2013 Hawaii University International Conferences
Education & Technology
Math & Engineering Technology
June 10th to June 12th
Ala Moana Hotel, Honolulu, Hawaii

An Algebraic Connection between Ordinary Lease-square Regression and Regression though the Origin

Dr. Xiaohui Zhong

University of Detroit Mercy, Detroit, Mi

An Algebraic Connection between Ordinary Lease-square Regression and Regression though the Origin

Xiaohui Zhong

University of Detroit Mercy, Detroit, USA

Keywords: Regression; Regression through the origin; Breakdown points; R squared.

Abstract

Ordinary least-square regression(OLS) or regression through the origin(RTO)? That is the question. This paper tries to establish an algebraic relation of the slopes and R squares between these two models and study the connection between them. One of the results can be used as a pedagogical tool to construct data set with breakdown point.

Introduction

In the first course of statistics, OLS model is always the one being introduced. However, in the reality, especially in some physical model, RTO may be more appreciate. For this special case of simple linear regression, most textbooks discuss it the same as a special case of the OSL by dropping the constant term while others cautiously warn that not to use this model unless necessary. Many literatures indicted that these models are not comparable. But is there any connection between them? Traditionally, the classical simple linear regression model

$$y = \alpha + \beta x \qquad (1)$$

was used to fit the observations for two variables x_i and y_i for $i = 1,2,...,n$, where the coefficients can be estimated as

$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2)$$

$$\alpha = y - \beta\bar{x} \qquad (3)$$

by the ordinary least squares (OLS) method, with $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. Here we use the same notation for the estimator of α and β for simplicity.

This model was widely used in a wide range of areas including engineering, social studies, bio-science, etc. In this model, one important statistics for checking whether the model is a "good" fit is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (4)$$

where $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ and $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. This identity is true because $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. A general rule of thumb is that the closer the R^2 is to 1, the better fit the model is to the observations. On the other hand, the same set of data can also be fitted to the linear model:

$$y = \beta x \qquad (5)$$

Such a model is called regression through the origin (RTO). In this model the coefficient is given by

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \qquad (6)$$

with the corresponding R^2 is given by

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} \hat{y}_i^{\,2}}{\sum_{i=1}^{n} y_i^{\,2}} \qquad (7)$$

However identity (4) is no longer true.

In many practical situations, RTO is necessary. Such necessity was well discussed by Casella[1983] and Eisenhauer [2003]. The appropriateness of adopting either OLS or RTO model was also discussed in these two papers. It was concluded that "regression through the origin is an important and useful tool in applied statistics, but it remains a subject of pedagogical neglect, controversy and confusion Eisenhauer [2003] because the R squares and other statistics are not comparable. While calculating the basic regression model it is now fairly easy using any statistical software or even handheld calculator, the RTO model is also just one click away. Practitioners in other fields are still not clear the differences between these two models. Even in SPSS there is such 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. Some also observed that in most of the cases, both R squares and F both increase from OLS to RTO. Is this always true? If so, what condition(s) will make them larger? What is the relationship between these statistics? What else can we look for while selecting an appropriate model? These are few questions we want to answer in this paper using simple algebra and calculus techniques.

Augmented Data

One way to study the relation between these two models is to add a special augmented point $(x_{n+1}, y_{n+1}) = \frac{n}{\sqrt{n+1}-1}(\bar{x}, \bar{y})$, Casella[1983]. The leverage impact of this point was studied and was made use for comparing the two models. To learn the connection between the OLS and RTO models, we investigate a broader model by introducing a general augmented point $(x_{n+1}, y_{n+1}) = (k\bar{x}, k\bar{y})$ to the data set with k being an arbitrary constant. If we denote $\bar{x}_{n+1} = \frac{\sum_{i=1}^{n+1} x_i}{n+1}$ and $\bar{y}_{n+1} = \frac{\sum_{i=1}^{n+1} y_i}{n+1}$, then simple algebra shows that

$$\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1}) = \sum_{i=1}^{n}(x_i y_i - rn\bar{x}\bar{y})$$

where
$$r = \frac{(n+1)-(k-1)^2}{n+1} \qquad (8)$$

The OLS model with the augmented point will have the following parameters:

