Research Journal in Advanced Sciences
ISSN: 2709-281X (Online) ISSN: 2709-2801 (Print)

Advanced Sciences

The benefits of teaching Inverse Regression alongside Least Squares Regression: Graphical and numerical comparisons

Review article

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

Article History:
Received: 2020-10-29
Revised: 2020-11-12
Accepted: 2020-12-09
Published: 2021-01-01

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: Graphical and numerical comparisons. *Research Journal in Advanced Sciences*, 2(1). Retrieved from https://royalliteglobal.com/rjas/article/view/457

Published in Nairobi, Kenya by Royallite Global in the **Research Journal in Advanced Sciences**

Abstract

A good, college-level first course in descriptive and inferential statistics must include the topic of Least Squares Regression. Indeed, most modern statistics textbooks discuss linear regression as soon as the concept of statistical correlation is understood. This article demonstrates how to enhance a general discussion of Least Squares Regression with the concept of Inverse Regression. In Part I of this bipartite exposition, foundations are laid, and graphical and numerical comparisons are developed. Deeper relationships are then explored in Part II of this series.

Scan & Read

Research Journal in Advanced Sciences
ISSN: 2709-281X (Online) ISSN: 2709-2801 (Print)

Advanced Sciences

Public Interest Statement

The study contrasts 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

The central idea behind Least Squares Regression of "**y** (dependent variable) on **x** (independent variable)" is to identify, in context, a "best-fitting" line associated with a plausible linear relationship between the coordinates of several non-collinear points: $(x_1, y_1), (x_2, y_2)\ldots\ldots\ldots\ldots, (x_n, y_n)$. The end result, attributed primarily to Johann Carl Frederick Gauss (1777-1855), is that the "best-fitting" line approximating this relationship is of the form $y = m^* x + b^*$, where m^* and b^* are functions only of the actual data values obtained. Stigler (1981) provides an excellent historical presentation, along with the many controversies surrounding the origins of Least Squares Regression. Least Squares Regression has been successfully used in a variety of fields of study and carefully taught for more than one hundred years. Peck (2015), Peck, Olsen and Devore (2015) and Weiss (2012) provide excellent, modern introductions to the topic of Least Squares Regression that are accessible to general undergraduate audiences. However, the companion application of Inverse Regression, which is sometimes also called the calibration problem, has received far less attention in collegiate curricula, and it is exposition of this concept that is of primary interest here. Consider the following two examples where Inverse Regression could be valuable:

Example 1:

It is known that a certain drug is successful in lowering LDL blood cholesterol levels and that the number of units **Y** that the LDL cholesterol reading is reduced by is a linear function of the quantity of drug, say **x**, administered in a given time interval. Suppose over a given time period, **n** patients suffering from this condition are monitored and treated at different levels of x_i. The observations are assumed to fit the simple, linear statistical model $Y_i = mx + b + \varepsilon_i$ for $i = 1, 2, \ldots, n$. However, in formulating a treatment regimen, a physician typically measures a patient's LDL cholesterol level and determines that it should be reduced by, say y_0, units. The main question is how many units of the drug, say x_0, should be given? That is, we want to estimate the drug quantity, x_0, that will reduce the patient's LDL level to roughly y_0 units. In a sense, this is an "inverse" type of statistical problem, where a value for an independent variable is sought.

Example 2:

