although the least squares estimation procedure and the formulae required to estimate the parameters for the "no-intercept" model are well documented in statistical and econometric textbooks, the calculation and interpretation of R^2 and the test of significance of the model are generally not discussed in detail. As a result, researchers often inadvertently use, interpret, and publish results which may erroneously show the no-intercept model to be "better" or statistically "more significant" than the corresponding unrestricted model (which includes an intercept). To add to this inadequate treatment of the subject in textbooks, many of the most widely used statistical packages compute the R^2 and the F or t statistics for the no-intercept model in a manner that is conducive to misleading conclusions. Some of the statistical textbooks (Myers 1986; Montgomery and Peck 1992; Kmenta 1986) and literature (Gordon 1981; Kvalseth 1985) make reference to this problem by way of a few

cautionary remarks. However, these references are not widely read among foresters, forest scientists and wood scientists. In addition, a few of the statistical packages provide some notes on misleading R^2 values (for example, SYSTAT page 476) (Wilkinson 1988). Unfortunately, manuals doubtlessly go unread by casual users.

The purpose of this paper is to discuss the correct interpretation of the no-intercept least squares models and to provide some guidelines for correcting the misleading statistics obtained by using statistical packages. Finally, a numerical example is offered to give the reader a better understanding of the problem at hand.

The Basic Models

To facilitate the discussion and understanding of this problem, the procedures presented in this paper will be restricted to simple linear regressions, though all of the formulae and interpretations shown here can easily be extended to multiple linear and some curvilinear cases. The example illustrated here is a simple linear relationship, common in wood science, between modulus of rupture (the dependent variable, y) and specific gravity (the independent variable, x) for air-dried wood.

Letting $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ represent n data points (a sample from a population, N), the simple linear relationship between the dependent and independent variables can be described by three distinct models:

$$\text{Full model: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (1)$$

$$\text{No intercept model: } y_i = \alpha_1 x_i + \epsilon_i \quad (2)$$

$$\text{Intercept-only model: } y_i = \alpha_0 + \epsilon_i \quad (3)$$

where the errors, ϵ_i , are independently and normally distributed with a zero mean and a σ^2 variance (it should be noted that normality is not required for estimation, only for hypothesis testing). For the full model, the least squares estimates, b_0 and b_1 , are calculated as follows for n pairs of observations:

$$b_0 = \bar{y} - b_1 \bar{x}$$
$$b_1 = SP_{xy} / SS_x$$

The equivalent least squares estimate for the no-intercept model, a_1 , is calculated as:

$$a_1 = \sum xy / \sum x^2$$

¹Professor and Associate Dean, Faculty of Forestry, the University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4.

²Graduate student in Wood Science, Faculty of Forestry, the University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4.

Last, the least squares estimate of α_0 (a_0) for the intercept-only model is calculated simply as:

$$a_0 = \sum y/n = \bar{y}$$

where: $\bar{y}$ = $\sum y/n$, the mean of the y observations
 $\bar{x}$ = $\sum x/n$, the mean of the x observations
 SP_{xy} = $\sum (x-\bar{x})(y-\bar{y})$, the corrected sum of products of the x and y observations
 SS_x = $\sum (x-\bar{x})^2$, corrected sum of squares of the x observations
 $\sum xy$ = the uncorrected sum of products of the x and y observations
 $\sum x^2$ = the uncorrected sum of squares of the x observations,

where the indices for all of the summations range in value from $i = 1$ to n (to simplify the notation, the subscript i has been omitted in the x and y values throughout this article).

The full model (1) is frequently referred to as the "maximum model." That is, of the three models, it accounts for the largest proportion of variation in the dependent variable. For like reasons, models (2) and (3) are referred to as "reduced models".

The physical constraints of the modulus of rupture and specific gravity relationship suggest the use of a simple linear no-intercept model. Thus, equation (2) was fit to estimate modulus of rupture for a sample of 20 air-dried specimens and compared to results obtained by using equation (1), the full model. Table 1 is a summary of the data, some relevant statistics, and estimates of the regression coefficients for both of the models. These statistics will be referred to throughout this paper.

Interpretation

Calculation of the slope or regression coefficient (a_1) for the no-intercept model is a procedure that is well-known to most users. Its derivation is clearly discussed in many of the textbooks. There are, however, two potential problems that users of the no-intercept model should be wary of: the interpretation of the coefficient of determination (R^2), and the interpretation of the test of significance.

(a) Coefficient of Determination (R^2)

Although some statisticians prefer the use of the lower case r^2 for the notation of the coefficient of determination in simple linear regressions, capital R^2 will be used in this paper. This is due mainly to the fact that the theory and formulae involved are the same for both simple linear and multiple regressions.