$$\beta(r) = \frac{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})}{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2} = \frac{\sum_{i=1}^{n}(x_i y_i - rn\bar{x}\bar{y})}{\sum_{i=1}^{n} x_i^2 - rn\bar{x}^2}$$

$$R^2(r) = \frac{SSR}{SST} = \frac{[\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})(y_i - \bar{y}_{n+1})]^2}{\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 \sum_{i=1}^{n+1}(y_i - \bar{y}_{n+1})^2}$$

$$= \frac{[\sum_{i=1}^{n}(x_i y_i - rn\bar{x}\bar{y})]^2}{(\sum_{i=1}^{n} x_i^2 - rn\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - rn\bar{y}^2)}$$

If $r = 1$, then $k = 1$, and

$$\beta(1) = \frac{\sum_{i=1}^{n}(x_i y_i - n\bar{x}\bar{y})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \quad \text{and} \quad R^2(1) = \frac{[\sum_{i=1}^{n}(x_i y_i - n\bar{x}\bar{y})]^2}{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}$$

which agree with (2) and (4) for the OLS. This means that the parameters above are those for the OLS with the original data set.

When $r = 0$, then $k = \sqrt{n+1}+1$, and

$$\beta(0) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \quad \text{and} \quad R^2(0) = \frac{[\sum_{i=1}^{n} x_i y_i]^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

which are exactly (6) and (7) for the RTO. Thus we restrict r to be between 0 and 1 with $k \in [1, \sqrt{n+1}+1]$ and the two models can be obtained by a unified OLS method with a parameter $r \in [0,1]$. We will study how the model parameters change from OLS to RTO by varying r from 1 to 0.

The slopes of the regression models and breakdown point

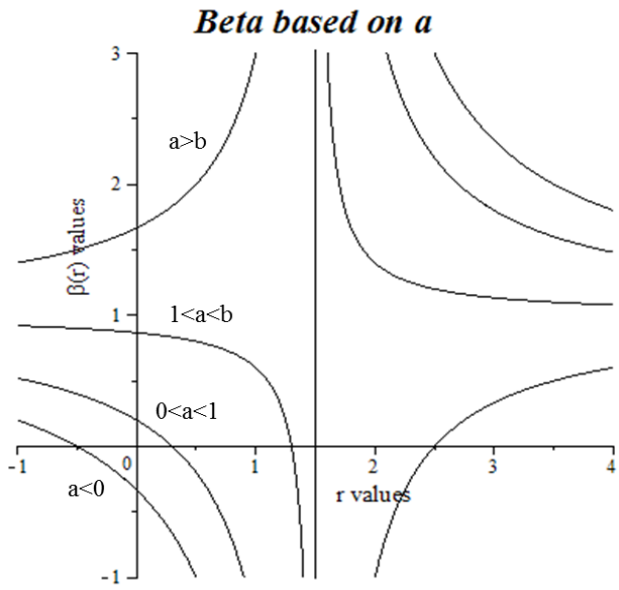
In both models, OSL and RTO, the slope $\beta(r)$ plays an important role in the prediction. The sign of the slope indicates the direction of the dependent variable linearly correlated with the independent variable and the magnitude tells the rate of change. For the purpose of this study, we will only consider the cases when $\bar{x} \neq 0$ and $\bar{y} \neq 0$. Actually, if both of these means are 0, the OSL will be exactly RTO already.

Now we can write

$$\beta(r) = \frac{\sum_{i=1}^{n}(x_i y_i - rn\bar{x}\bar{y})}{\sum_{i=1}^{n} x_i^2 - rn\bar{x}^2} = m\frac{r-a}{r-b} \qquad (9)$$

where $a = \frac{\sum_{i=1}^{n}(x_i y_i)}{n\bar{x}\bar{y}}$, $b = \frac{\sum_{i=1}^{n} x_i^2}{n\bar{x}^2}$, and $m = \frac{\bar{y}}{\bar{x}}$.