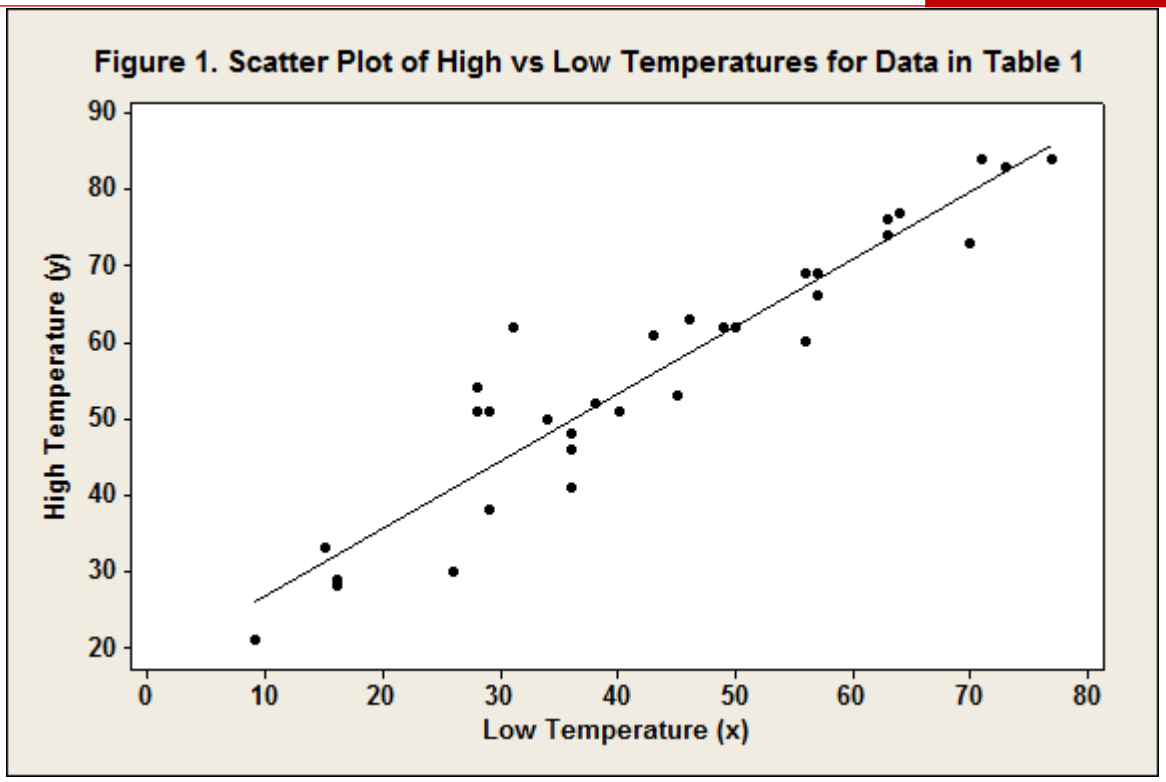
Suppose the linear statistical model $Y_i = a_1 t + a_0 + \varepsilon_i \quad (i = 1, 2, ..., n)$ governs the relationship between the weight **Y,** (in pounds) of a given breed of turkeys and time **t** (in weeks since birth, $0 \le t \le 15$), while being nourished on a certain feed supply. Here, a_0 is regarded as the typical turkey weight at birth and a_1 denotes the weekly weight growth rate. If a turkey processor receives a rafter of young turkeys that have been fed this particular feed supply and records their weights, it is important to estimate the (usually) unknown number of weeks, say $\mathbf{t_0}$, that this rafter has been on the feeding program, before additional feeding and processing is initiated. Again, this is an "inverse" type of statistical problem, where a value for an independent variable is sought. In each of the examples just presented, the response variable, **Y,** is easy to compute once data have been collected. However, primary interest focuses on estimating the corresponding value of the explanatory variable, **X.** This problem is distinctly different from the usual Least Squares Regression set-up, and it is the one explored in the article.

2. Brief Review of Least Squares Regression with Low and High Temperature Data

Table 1 gives the low and high temperatures (in degrees Farenheit) for thirty-two U.S. cities on a given winter day. Figure 1 is the associated scatter plot with the accompanying Least Squares regression line of high temperature on low temperature.