For the full model the following identity exists:

$$\text{Corrected total sum of squares} = \text{corrected regression sum of squares} + \text{residual sum of squares} \quad (4)$$

or

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2 \quad (5)$$

with degrees of freedom of $n-1 = 1 + n-2$.

The two components on the right-hand side of equation (5) are logical and meaningful. The corrected regression sum of squares (the first term) is the measure of the "variation" (in terms of the sum of squares) of the regression line (or surface) around the mean. The residual sum of squares (the second term) is the measure of the "variation" of the observations around the regression line (or surface). Based on these definitions, the coefficient of determination is seen to be:

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} \quad (6)$$

which represents the proportion of the sum of squares of the dependent variable accounted for by the regression equation. That is, R^2 is the proportion of variation "explained" by the independent variable(s). From equation (5), it can also be seen that R^2 can alternatively be calculated as:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2} \quad (7)$$

It can also be seen from identity (5) that the R^2 value for the intercept-only model (3) is zero. A comparison of the R^2 values obtained from the three basic models (1, 2 and 3) shows considerable variation. Given identity (5), the full model (1) is seen to have the highest R^2 value. This holds unless $\beta_0 = 0$ or $\beta_1 = 0$, since dropping β_0 or β_1 or both from model (1) results in a higher residual sum of squares.

This leaves one question. What value of R^2 can be used for the no-intercept model (2)? The main problem in calculating R^2 originates from the fact that the corrected total sum of squares ($\sum (y-\bar{y})^2$) cannot be partitioned as it was in equation (5). Instead, the uncorrected sum of squares must be partitioned as follows:

$$\sum y^2 = \sum \hat{y}_0^2 + \sum (y-\hat{y}_0)^2 \quad (8)$$

with degrees of freedom of $n = 1 + n - 1$

where: $\hat{y}_0$ represents the predicted values from the no-intercept model; and

$$\sum \hat{y}_0^2 = a_1^2 \sum xy, \text{ the regression sum of squares.}$$

Though equation (8) is algebraically correct, it does not have the same meaning as equation (5). Note that the left-hand side, the uncorrected sum of squares of y (the sum of squares around zero, instead of around the mean of y), is partitioned into the uncorrected regression sum of squares (again, the regression sum of squares around zero instead of around the mean of y) and the residual sum of squares. Of the three terms in equation (8), only the residual sum of squares is comparable to the terms in equation (5), unless the mean of y equals zero. From equation (8), the R^2 analog for the no-intercept model may be calculated as:

$$R^2_0 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{\sum \hat{y}_0^2}{\sum y^2} \quad (9)$$

In fact, this is how many of the computer packages calculate the coefficient of determination for the no-intercept model. One should note, however, that there are distinct differences between the of R^2_0 (9) and R^2 (6) expressions. In the no-intercept model (equation 9), the numerator and denominator represent the sum of squares around zero (uncorrected sum of squares), while in the case of the full model (equation 6), the numerator and denominator represent the sum of squares around the mean of the dependent variable ($\bar{y}$). It is easy to see that a direct comparison of R^2_0 and R^2 is like "comparing apples and oranges."

Using equations (6) or (7) for the full model, the R^2 value obtained in our example is 0.9385. The no-intercept model yields a higher R^2 value of 0.9893 using equation (9). While it is indisputable that, for a given set of observations, the full model explains more of the variation of the dependent variable than any of the reduced models (including the no-intercept one), coefficients of determination greater than 0.9385 (such as the 0.9893 value that was obtained) are misleading and invite researchers to draw false conclusions.

As was previously stated, the residual sum of squares in equation (8) is comparable to the equivalent term in equation (5). This would suggest the use of formula (7) in calculating the coefficient of determination for the no-intercept model:

$$R^2_1 = 1 - \frac{\sum (y - \hat{y}_0)^2}{\sum (y - \bar{y})^2} \quad (10)$$

As well, since $\sum (y - \hat{y}_0)^2 \geq \sum (y - \bar{y})^2$, the R^2_1 calculated from equation (10) will always be less than or equal to the R^2 obtained by the full model (as it should be). For our modulus of rupture and specific gravity example, the coefficient of determination from equation (10) is 0.9079. This value is somewhat less than the R^2 of 0.9385 calculated from the full model. Although we cannot be positive, the reason why many computer packages do not use formula (10) to calculate the correct coefficient of determination may be due to the fact the formula will produce negative values for R^2_1 when:

$$\sum (y - \hat{y}_0)^2 > \sum (y - \bar{y})^2 \quad (11)$$

And how would it look to have a negative coefficient of determination in a computer output? In our opinion, no worse than obtaining an R^2 larger than that of the full model.