Without lost of generality, we will consider the case when $M > 0$ the results will be opposite when $M < 0$. In addition, we will avoid the rare case when $a = b$. Notice that $b > 1$, we will analyze how the slope changes from OSL to RTO by simply looking at the graph of the rational function $\beta(r) = m(1 + \frac{b-a}{r-b})$.



From the graph, it can be seen that $\beta(r)$ is always a monotonic function of $r \in [0,1]$. The location of the statistics a relative to this interval determines the sign of $\beta(r)$ and whether it is increasing or decreasing. The different cases can be summarized as follows.

Case a: The slope of OLS is positive and greater than the slope of RTO, or

$0 < \beta(0) < \beta(r) < \beta(1)$ for $r \in (0,1)$, iff $a = \dfrac{\sum_{i=1}^{n} x_i y_i}{n\overline{xy}} > b = \dfrac{\sum_{i=1}^{n} x_i^2}{n\overline{x}^2}$.

We see that the graph of the rational function (9) is above m and increasing for $r \in [0,1]$. Translating to the OSL and RTO model, it can be interpreted as that the slope of OSL is greater than $m = \dfrac{\overline{y}}{\overline{x}}$, as it decreases toward the slope of RTO which is also greater than m.

Notice that the line obtained with $r = 0$ is not the real RTO, but it is a line parallel to the RTO.

Case b: The slope of OLS is positive and smaller than that of the RTO , or

$\beta(0) > \beta(r) > \beta(1) > 0$ for $r \in (0,1)$, iff $1 < a < b$.

Case c: The slope of OLS is negative and smaller than that of the RTO, or

$0 > \beta(0) > \beta(1) > \beta(1)$ for $r \in (0,1)$, iff $a < 0$.

Case d: This is the most interesting case: The sign of the slope of OLS is negative and the sign of the slope of RTO is positive, or $\beta(0) < 0 < \beta(1)$, iff $0 < a < 1$. Furthermore $\beta(a) = 0$, this particular regression line will have a slope 0.

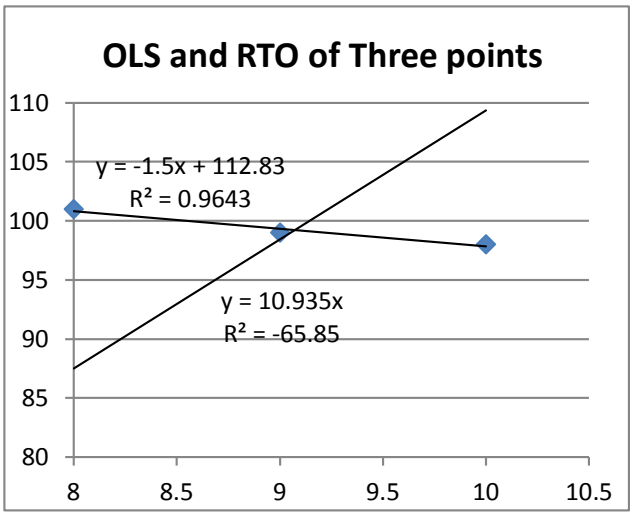
These observations may not be very significant now. But the last case can be used as a pedagogy method to construct the data set with breakdown points. With this method, simple algebra can be used construct the data. There are many ways to define breakdown points. They are the points in a set of data so "contaminated" such that the model does not convey useful information. One simple definition is found in Chan[2001].

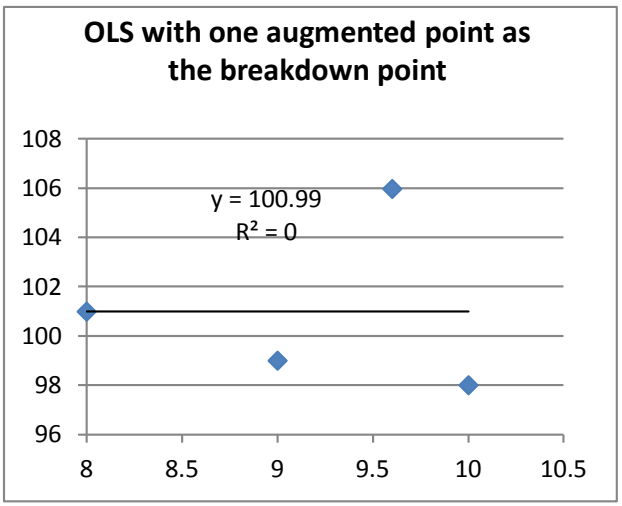
Definition: The breakdown point of a procedure for fitting a line to n pairs of bivariate data is defined as the minimum proportion of the sample that, when replaced by suitably chosen outliers, can drive the slope to infinity or zero, Chan [2001].