Table 1. High and Low Temperatures for 32 American cities on Christmas Day of 2015

City, State	High Temp	Low Temp	City, State	High Temp	Low Temp
Amarillo, TX	51	29	Miami, FL	84	77
Atlanta, GA	76	63	Milwaukee, WI	38	29
Bangor, ME	54	28	Minneapolis, MN	30	26
Billings, MT	21	9	Mobile, AL	73	70
Birmingham, AL	77	64	Montpelier, VT	51	28
Boston, MA	62	49	Nashville, TN	69	56
Buffalo, NY	48	36	New Orleans, LA	63	46
Charlotte, NC	74	63	New York, NY	66	57
Chicago, IL	46	36	Phoenix, AZ	61	43
Cincinnati, OH	53	45	Pittsburgh, PA	62	50
Concord, NH	62	31	Salt Lake City, UT	29	16
Denver, CO	28	16	San Diego, CA	60	56
Detroit, MI	50	34	San Francisco	51	40
Houston, TX	83	73	Seattle, WA	41	36
Jacksonville, FL	84	71	St. Louis, MO	52	38
Lincoln, NE	33	15	Washington, DC	69	57



In this context, "best-fitting " is taken to mean that the total (vertical) deviations from the given data points to the regression line be minimized in the squared-error sense. See Peck (2015) and Peck, et al., (2015). More precisely, the "best-fitting" line, say $y = m_{y|x}x + b_{y|x}$, in the Least Squares sense is the unique line having slope $m_{y|x}$ and intercept $b_{y|x}$ satisfying

$$\sum_{i=1}^{n}\left(y_i - (m_{y|x}x_i + b_{y|x})\right)^2 \le \sum_{i=1}^{n}\left(y_i - (\boldsymbol{m}x_i + \boldsymbol{b})\right)^2 \qquad (1)$$

where

$$m_{y|x} = r_{xy}\left(\frac{S_y}{S_x}\right) \qquad \text{and} \qquad b_{y|x} = \bar{y} - m_{y|x}\bar{x}, \qquad (2)$$

with

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

$$\text{and} \quad r_{xy} = \frac{S_{xy}}{S_x \times S_y}. \qquad (3)$$

for all real numbers **m** and **b**. Here, r_{xy} is the sample Pearson Correlation Coefficient (Weiss, 2012) associated with the ordered pairs $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots\ldots, (x_n, y_n)$. Denoting

the high temperatures as the "**y**-data", or response data, and the low temperatures as the "**x**-data", or explanatory data, equations (3) provide, $\bar{x} = 43.34^\circ\text{F},\ \bar{y} = 56.28^\circ\text{F},\ S_x^2 = 340.943\left(^\circ\text{F}\right)^2,\ S_y^2 = 296.144\left(^\circ\text{F}\right)^2,\ S_{xy} = 299.191\left(^\circ\text{F}\right)^2,$ and

$$r_{xy} = \frac{S_{xy}}{S_x \times S_y} = \frac{299.191}{\sqrt{340.943 \times 296.144}} = 0.9416 \tag{4}$$

where the bivariate sample size is $n = 32$. Substituting the statistics in equations (4) into equations (3) gives

$$m_{y|x} = r_{xy}\left(\frac{S_y}{S_x}\right) = (0.9416)\sqrt{\left(\frac{296.144}{340.943}\right)} = 0.8776 \quad \text{and}$$

$$b_{y|x} = \bar{y} - m_{y|x}\bar{x} = 56.28^\circ\text{F} - 0.8776 \times 43.34^\circ\text{F} = 18.2448^\circ\text{F}. \tag{5}$$

Equations (5) permits writing the Least Squares Regression of high temperature on low temperature as

$$\hat{y}^\circ\text{F} = (0.878)x^\circ\text{F} + 18.245^\circ\text{F}$$

or

$$predicted\ high\ temperature^\circ\text{F} = (0.878) \times (actual\ low\ temperature)^\circ\text{F} + 18.245^\circ\text{F} \tag{6}$$

using $\mathbf{10^{-3}}$ precision. The Least Squares regression line depicted in Figure 1 is specified by equation (6),

