The benefits of teaching inverse regression alongside Least Squares Regression: Deeper comparisons for undergraduate research

Review article

Published in Nairobi, Kenya by Royallite Global in the ***Research Journal in Mathematics, Econometrics and Statistics,*** Volume 2, Issue 1, 2021

Article Information
Submitted: 2nd November 2020
Accepted: 29th January 2021
Published: 11th February 2021
Conflict of Interest: No conflict of interest was reported by the authors
Funding: None

Additional information is available at the end of the article

To read the paper online, please scan this QR code

Di Gao [1] & Stephen M. Scariano [2]
Department of Mathematics and Statistics, Sam Houston State University, USA
Correspondence: di.gao@shsu.edu
https://orcid.org/0000-0002-3596-1456

Abstract

This article is a continuation of the authors' previously published article, later referred as "Part I", and entitled, "The benefits of teaching inverse regression alongside Least Squares Regression: Graphical and numerical comparisons". In Part I of this companion series, a foundational exposition comparing Inverse Regression and Least Squares Regression was undertaken using temperature data for thirty-two American cities. Deeper relationships are explored in this article (Part II of this series). The goal is to contrast the estimates provided by both regression methods using a collection of corollaries that are accessible to undergraduate mathematics and science students who have studied Least Squares Regression. Collectively, these two articles demonstrate how to purposely enhance a general discussion of Least Squares Regression.

Keywords: inverse regression, least squares, regression, temperature data, undergraduates

How to Cite:
Gao, D., & Scariano, S. (2021). The benefits of teaching inverse regression alongside Least Squares Regression: Deeper comparisons for undergraduate research. *Research Journal in Mathematics, Econometrics and Statistics*, 2(1). Retrieved from https://royalliteglobal.com/jmes/article/view/438

Public Interest Statement

In Part I of this companion series, a foundational exposition comparing Inverse Regression and Least Squares Regression was undertaken using temperature data for thirty-two American cities. Deeper relationships are explored here in Part II of this series. The goal is to contrast the estimates provided by both regression methods using a collection of corollaries that are accessible to undergraduate mathematics and science students who have studied Least Squares Regression.

1. Introduction

From Part I, given a collection of points $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots\ldots\ldots, (x_n, y_n)$ for which a linear trend model is reasonable, the Inverse Regression and Least Squares "x on y" Regression techniques provide equations for estimating an "x" variable from knowledge of a corresponding "y" variable associated with the trend. Recall from that discussion that the Least Squares "x on y" Regression actually uses the Least Squares methodology with the transposed coordinates $(y_1, x_1), (y_2, x_2) \ldots\ldots\ldots\ldots\ldots, (y_n, x_n)$, while Inverse Regression uses the original ordered pairs $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots\ldots\ldots, (x_n, y_n)$ with the usual Least Squares "y on x" Regression equation simply inverted. In general, the estimating equations are:

Least Squares "x on y" Regression: ***Inverse Regression:***

$$\hat{x}_i = m_{x|y} y_i + b_{x|y} \qquad \tilde{x}_i = m_{Inv} y_i + b_{Inv} \qquad (1)$$

where

$$m_{x|y} = r_{yx}\left(\frac{S_x}{S_y}\right),\ b_{x|y} = \bar{x} - m_{x|y}\bar{y} \ \ \textit{and} \ \ m_{Inv} = \left(\frac{1}{m_{y|x}}\right) = \left(\frac{S_x}{r_{xy}S_y}\right),\ b_{Inv} = \left(\frac{-b_{y|x}}{m_{y|x}}\right) = \bar{x} - \left(\frac{S_x}{r_{xy}S_y}\right)\bar{y}$$

with

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,\ S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)}, S_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2,}{(n-1)}\ S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)},$$

and

$$r_{xy} = r_{yx} = r = \frac{S_{xy}}{S_x \times S_y}.$$

For the entire discussion, the "tilde" symbol is associated with estimates or statistics computed using the Inverse Regression methodology, while the "hat" symbol is associated with similar quantities computed using the Least Squares technique. Remaining notation is consistent with that defined in Part I.