It should be noted that, if the above inequality (11) does, in fact, occur, the restricted no-intercept model should not be fitted at all, since the variation around the regression line (or surface) is greater than the variation around the mean. It would be possible for computer packages to use formula (10) to determine the R^2 for no-intercept models, if the printing of any statistics is restricted by the condition that the residual sum of squares is less than the corrected total sum of squares of y . If this condition is not met, the computation should be terminated with a message stating that the no-intercept model is not justified.

(b) Test of Significance

To test the significance of the no-intercept model, an F (or t , the square root of F) test is recommended by the textbooks, and utilized by most statistical packages (from identity (8)):

$$F_{(1,n-1)} = \frac{(a_1 \sum xy) / 1}{\sum (y - \hat{y}_0)^2 / (n-1)} \quad (12)$$

This test seems to be equivalent to the F (or t) test [from identity (5)] used for the full model:

$$F_{(1,n-2)} = \frac{\sum (\hat{y} - \bar{y})^2 / 1}{\sum (y - \hat{y})^2 / (n-2)} \quad (13)$$

If we consider that $a_1 \sum xy = \sum \hat{y}^2_0$, tests (12) and (13) differ for the same reasons that R^2_0 and R^2 do. For the no-intercept

model, the numerator of (12) is the sum of squares around the abscissa (uncorrected regression sum of squares). For the full model, the numerator of (13) is the sum of squares around the mean, $\bar{y}$ (corrected regression sum of squares). The F value obtained in our example is 274.55 with 1 and 18 degrees of freedom using test (13) for the full model, and 1751.15 with 1 and 19 degrees of freedom using test (12) for the no-intercept model. For the uninformed user, these values indicate a "more significant" or better relationship when the no-intercept model is employed.

For most users, a "significant regression" means a good, well-defined relationship between the dependent and independent variables. In the case of the full model, this interpretation is correct when there exists no lack of fit. In any case, when the null hypothesis of $\beta_1 = 0$ is rejected (i.e. a significant regression), it can be stated that the regression equation accounts for a significant portion of the variation of the dependent variable. However, in the case of the no-intercept model, a significant F or t value does not necessarily mean a good, well defined relationship between the dependent and independent variables. Because of the nature of the F -test (12) for the no-intercept model, the size of the F -value depends on the position of y values. If the y values are far from zero, the chances of obtaining a significant F -value are very good, even if the no-intercept model describes the data very poorly.

Consider the following four pairs of hypothetical observations:

x	y
3	4
4	4
4	3
5	3

The full model for these observations is $\hat{y} = 5.5 - 0.5x$. Note that the slope is negative, and that the relationship is not significant ($F_{(1,2)} = 1.00$). The no-intercept model for the above data is $\hat{y}_0 = 0.83x$. Applying formula (12), the slope (0.83) is seen to be significantly different from zero ($F_{(1,3)} = 33.07$), not because the model describes the x and y relationship well, but because the points are far enough from the origin (both x -wise and y -wise) (Figure 1). We can move these points closer to the origin such that their relative position to each other remains the same ($x' = x - 3$, $y' = y - 2$) and produce the same slope (0.83) for a no-intercept model (Fig. 1).

x'	y'
0	2
1	2
1	1
2	1

Note that the significance of the full model remains the same ($F_{(1,2)} = 1.00$), while the significance of the no-intercept model changes to ($F_{(1,3)} = 2.14$). This F -value now indicates that the slope of 0.83 of the no-intercept model is not significantly different from zero.

The above example clearly shows that the F (or t) test (12) for the no-intercept model does not indicate the "goodness" of the relationship. As such, the researcher must handle and interpret these tests with care. There are, however, two equivalent tests available to the user for verifying the "goodness" of no-intercept models. These two tests determine whether or not the no-intercept model differs significantly from the full

Figure 1. An example of observations affecting the test of significance of the no-intercept model.

model. If the full model is significant ($\beta_1 \neq 0$) and if there is no significant difference between the full and no-intercept models, one can safely use the no-intercept model to describe the relationship between x and y . The two equivalent tests are:

(i) *Is the intercept of the full model significantly different from zero?* Most statistical packages provide a t value for this test. If not, it can be calculated as:

$$t_{(n-2)} = \frac{b_0}{SE(b_0)} \quad (14)$$

where: $SE(b_0)$ is the standard error of the intercept

$$SE(b_0) = \sqrt{\frac{S^2_{y,x} \sum x^2}{n SS_x}}$$

$S^2_{y,x}$ = the residual variance or MS_{Res} (residual mean squares); or

(ii) *Is the residual sum of squares of the no-intercept model significantly greater than the residual sum of squares of the full model?*

$$F_{(1,n-2)} = \frac{(SS_{Res0} - SS_{Res}) / 1}{SS_{Res} / (n-2)} \quad (15)$$

where: SS_{Res0} = residual sum of squares of the no-intercept model; and

SS_{Res} = residual sum of squares of the full model.