Here, we intentionally use the notation $m_{y|x}$ and $b_{y|x}$ to denote the slope and intercept of the Least Squares Regression line for two reasons: (i) to explicitly demonstrate the dependence of the response variable, y, on the predictor variable, x, and (ii) to avoid confusion when discussing the slope, $m_{x|y}$, intercept, $b_{x|y}$, of the Least Squares Regression of "**x** on **y**", which will be considered later. Using equation (6), Table 2 gives the Least Squares predicted high temperatures for each city identified in Table 1. The "Residual" column gives the estimation error, defined by

$$\text{Residual} = Actual\ High\ Temperature - Least\ Squares\ Predicted\ High\ Temperature. \tag{7}$$

Each residual provides both the magnitude and direction of the estimation error for that particular city. A positive residual is associated with under-prediction of the actual high temperature while a negative residual is affiliated with an over-prediction of the actual high temperature.

Table 2. Least Squares Predicted High Temperatures with Residuals using Table 1 data.

City, State	High	Low	Predicted High	Residual	City, State	High	Low	Predicted High	Residual
Amarillo, TX	51	29	43.707	7.293	Miami, FL	84	77	85.851	-1.851
Atlanta, GA	76	63	73.559	2.441	Milwaukee, WI	38	29	43.707	-5.707
Bangor, ME	54	28	42.829	11.171	Minneapolis, MN	30	26	41.073	-11.073
Billings, MT	21	9	26.147	-5.147	Mobile, AL	73	70	79.705	-6.705
Birmingham, AL	77	64	74.437	2.563	Montpelier, VT	51	28	42.829	8.171
Boston, MA	62	49	61.267	0.733	Nashville, TN	69	56	67.413	1.587
Buffalo, NY	48	36	49.853	-1.853	New Orleans, LA	63	46	58.633	4.367
Charlotte, NC	74	63	73.559	0.441	New York, NY	66	57	68.291	-2.291
Chicago, IL	46	36	49.853	-3.853	Phoenix, AZ	61	43	55.999	5.001
Cincinnati, OH	53	45	57.755	-4.755	Pittsburgh, PA	62	50	62.145	-0.145
Concord, NH	62	31	45.463	16.537	Salt Lake City, UT	29	16	32.293	-3.293
Denver, CO	28	16	32.293	-4.293	San Diego, CA	60	56	67.413	-7.413
Detroit, MI	50	34	48.097	1.903	San Francisco	51	40	53.365	-2.365
Houston, TX	83	73	82.339	0.661	Seattle, WA	41	36	49.853	-8.853
Jacksonville, FL	84	71	80.583	3.417	St. Louis, MO	52	38	51.609	0.391
Lincoln, NE	33	15	31.415	1.585	Washington, DC	69	57	68.291	0.709

The aggregate estimation error is usually called the Standard Error about the Regression, denoted by S_{REG} and defined by

$$S_{REG} = \sqrt{\frac{\sum_{i=1}^{n} \left(\text{Residual}_i\right)^2}{(n-2)}} \tag{8}$$

The data in Table 2 yield

$$S_{REG} = \sqrt{\frac{\sum_{i=1}^{32} \left(\text{Residual}_i\right)^2}{(n-2)}} = \sqrt{\frac{1,041.387}{30}} = 5.892^{\circ}\,\text{F},$$

which can be interpreted as the "typical" (vertical) deviation of an actual data point from the regression line. The square of the sample Pearson Correlation Coefficient, r_{xy}^2, called the sample Coefficient of Determination, is oftentimes used as a "goodness of fit" measure for Least Squares regression; here $r_{xy}^2 = (0.9416)^2 = 88.66\%$, suggesting that approximately 89% of the variation observed in the actual high temperatures for these thirty-two cities can be explained by the Least Squares Regression model of high temperature on low temperature.