2. Comparing Inverse Regression and Least Squares "x on y" Regression

The purpose in the section is to explore relationships between these methods of estimation. Using direct substitution into the equations in (1),

$$\begin{aligned} \hat{x}_i &= m_{x|y} y_i + b_{x|y} & \tilde{x}_i &= m_{Inv} y_i + b_{Inv} \\ \hat{x}_i &= m_{x|y} y_i + (\bar{x} - m_{x|y}\bar{y}) & \tilde{x} &= \left(\frac{1}{m_{y|x}}\right) y_i + \left(\frac{-b_{y|x}}{m_{y|x}}\right) \\ \therefore (\hat{x}_i - \bar{x}) &= m_{x|y}(y_i - \bar{y}) & \therefore m_{y|x}(\tilde{x}_i - \bar{x}) &= (y_i - \bar{y}) \end{aligned} \qquad (2)$$

Note also that

$$m_{x|y} m_{y|x} = \left[r_{yx}\left(\frac{S_x}{S_y}\right)\right]\left[r_{xy}\left(\frac{S_y}{S_x}\right)\right] = r^2 \qquad (3)$$

(Product of slopes is the square of sample Pearson Correlation Coefficient),

and

$$(\hat{x}_i - \bar{x}) = m_{x|y}(y_i - \bar{y}) = \ m_{x|y} m_{y|x}(\tilde{x}_i - \bar{x}) = r^2(\tilde{x}_i - \bar{x}) \qquad (4)$$

using equations (2) and (3).

Before proceeding further, recall that the usual Least Squares Decomposition of the Total Sum of Squares (SST), (See Peck (2015)), in this context is

$$\mathrm{SST} := \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2 + \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 := \mathrm{S\hat{S}E} + \mathrm{S\hat{S}R},$$
(5)

where the defined terms $\mathrm{S\hat{S}E}$ and $\mathrm{S\hat{S}R}$ denote the usual residual (or error) and regression sums of squares, respectively. Using equations (1) – (4), it is easy to verify decomposition (5) directly since

$$S\hat{S}R = \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 = \sum_{i=1}^{n} m_{x|y}^2 (y_i - \bar{y})^2 = \left(\frac{rS_x}{S_y}\right)^2 \sum_{i=1}^{n}(y_i - \bar{y})^2 = \left(\frac{r^2 S_x^2}{S_y^2}\right)(n - 1)S_y^2 = r^2 \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

and (6)

$$\begin{aligned} S\hat{S}E &= \sum_{i=1}^{n}(x_i - \hat{x}_i)^2 = \sum_{i=1}^{n}\left[(x_i - \bar{x}_i) - m_{x|y}(y_i - \bar{y})\right]^2 = (n-1)S_x^2 - 2(n-1)m_{x|y} r S_x S_y + (n-1)(m_{x|y})^2 S_y^2 \\ &= (n-1)S_x^2 - 2(n-1)\left(\frac{rS_x}{S_y}\right)\left(rS_x S_y\right) + (n-1)\left(\frac{rS_x}{S_y}\right)^2 \left(S_y^2\right) = (n-1)S_x^2 - 2(n-1)r^2 S_x^2 + (n-1)r^2 S_x^2 \\ &= (n-1)S_x^2\left(1 - r^2\right) = (1 - r^2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2. \end{aligned}$$

Therefore, the Least Squares Decomposition of the Total Sum of Squares can alternately be written as

$$SST = \sum_{i=1}^{n}(x_i - \bar{x})^2 = (1 - r^2)\sum_{i=1}^{n}(x_i - \bar{x})^2 + r^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = S\hat{S}E + S\hat{S}R$$

(7)