It was concluded that the slope β in model (1) cannot tolerate with even one outlier with a breakdown point of $1/n$. Chan[2001] gave an algebraic method by trial and error of asymptotic breakdown point. Here we can provide another method of constructing a set of data with a very good OLS fit, but with one outlier, the regression line becomes totally useless, meaning that the slope becomes exactly zero. To construct such an example, we can start with a set that obviously have a negative slope in OLS and positive slope in RTO. Such data set can be generated with any line such $y_i = \beta x_i + \alpha + \varepsilon_i$, where $\beta < 0$ and $\alpha > 0$, and $\varepsilon_i \sim N(0, \sigma)$ for positive pairs of (x_i, y_i). An RTO for this set will have a positive slope because negative slope will produce all negative predicted y values. On the other hand the slope of OLS should be negative because it is estimating β_0. Thus, from the result in case d, it can be concluded that the statistics a must be between 0 and 1. Following example illustrates the process.

Example: We start with the simplest set of data: (8, 101), (9, 99), (10, 98). The OLS model is

$y = -1.5x + 112.83$ with $R^2 = 0.9643$ which is considered a very good fit. If RTO is introduced, we will have $y = 10.935x$. Using Trendline comment in Excel will produce a meaningless result $R^2 = -65.85$. This phenomena is the result of inccorect implementation of the R squared formula discussed in Eisenhauer [2003]. However, the statistics $a = 0.998881$ will help finding an outlier. Letting $r = a = 0.999881$ and solve for k from (8), we have $k = \sqrt{(1-a)(n+1)} + 1 = 1.06689$. The augmented point is $k(\bar{x}, \bar{y}) \approx (9.6,\ 106)$. Now the OSL on the set of data (8, 101), (9, 99), (9.6, 106), (10, 98) will have a slope 0 which no one wants to have.





The advantage of this method is to have algebraic steps to create a set that show that the breakpoint of the slope of a SLO has a break point of $\frac{1}{n}$. Since the slope $\beta(r)$ varied continuously between $\beta(0)$ and $\beta(1)$, this means that we can also create a set of data with one extra point to have the slope that is between these two slopes in a similar fashion. But how well is the fitting? We will discuss this by investigating change of an indicator R^2 next.

The R^2

Applicators always get excited when they saw that R^2 becomes bigger while moving from OLS to RTO in most of the observations [quote about the website]. However, it may not be very meaningful to have a bigger R^2. Following, we will find out some cases that we are certain that

the R^2 becomes bigger, while other cases it can be affirmed that it is smaller. More important, we will use the correct definition so that R^2 will not become negative.

The formula of R^2 should be very familiar to most students even though it is slightly complicated than the slope formula. Recall that for a set of data with augmented point defined in section 2,

$$R^2(r) = \frac{(\sum_{i=1}^{n} x_i y_i - rn\overline{xy})^2}{(\sum_{i=1}^{n} x_i^2 - nr\overline{x}^2)(\sum_{i=1}^{n} y_i^2 - nr\overline{y}^2)} = \frac{(r-a)^2}{(r-b)(r-c)} \qquad (10)$$

where a and b are defined in last section, and $c = \frac{\sum_{i=1}^{n} y_i^2}{n\overline{y}^2}$.