2. Inverse Regression with Low and High Temperature Data

Let us now revisit the Low and High Temperature Data in Table 1 from a slightly different perspective. That is, given a value for the high temperature, invert the Least Squares Regression equation to estimate the low temperature that gave rise to that particular city's high temperature. For example, New Orleans, LA had an actual high temperature reading of 63°F and an actual low temperature reading of 46°F on Christmas Day of 2015. Inputting this high temperature directly into equation (6) and inverting produces:

$$63^\circ\,\mathrm{F} = (0.878)\hat{x}^\circ\,\mathrm{F} + 18.245^\circ\,\mathrm{F} \;\Rightarrow\; \hat{x}^\circ\mathrm{F} = \frac{(63^\circ\,\mathrm{F} - 18.245^\circ\mathrm{F})}{0.878} = 50.974^\circ\mathrm{F} \approx 51^\circ\mathrm{F} \qquad (9)$$

using integer precision. So, given a high temperature, it is a simple matter to **invert** the Least Squares Regression equation to produce an estimate of a city's corresponding low temperature on that day. Table 3 gives the **Inverse Regression** estimated low temperatures for all thirty-two American cities shown in Table1, along with their corresponding residuals.

Table 3. Inverse Least Squares Predicted LowTemperatures with Residuals using Table 1 data.

City, State	High	Low	Predicted Low	Residual	City, State	High	Low	Predicted Low	Residual
Amarillo, TX	51	29	37.325	-8.325	Miami, FL	84	77	74.929	2.071
Atlanta, GA	76	63	65.813	-2.813	Milwaukee, WI	38	29	22.512	6.488
Bangor, ME	54	28	40.744	-12.744	Minneapolis, MN	30	26	13.396	12.604
Billings, MT	21	9	3.140	5.860	Mobile, AL	73	70	62.394	7.606
Birmingham, AL	77	64	66.952	-2.952	Montpelier, VT	51	28	37.325	-9.325
Boston, MA	62	49	49.860	-0.860	Nashville, TN	69	56	57.836	-1.836
Buffalo, NY	48	36	33.907	2.093	New Orleans, LA	63	46	50.999	-4.999
Charlotte, NC	74	63	63.534	-0.534	New York, NY	66	57	54.418	2.582
Chicago, IL	46	36	31.628	4.372	Phoenix, AZ	61	43	48.720	-5.720
Cincinnati, OH	53	45	39.604	5.396	Pittsburgh, PA	62	50	49.860	0.140
Concord, NH	62	31	49.860	-18.860	Salt Lake City, UT	29	16	12.256	3.744
Denver, CO	28	16	11.117	4.883	San Diego, CA	60	56	47.581	8.419
Detroit, MI	50	34	36.186	-2.186	San Francisco	51	40	37.325	2.675
Houston, TX	83	73	73.789	-0.789	Seattle, WA	41	36	25.930	10.070
Jacksonville, FL	84	71	74.929	-3.929	St. Louis, MO	52	38	38.465	-0.465
Lincoln, NE	33	15	16.814	-1.814	Washington, DC	69	57	57.836	-0.836

Although Inverse Regression is a natural way to estimate an abscissa value using an ordinate value and inverting a previously obtained Least Squares Regression equation $y = m_{y|x}x + b_{y|x}$, it must be clearly understood that such estimation is not equivalent to Least Squares Regression of "**x** on **y**", where the roles of **x** and **y** are interchanged in equation (1). Nonetheless, Inverse Regression should be adequate in many practical situations, at least for a quick estimate of an abscissa.

To formalize the concept of Inverse Regression, assume that a reasonable linear trend exists between a response variable **y** and a predictor variable **x**, and Least Squares

Regression of "**y** on **x**" has already been quantified as $y = m_{y|x}x + b_{y|x}$, using equations (2).