Now, by its very definition, the estimate $\hat{x}_i$ is superior to estimate $\tilde{x}_i$ in the Least Squares sense (and a proof is given in Corollary 2 below). That is,

$$S\hat{S}E := \sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \le \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 := S\tilde{S}E, \qquad (8)$$

where $S\hat{S}E := \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$is defined to be the residual (or error) sum of squares for the Least Squares "x on y" Regression, and $S\tilde{S}E := \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2$is defined to be the residual (or error) sum of squares for the Inverse Regression of **y** on **x**. However, given the computational simplicity of the Inverse Regression technique, a natural question is: "Is the difference $S\tilde{S}E - S\hat{S}E$ or percentage error of much practical importance?" That question is considered next with the aid of two corollaries.

Corollary 1: *The residual sum of squares,* $S\tilde{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2$, *from the Inverse Regression of* ***y*** *on* ***x***
satisfies

$$S\tilde{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 = \left(\frac{1}{r^2} - 1\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 \ge 0. \qquad (9)$$

The regression sum of squares, $S\tilde{S}R = \sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2$, *from the Inverse Regression of* ***y*** *on* ***x***
satisfies

$$S\tilde{S}R = \sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2 = \left(\frac{1}{r^2}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 \ge 0. \qquad (10)$$

provided $r \neq 0$.

Proof: Using equations (1) – (4) and expanding,

$$S\tilde{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 = \sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - \tilde{x}_i)]^2$$

$$= \sum_{i=1}^{n}[(x_i - \bar{x})^2 - 2(x_i - \bar{x})(\tilde{x}_i - \bar{x}) + (\tilde{x}_i - \bar{x})^2]$$

$$= \sum_{i=1}^{n}\Bigg[(x_i - \bar{x})^2 - \frac{2}{r^2}(x_i - \bar{x})(\widehat{x}_i - \bar{x})$$

$$+ \frac{1}{r^4}(\widehat{x}_i - \bar{x})^2\Bigg] \qquad \text{using (4)}$$

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

$$+ \frac{m_{x|y}^2}{r^4}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{using (2)}$$

$$= (n-1)S_x^2 - \left(\frac{2}{r^2}\right)\left(\frac{rS_x}{S_y}\right)\left[(n-1)S_{xy}\right] + \left(\frac{1}{r^4}\right)\left(\frac{rS_x}{S_y}\right)^2 (n-1)S_y^2 \qquad \text{using (1)}$$

$$= (n-1)S_x^2 - \left(\frac{2}{r^2}\right)\left(\frac{rS_x}{S_y}\right)\left[(n-1)rS_xS_y\right] + \left(\frac{1}{r^4}\right)\left(\frac{rS_x}{S_y}\right)^2 (n-1)S_y^2$$

$$= \left(\frac{1}{r^2} - 1\right)(n-1)S_x^2$$

$$= \left(\frac{1}{r^2} - 1\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0$$

□

Now, using equation (4)

$$S\tilde{S}R = \sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2 = \sum_{i=1}^{n}\left[\left(\frac{1}{r^2}\right)(\hat{x}_i - \bar{x})\right]^2$$

$$= \left(\frac{1}{r^4}\right)\sum_{i=1}^{n}(\hat{x}_i$$

$$- \bar{x})^2 \qquad \text{using (6)}$$

$$= \left(\frac{1}{r^4}\right)\left[r^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \left(\frac{1}{r^2}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

□

This corollary is numerically verified in the example below.

Example 1: Recall from Part I that $S_x^2 = (32-1) * 340.943 = 10{,}569.233$ and $r = 0.9416$. Also, the bottom portion of Table 4 of Part I gives $\mathrm{S\tilde{S}E} = 1{,}352.181$, which roughly agrees with

$$S\tilde{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 = \left(\frac{1}{r^2} - 1\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 \simeq \left(\frac{1}{0.9416^2} - 1\right) \times 10{,}569.233$$
$$= 1{,}351.709,$$

aside from some slight round-off error in the table itself. Likewise,

$$S\tilde{S}R = \sum_{i=1}^{n}(\tilde{x}_i - \bar{x})^2 = \left(\frac{1}{r^2}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{0.9416^2}\right) \times 10{,}569.233$$
$$= 11{,}920.942$$

Note that equations (6) and (7) are also consistent with the computed results given in Table 4 of Part I since,

$$S\hat{S}E = (1 - r^2)\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 = (1 - 0.9416^2) \times 10.569.233 = 1{,}198.439$$

and

$$S\hat{S}R = r^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = (0.9416^2) \times 10{,}569.233 = 9{,}370.794$$

again, aside from slight round-off error.