For $r \le 1$, the R square $R^2(r)$ is that for the a set of $(n+1)$ data, thus it is always true that $0 \le R^2(r) \le 1$. Using this definition, the R square in the example is 0.989. To study this rational function, we will consider all r except the values of b and c where this function has veridical asymptotes, but the value of the function will exceed 1 or become negative when $r > 1$. We will also avoid the very rare and special cases when $a = b$, $b = c$, $a = c$ and $a = \frac{b+c}{2}$, because they are all equivalent and it happens if and only if $y_i = mx_i$ for all $i = 1,2,..,n$. Without lost of generality, we assume that $b < c$. Elementary calculus reveals that $R^2(r)$ should have two critical points $r_1 = a$ and $r_2 = \frac{2bc - a(b+c)}{(b+c) - 2a}$. We will see that the location of these two critical points determines how $R^2(r)$ changes on the interval [0, 1]. First, we see that one and only one of the critical points must be between b and c. The other critical must be smaller than b, otherwise, $R^2(r)$ will become greater than 1 for all $r < 1$. As a by-product, this result can be summarized as a theorem:

Theorem: Let $\{(x_i, y_i) \mid i = 1,2,..n\}$ be n pairs of points of reals such that $\overline{x} \neq 0$, $\overline{y} \neq 0$, and $y_i \neq mx_i$ for all $i = 1,2,...,n$, then one of the following must be true:

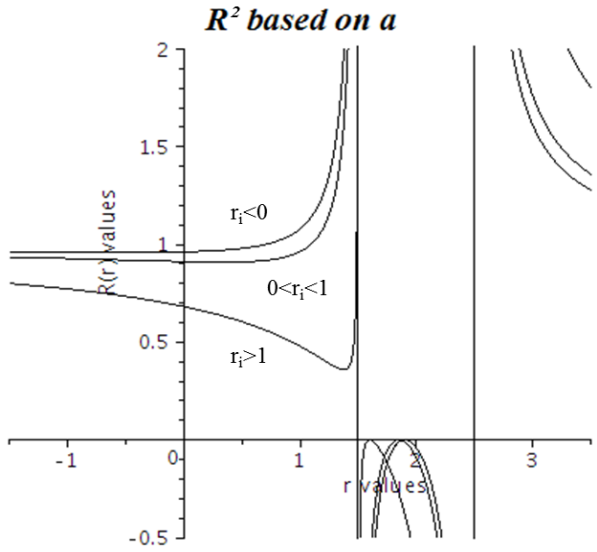
$$a < \min(b,c) < \frac{2bc - a(b+c)}{(b+c) - a} < \max(b,c)$$

or

$$\frac{2bc - a(b+c)}{(b+c) - a} < \min(b,c) < a < \max(b,c)$$

where $a = \frac{\sum_{i=1}^{n} x_i y_i}{n\bar{x}\bar{y}}, b = \frac{\sum_{i=1}^{n} x_i^2}{n\bar{x}^2}$, and $c = \frac{\sum_{i=1}^{n} y_i^2}{n\bar{y}^2}$.

With the result of this theorem, the changes of R squares from OLS to RTO can be summarized as follows by investigating the graph of the rational function defined by (10).



Case I: R square is smaller when moving from OLS to RTO, meaning $R^2(r)$ is increasing when r changes from 0 to 1 iff $r_i < 0$, $i = 1$ or 2.

Case II: R square is greater when moving from OLS to RTO, meaning $R^2(r)$ is decreasing when r changes from 0 to 1 iff $r_i > 1$, $i = 1$ and 2.

Case III: If $0 < r_i < 1$, $i = 1$ or 2, then $R^2(r)$ will reach a minimum at r_i. We will not be able to judge which R square is greater without actual calculating them. But in such a case, it is possible to construct a data set that both $R^2(0)$ and $R^2(1)$ are the same.

Conclusion:

OLS and RTO models remain to be a myth. However, with basic algebra, we can establish certain relationship between their slopes and R square. Such relationship can be used as a pedagogy tool to reveal to student the fact that R square is not always bigger with RTO. The future work will be to establish similar relationship between the *F* statistics.

References

Casella, G. (1983). Leverage and Regression through the Origin. *American Statistician*,

37(2), 147–52.

Chan, W. (2001), Teaching the Concept of Breakdown Point in Simple Linear Regression, *International Journal of Mathematical Education in Science and Technology*, Volume 32, No.5, 745-794.

Eishenhauer, J. (Autumn 2003), Regression through the Origin, *Teaching Statistics*, Volume 25, Number 3, 76-80