Inverting produces

$$x = \frac{y - b_{y|x}}{m_{y|x}} = \left(\frac{1}{m_{y|x}}\right)y + \left(\frac{-b_{y|x}}{m_{y|x}}\right), \qquad (10)$$

so that the Inverse Regression estimating equation is

$$\hat{x} = m_{Inv}y + b_{Inv}, \qquad (11)$$

where $m_{Inv} = \left(\frac{1}{m_{y|x}}\right) = \left(\frac{1}{r_{xy}\left(\frac{S_y}{S_x}\right)}\right) = \left(\frac{S_x}{r_{xy}S_y}\right)$ and

$$b_{Inv} = \left(\frac{-b_{y|x}}{m_{y|x}}\right) = \left(\frac{m_{y|x}\bar{x} - \bar{y}}{m_{y|x}}\right) = \bar{x} - \left(\frac{S_x}{r_{xy}S_y}\right)\bar{y} \qquad (12)$$

respectively, denote the Inverse Regression slope and intercept coefficients, provided

$$m_{y|x} \neq 0.$$

The results of equations (4), (5) and (6) substituted into equations (12) yield

$$m_{Inv} = \left(\frac{1}{m_{y|x}}\right) = \left(\frac{1}{0.8776}\right) = 1.1395 \text{ and } b_{Inv} = \left(\frac{-b_{y|x}}{m_{y|x}}\right) = \left(\frac{-18.2448}{0.8776}\right) = -20.7894 \qquad (13)$$

So, the Inverse Regression estimating equation for low temperature from high temperature is

$$\hat{x}^\circ \text{F} = (1.1395)y^\circ \text{F} - 20.7894^\circ\text{F}$$

or (14)

$$\text{predicted low temperature}^\circ \text{F} = (1.1395) \times (\text{actuall high temperature})^\circ \text{F} - 20.7894^\circ\text{F},$$

which is the equation used to compute the low temperature estimates, and their residuals, shown in Table 3. For New Orleans, LA, equation (14) gives

$$\hat{x}^\circ \text{F} = (1.1395)63^\circ \text{F} - 20.7894^\circ \text{F} = 50.9991^\circ \text{F} \approx 51^\circ \text{F}$$

which matches the result in equation (9), allowing for slight round-off error. The residual is

$$\text{Residual}_{New\ Orleans} = 46^\circ\text{F} - 50.9991^\circ\text{F} = -4.9991^\circ\text{F},$$

matching the corresponding entry in Table 3.

3. Comparing and Contrasting Inverse Regression with Least Squares Regression

For Least Squares Regression of "**x** on **y**", the roles of the response and predictor variables are interchanged, and this is another way to view the low and high temperature data in Table 1. That is, given data points $(y_1, x_1), (y_2, x_2) \ldots\ldots\ldots\ldots, (y_n, x_n)$, equations (1) – (3) become

$$\sum_{i=1}^{n}\left(x_i - (m_{x|y}y_i + b_{x|y})\right)^2 \le \sum_{i=1}^{n}\left(x_i - (\boldsymbol{m}y_i + \boldsymbol{b})\right)^2 \tag{15}$$

where

$$m_{x|y} = r_{yx}\left(\frac{S_x}{S_y}\right) \qquad \text{and} \qquad b_{x|y} = \bar{x} - m_{x|y}\bar{y}, \tag{16}$$

with

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i,\ \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i,\ \ S_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)},\ S_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)}\ S_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)},$$
$$\text{and}\ \ r_{yx} = \frac{S_{yx}}{S_y \times S_x}. \tag{17}$$

where $m_{x|y}$ and $b_{x|y}$ are the unique constants minimizing equation (15) for all real numbers **m** and **b**.

Since $S_{yx} = S_{xy}$, it follows that $r_{yx} = r_{xy}$. Using the results in equations (4),

$$m_{x|y} = (0.9416)\sqrt{\frac{340.943}{296.144}} = 1.0103 \qquad \text{and}$$
$$b_{x|y} = 43.34^\circ\text{F} - (1.0103 \times 56.28^\circ\text{F}) = -13.5197^\circ\text{F} \tag{18}$$

Equations (18) give rise to the Least Squares Regression equation of low temperature from high temperature as

$$\hat{x}^\circ\text{F} = (1.0103)y^\circ\text{F} - 13.5197^\circ\text{F}$$
$$\text{or}$$
$$predicted\ low\ temperature^\circ\text{F} = (1.0103) \times (actual\ high\ temperature)^\circ\text{F} - 13.5197^\circ\text{F} \tag{19}$$

using 10^{-3} precision.