Corollary 2: Provided $r \neq 0$, the residual sum of squares, $\mathrm{S\hat{S}E} = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$, from the Least Squares Regression of **x** on **y** and the residual sum of squares, $\mathrm{S\tilde{S}E} = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2$, from the Inverse Regression of **y** on **x** satisfy

$$S\tilde{S}E - S\hat{S}E = \left[\frac{1}{r} - r\right]^2 \sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \geq 0,$$

with equality prevailing if and only if $r = \pm 1$.

Proof: Using Corollary 1, equations (6) and expanding gives

$$S\tilde{S}E - S\hat{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 - \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

$$= \left(\frac{1}{r^2} - 1\right)\sum_{i=1}^{n}(x_i - \bar{x})^2 - (1 - r^2)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \left[\left(\frac{1}{r^2} - 1\right) - (1 - r^2)\right]\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left[r^2 + \frac{1}{r^2} - 2\right]\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \left[\frac{(1 - r^2)^2}{r^2}\right]\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left[\frac{1}{r} - r\right]^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0,$$

and it is simple to show that $\left[\frac{1}{r} - r\right]^2 = 0 \Leftrightarrow r = \pm 1.$ □

Example 2: Using Corollary 2 and the numerical results of Example 1, the absolute difference in these error sums of squares for the 32-city low and high temperature data in Part I is

$$\left|S\tilde{S}E - S\hat{S}E\right| = \left[\frac{1}{r} - r\right]^2\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2 = \left[\frac{1}{0.9416} - 0.9416\right]^2 \times 10{,}569.233$$

$$= 153.270.$$

Corollary 3: *The percentage error,* PE_{SSE}, *, in* $S\tilde{S}E = \sum_{i=1}^{n}(x_i - \tilde{x}_i)^2$ *over* $S\hat{S}E = \sum_{i=1}^{n}(x_i - \hat{x}_i)^2$*is*

$$PE_{SSE} = 100 \times \left[\frac{1}{r^2} - 1\right] \geq 0,$$

provided $r \neq 0$, *with equality prevailing if and only if* $r = \pm 1$.

Proof: Using the two previous corollaries,

$$100\% \times \left|\frac{S\tilde{S}E - S\hat{S}E}{S\hat{S}E}\right| = 100\% \times \left|\frac{\left[\frac{1}{r} - r\right]^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{(1 - r^2)\sum_{i=1}^{n}(x_i - \bar{x})^2}\right| = 100\% \times \left|\frac{\left[\frac{1}{r} - r\right]^2}{(1 - r^2)}\right|$$

$$= 100\% \times \left|\frac{\frac{1}{r^2}(1 - r^2)^2}{(1 - r^2)}\right| = 100\% \times (\frac{1}{r^2} - 1) \geq 0$$

with equality clearly prevailing if and only if $r = \pm 1$. □

Example 3: Using Corollary 3 with $r = 0.9416$,

$$PE_{SSE} = 100\% \times \left[\frac{1}{r^2} - 1\right] = 100\% \times \left[\frac{1}{0.9416^2} - 1\right] = 100\% \times (0.12789)$$
$$= 12.789\%.$$

Also, direct computation from Table 4 of Part 1 gives,

$$100\% \times \left|\frac{S\tilde{S}E - S\hat{S}E}{S\hat{S}E}\right| = 100\% \times \left|\frac{1{,}352.181 - 1{,}198.904}{1{,}198.904}\right| = 12.785\%,$$

with the results agreeing aside from slight round-off error.