For our example, the t value for testing the significance of the intercept is 2.99, while the F value for testing the difference between the two residual sums of squares is 8.93 or t^2 (Table 1). Thus, in our case, the no-intercept model does not provide a "good relationship" between the dependent and independent variables.

Recommendations

Users of statistical packages can easily remedy any misleading results obtained by determining the no-intercept solution. To do so, they must obtain output for both the full and the no-intercept models as well as their corresponding statistics. If the value of the coefficient of determination for the no-intercept model is greater than for the full model, it should be corrected.

Table 1. Data used for the example and some relevant statistics

Specific gravity (x)	Modulus of rupture (1,000 lbs/inch ²) (y)	
0.680	15.72	$\sum x = 9.449$
0.631	14.86	
0.348	6.14	$\sum y = 212.02$
0.413	8.21	
0.698	17.07	$\sum x^2 = 4.829139$
0.368	9.07	
0.571	14.68	$\sum y^2 = 2,544.2952$
0.433	10.37	
0.252	5.18	$\sum xy = 110.24905$
0.387	9.36	
0.409	7.61	$SS_x = 0.36496$
0.456	10.43	
0.375	8.93	$SS_y = 296.6712$
0.495	10.33	
0.550	14.46	$SP_{xy} = 10.0802$
0.576	12.39	
0.265	4.06	$n = 20$
0.299	4.67	
0.612	13.73	
0.631	14.75	

	Full model	No-intercept model
$\hat{y}$	$= -2.448 + 27.620x$	$\hat{y}_0 = 22.830x$
$S_{y,x}$	1.007	1.200
R^2	0.9385	0.9079 (0.9893) ¹
F	274.55	-(1751.15)
SS_{Res}	18.2561	27.3094
SS_{Reg}	278.415	-(2516.9858)
Std. error of slope	1.667	0.546
Std. error of intercept	0.819	-

$$F_{(1,18)} = \frac{(27.3094 - 18.2561) / 1}{18.2561 / 18} = 8.93$$

¹Values in parentheses are based on the no-intercept formulae found in most statistical packages.

Interpretation of the F or t value provided by the package for testing the significance of the regression should always be avoided. The following steps are recommended for correcting the results:

- The corrected total sum of squares, $\sum (y - \bar{y})^2$ is usually available from the output of the full model.
- The residual sum of squares for the no-intercept model is usually provided as part of the output for this model. If not, the residual sum of squares can be calculated using the standard error of estimate ($S_{y,x0}$):

$$\sum (y - \hat{y}_0)^2 = (n - 1) S_{y,x0}^2$$

- Using the obtained values, the R^2 can be corrected as:

$$R^2_1 = 1 - \frac{\sum (y - \hat{y}_0)^2}{\sum (y - \bar{y})^2}$$

- If a test is not available for the significance of the intercept in the full model output, the F -test computed by formula (15) can be carried out, as the residual sums of squares for both of the models are usually available from the respective outputs.

Readers interested in a more theoretical discussion regarding the effect of the above tests on the mean squared error performance of the non-intercept model are referred to Gregoire *et al.* (1992).

As a final precautionary note, it is recommended that no-intercept models be used only when theoretically justified. Furthermore, there should be no significant difference between the residual variances of the full and no-intercept models. Also, if possible, it is recommended that observations near the origin be collected if the objective is to fit a no-intercept model. In our example, theoretical justification warrants the use of a no-intercept model. However, the residual variances are significantly different (see the *F*-test in Table 1), indicating that either the no-intercept model is not justified for the range of data used to fit the models, or that the relationship between modulus of rupture and specific gravity is not linear for the full range of observations.

References

Gordon, H.A. 1981. Errors in computer packages. Least squares regression through the origin. *The Statistician* 30: 23-29.

Gregoire, T.G., A.A. Arabatzis and M.R. Reynolds, Jr. 1992. Mean squared error performance of simple linear regression conditional upon the outcome of pretesting the intercept. *The American Statistician* 46(2): 89-93.

Kmenta, J. 1986. *Elements of Econometrics*. MacMillan Publishing Company, New York. 786 p.

Kvalseth, T.O. 1985. Cautionary note about R^2 . *The American Statistician* 39(4): 279-285.

Montgomery, D.C. and E.A. Peck. 1992. *Introduction to linear regression analysis*. Second edition. John Wiley, New York. 527 p.

Myers, R.H. 1986. *Classical and Modern Regression with Applications*. Duxbury Press, Boston. 359 p.

Wilkinson, L. 1988. *SYSTAT: The System for Statistics*. Evanston, IL: SYSTAT, Inc. 822 p.