For New Orleans, LA, equation (19) gives

$$\hat{x}^{\circ}\mathrm{F} = (1.0103)\times 63^{\circ}\mathrm{F} - 13.5197^{\circ}\mathrm{F} = 50.1292^{\circ}\mathrm{F} \approx 50^{\circ}\mathrm{F}.$$

with residual (20)

$$\text{Residual}_{New\ Orleans} = 46^{\circ}\mathrm{F} - 50.1292^{\circ}\mathrm{F} = -4.1292^{\circ}\mathrm{F},$$

For direct city-by-city comparison, Table 4 provides results for both the Inverse Regression and Least Squares Regression of "**x** (Low temperature) on **y** (High temperature)".

Table 4. Comparisons of Inverse Regression and Least Squares Regression of Low on High Temperature using Table 1 data.

City, State	High	Low	Inverse Regression Predicted Low	Inverse Regression Residual	Inverse Regression Squared Residual	LS x on y Regression Predicted Low	LS x on y Regression Residual	LS x on y Regression Squared Residual
Amarillo, TX	51	29	37.325	-8.325	69.307	38.006	-9.006	81.101
Atlanta, GA	76	63	65.813	-2.813	7.911	63.263	-0.263	0.069
Bangor, ME	54	28	40.744	-12.744	162.399	41.037	-13.037	169.950
Billings, MT	21	9	3.140	5.860	34.338	7.697	1.303	1.699
Birmingham, AL	77	64	66.952	-2.952	8.715	64.273	-0.273	0.075
Boston, MA	62	49	49.860	-0.860	0.739	49.119	-0.119	0.014
Buffalo, NY	48	36	33.907	2.093	4.382	34.975	1.025	1.051
Charlotte, NC	74	63	63.534	-0.534	0.285	61.243	1.758	3.089
Chicago, IL	46	36	31.628	4.372	19.118	32.954	3.046	9.278
Cincinnati, OH	53	45	39.604	5.396	29.116	40.026	4.974	24.739
Concord, NH	62	31	49.860	-18.860	355.685	49.119	-18.119	328.295
Denver, CO	28	16	11.117	4.883	23.848	14.769	1.231	1.516
Detroit, MI	50	34	36.186	-2.186	4.777	36.995	-2.995	8.972
Houston, TX	83	73	73.789	-0.789	0.623	70.335	2.665	7.101
Jacksonville, FL	84	71	74.929	-3.929	15.434	71.346	-0.346	0.119
Lincoln, NE	33	15	16.814	-1.814	3.291	19.820	-4.820	23.234
Miami, FL	84	77	74.929	2.071	4.291	71.346	5.655	31.973
Milwaukee, WI	38	29	22.512	6.488	42.099	24.872	4.128	17.043
Minneapolis, MN	30	26	13.396	12.604	158.871	16.789	9.211	84.837
Mobile, AL	73	70	62.394	7.606	57.850	60.232	9.768	95.410
Montpelier, VT	51	28	37.325	-9.325	86.957	38.006	-10.006	100.112
Nashville, TN	69	56	57.836	-1.836	3.371	56.191	-0.191	0.036
New Orleans, LA	63	46	50.999	-4.999	24.991	50.129	-4.129	17.050
New York, NY	66	57	54.418	2.582	6.669	53.160	3.840	14.745
Phoenix, AZ	61	43	48.720	-5.720	32.720	48.109	-5.109	26.098
Pittsburgh, PA	62	50	49.860	0.140	0.020	49.119	0.881	0.776
Salt Lake City, UT	29	16	12.256	3.744	14.017	15.779	0.221	0.049
San Diego, CA	60	56	47.581	8.419	70.886	47.098	8.902	79.240
San Francisco	51	40	37.325	2.675	7.155	38.006	1.994	3.978
Seattle, WA	41	36	25.930	10.070	101.403	27.903	8.097	65.568
St. Louis, MO	52	38	38.465	-0.465	0.216	39.016	-1.016	1.032
Washington, DC	69	57	57.836	-0.836	0.699	56.191	0.809	0.654
			Sum of Squared Residuls =		1352.181			1198.904
			Standard Error about the Regression Line =		6.714° F			6.322° F