Corollary 4: Suppose $\mathrm{r} \neq 0$. If $\hat{\mathrm{S}}_{\mathrm{e}} = \sqrt{\frac{\mathrm{S\hat{S}E}}{(\mathrm{n}-1)}}$ denotes the sum of squares about the regression line, from the Least Squares Regression of **x** on **y**, and $\tilde{\mathrm{S}}_{\mathrm{e}} = \sqrt{\frac{\mathrm{S\tilde{S}E}}{(\mathrm{n}-1)}}$ denotes the sum of squares about the regression line, from the Inverse Regression of **y** on **x**, then the percentage error in the difference between these quantities, $\mathrm{PE}_{\mathrm{regression}}$, is

$$\mathrm{PE}_{\mathrm{regression}} = 100\% \times \left[\frac{\tilde{\mathrm{S}}_{\mathrm{e}} - \hat{\mathrm{S}}_{\mathrm{e}}}{\hat{\mathrm{S}}_{\mathrm{e}}}\right] = 100\% \left[\frac{1}{|\mathrm{r}|} - 1\right] \geq 0,$$

with equality prevailing if and only if $\mathrm{r} = \pm 1$.

Proof: Using the previous corollaries,

$$PE_{regression} = 100\% \times \left[\frac{\tilde{S}_e - \hat{S}_e}{\hat{S}_e}\right] = 100\% \times \left[\frac{\sqrt{\frac{S\tilde{S}E}{(n-1)}} - \sqrt{\frac{S\hat{S}E}{(n-1)}}}{\sqrt{\frac{S\hat{S}E}{(n-1)}}}\right]$$

$$= 100\% \times \left[\frac{\sqrt{S\tilde{S}E}}{\sqrt{S\hat{S}E}} - 1\right]$$

$$= 100\% \times \left[\sqrt{\frac{\left(\frac{1}{r^2}-1\right)\sum_{i=1}^{n}(x_i-\bar{x})^2}{(1-r^2)\sum_{i=1}^{n}(x_i-\bar{x})^2}} - 1\right] = 100\% \times \left[\sqrt{\frac{\left(\frac{1}{r^2}-1\right)}{(1-r^2)}} - 1\right]$$

$$= 100\% \times \left[\frac{1}{|r|} - 1\right] \geq 0$$

with equality clearly prevailing if and only if $r = \pm 1$. □

<u>Example</u> 4:

Using Corollary 4 with $r = 0.94158$,

$$PE_{regression} = 100\% \left[\frac{1}{|0.9416|} - 1\right] = 100\% \times 0.062 = 6.20\%$$

Also, direct computation from Table 4 of Part 1 gives,

$$PE_{regression} = 100\% \times \left[\frac{6.714 - 6.322}{6.322}\right] = 100\% \times 0.0621 = 6.21\%$$

with the results agreeing aside from slight round-off error. □

It is fascinating to see the singularly important role played by the sample Pearson Correlation Coefficient in Corollaries (1) through (4). Interestingly, Table 1 shows precisely how the percent errors proven in Corollaries 3 and 4 vary as a function of selected values of the sample Pearson Correlation Coefficient, **r**.

When comparing Inverse Regression of **y** on **x** with Least Squares Regression of **x** on **y**, the percent error in the Standard Error about the Regression Line remains below 11.11% as long as the sample Pearson Correlation Coefficient $r \geq 0.90$. For positive values of **r** much smaller than 0.90, Inverse Regression of **y** on **x** is quite inferior to Least Squares Regression of **x** on **y**.

Table 1. **Percent Error in Residual Sums of Squares and Percent Error in the Standard Error about the Regression Line for Inverse Regression over Least Squares Regression**

Sample Pearson Correlation Coefficient, r	*Percent Error in Residual Sums of Squares*	*Percent Error in the Standard Error about the Regression Line*
1.00	0.00	0.00
0.98	4.12	2.04
0.96	8.51	4.17
0.94	13.17	6.38
0.92	18.15	8.70
0.90	23.46	11.11
0.86	35.21	16.28
0.84	41.72	19.05
0.82	48.72	21.95
0.80	56.25	25.00

Although these corollaries provide global comparisons between Inverse Regression and Least Squares "x on y" Regression, Corollary 5 targets the absolute difference in individual estimates.

Corollary 5: Let $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots\ldots\ldots, (x_n, y_n)$ be a collection of data points which have been used to construct Inverse Regression and Least Squares "x on y" Regression estimates, say $\tilde{x}^*$ and $\tilde{x}^*$, which are both estimated at y^*, then

(a) $|\tilde{x}^* - \hat{x}^*| = \left|\frac{1}{r} - r\right| \times \left(\frac{S_x}{S_y}\right) |y^* - \bar{y}|$

(b) $\lim\limits_{r \to \pm 1} |\tilde{x}^* - \hat{x}^*| = 0$

(c) $|z_{\tilde{x}^*} - z_{\hat{x}^*}| = \left|\frac{1}{r} - r\right| \times z_{y^*}$

where the summary statistics $\bar{y}, S_x, S_y$ and r are as previously defined, and $z_{\tilde{x}^*} = \frac{(\tilde{x}^* - \bar{x})}{S_x}$, $z_{\hat{x}^*} = \frac{(\hat{x}^* - \bar{x})}{S_x}$ and $z_{y^*} = \frac{(y^* - \bar{y})}{S_y}$ are the sample Z (standard) scores associated with **x** and **y** data, respectively.

Proof: For (a), use equations (2), (3) and (4) to get

$$|\tilde{x}^* - \hat{x}^*| = |(\tilde{x}^* - \bar{x}) - (\hat{x}^* - \bar{x})| = |(\tilde{x}^* - \bar{x}) - r^2(\tilde{x}^* - \bar{x})|$$

$$= |(1 - r^2)(\tilde{x}^* - \bar{x})| = \left|\left[\frac{(1 - r^2)}{m_{y|x}}\right](y^* - \bar{y})\right|$$

$$= \left|\left[\frac{(1 - r^2)S_x}{rS_y}\right](y^* - \bar{y})\right| = \left|\frac{1}{r} - r\right| \times \left(\frac{S_x}{S_y}\right) \times |y^* - \bar{y}|$$

For (b), note that

$$\lim_{r\to\pm 1}|\tilde{x}^* - \hat{x}^*| = \lim_{r\to\pm 1}\left|\frac{1}{r} - r\right| \times \left(\frac{S_x}{S_y}\right) \times |y^* - \bar{y}| = \left(\frac{S_x}{S_y}\right) \times |y^* - \bar{y}| \times \lim_{r\to\pm 1}\left|\frac{1}{r} - r\right|$$

$$= 0$$

□

To show (c), algebraically rewrite (a) as follows,

$$|\tilde{x}^* - \hat{x}^*| = \left|\frac{1}{r} - r\right| \times \left(\frac{S_x}{S_y}\right) \times |y^* - \bar{y}| \;\Leftrightarrow\; |(\tilde{x}^* - \bar{x}) - (\hat{x}^* - \bar{x})| = \left|\frac{1}{r} - r\right| \times \left(\frac{S_x}{S_y}\right) \times$$

$$|y^* - \bar{y}|$$

$$\Leftrightarrow\; \left|\frac{(\tilde{x}^* - \bar{x})}{S_x} - \frac{(\hat{x}^* - \bar{x})}{S_x}\right| = \left|\frac{1}{r} - r\right| \times \left(\left|\frac{(y^* - \bar{y})}{S_y}\right|\right) \;\Leftrightarrow\; |z_{\tilde{x}^*} - z_{\hat{x}^*}| = \left|\frac{1}{r} - r\right| \times |z_{y^*}|.$$

Numerical verification of Corollary 5 for the city of New Orleans is shown in Example 5.

Example 5: Table 4 of Part I shows the actual low and high temperatures on a given day in the U.S.A. to be 46 °F and 63 °F, respectively, for New Orleans, LA. The Least Squares "Low on High" estimate is seen there to be $\hat{x} = 50.129$°F, while the Inverse Regression estimate is $\tilde{x} = 50.999$°F. So, the absolute difference is

$$|\tilde{x} - \hat{x}| = |50.999 - 50.129| = 0.870 \tag{11}$$

Yet, recall that the summary statistics Table 4 of Part I are

$$S^2_{Low} = 340.943, S^2_{High} = 294.144, \bar{y} = \overline{High} = 56.28 \text{ and } r = 0.9416,$$

and inputting these statistics into the right-hand side of result (a) of Corollary 5 gives

$$|\tilde{x} - \hat{x}| = \left|\frac{1}{0.9416} - 0.9416\right| \times \sqrt{\frac{340.943}{294.144}}|63 - 56.28| = 0.871, \tag{12}$$

which agrees with equation (11) aside from some slight round-off error.

Conclusions

Parts I and II of this companion series of articles address comparison of Inverse Regression of **y** on **x** with Least Squares Regression of **x** on **y**. Although Inverse Regression of **y** on **x** is inferior to Least Squares Regression of **x** on **y** in a "squared error" sense, there are instances, depending on the absolute value of the magnitude of the sample Pearson Correlation Coefficient, $|r|$, when the two methods yield comparable estimates for the calibration problem. Part I of this bipartite series is primarily concerned with a description on the calibration problem as the genesis for Inverse Regression, its development, and rudimentary comparisons with low and high temperature data for thirty-two American cities. On the other hand, Part II presents some theoretical comparisons that culminate in a collection of five curious corollaries suitable for an undergraduate research project. The results of the corollaries are accessible to students who have studied applied Least Squares Regression, while their proofs can be broached with mathematically more sophisticated students who are familiar with Least Squares Regression. In tandem, these articles offer many opportunities for instructors and students to delve deeper into the topic of Regression.

Funding: This research received neither internal nor external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Disclaimer Statement: This is original research unaffiliated with other work.

Author Biographies

Di Gao, Ph.D. Assistant Professor of Statistics, Department of Mathematics and Statistics in the Department of Mathematics and Statistics, Sam Houston State University, Huntsville, TX, USA. His areas of interest include Bayesian Inference, Biostatistics, Dimension Reduction, Statistical Learning with Sparsity, Statistical Consulting.

Stephen M. Scariano, Ph.D. Professor of Mathematics and Statistics, Department of Mathematics and Statistics, Sam Houston State University, Huntsville, TX, USA. His areas of interest include Quality Control, Time Series, Regression Analysis, Design of Experiments, Multivariate Analysis, Statistics Education.

Authorship and Level of Contribution: The authors equally contributed in the research, writing and further revised it for consideration for publishing.

References

Gao, D., & Scariano, S. M. (2021). The Benefits of Teaching Inverse Regression Alongside Least Squares Regression: Graphical and Numerical Comparisons, *Research Journal in Advanced Sciences*, 2(1), https://royalliteglobal.com/rjas/article/view/457.

Harter, W. L. (1974). The Method of Least Squares and Some Alternatives: Part I, 42(2), 147-174.

Peck, R. (2015). Statistics: Learning from Data, Cengage Learning, Stamford, CT.

Peck, R., Olsen, C., & Devore, J. (2015). Statistics and Data Analysis, 5th ed. Cengage Learning, Boston, MA.

Weiss, N. (2012). Elementary Statistics, 8th ed. Addison-Wesley, Boston, MA

Stigler, S. M. (1981). Gauss and the Invention of Least Squares. *Annals of Statistics*, 9(2), 465-474. https://projecteuclid.org/download/pdf_1/euclid.aos/1176345451