Of course, it is not surprising that the estimated standard error about the regression line value of **6.332** °F for the Least Squares Regression of "**x** on **y**" is smaller than the estimated standard error about the regression line value of **6.714**° F for the Inverse

Regression of **"x on y"**. This is so precisely because the Least Squares criterion demands minimizing the sum of square residuals, which is not the criterion used for Inverse Regression. However, in this particular application the absolute difference $|6.332\ ^\circ F - 6.714\ ^\circ F| = .382\ ^\circ F < \frac{1}{2}\ ^\circ F$ is of little practical importance.

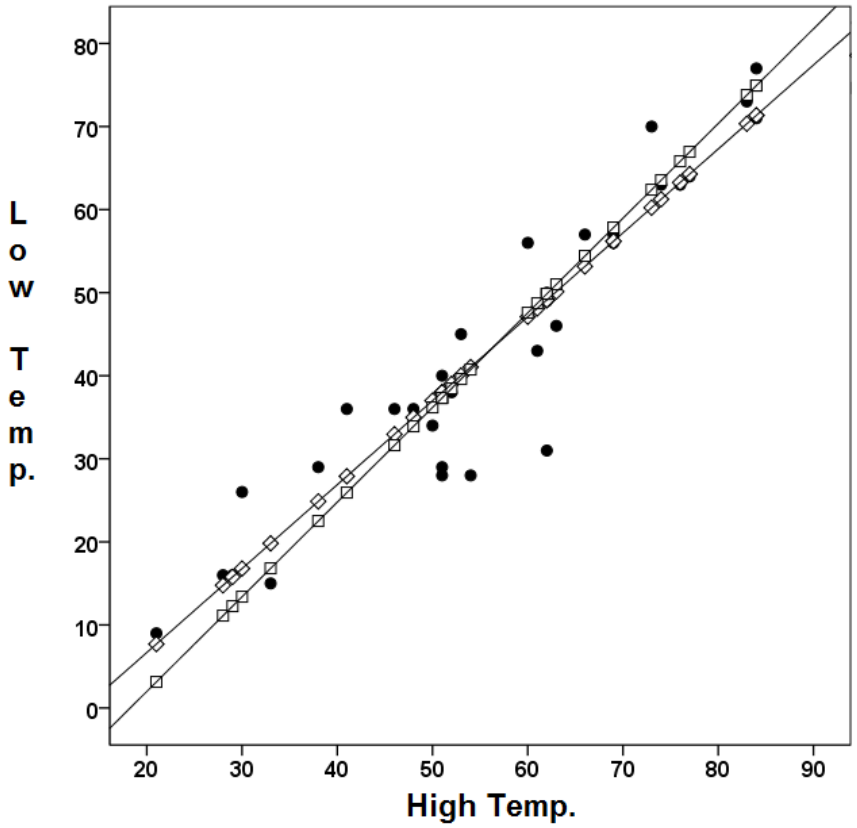
Table 5 gives a side-by-side comparison of the slopes and intercepts for two regression models,

Table 5. Comparison of Slopes and Intercepts of Estimating Equations

	Slope	Intercept
Inverse Regression of High Temperature on Low Temperature	1.1395	-20.7894
Least Squares Regression of Low Temperature on High Temperature	1.0103	-13.5197

while Figure 2 provides the corresponding overlay scatter plots, along with the data points. The Inverse

Figure 2. Inverse Regression and Least Squares Regression lines superimposed on (High, Low) temperature Scatter Plot.

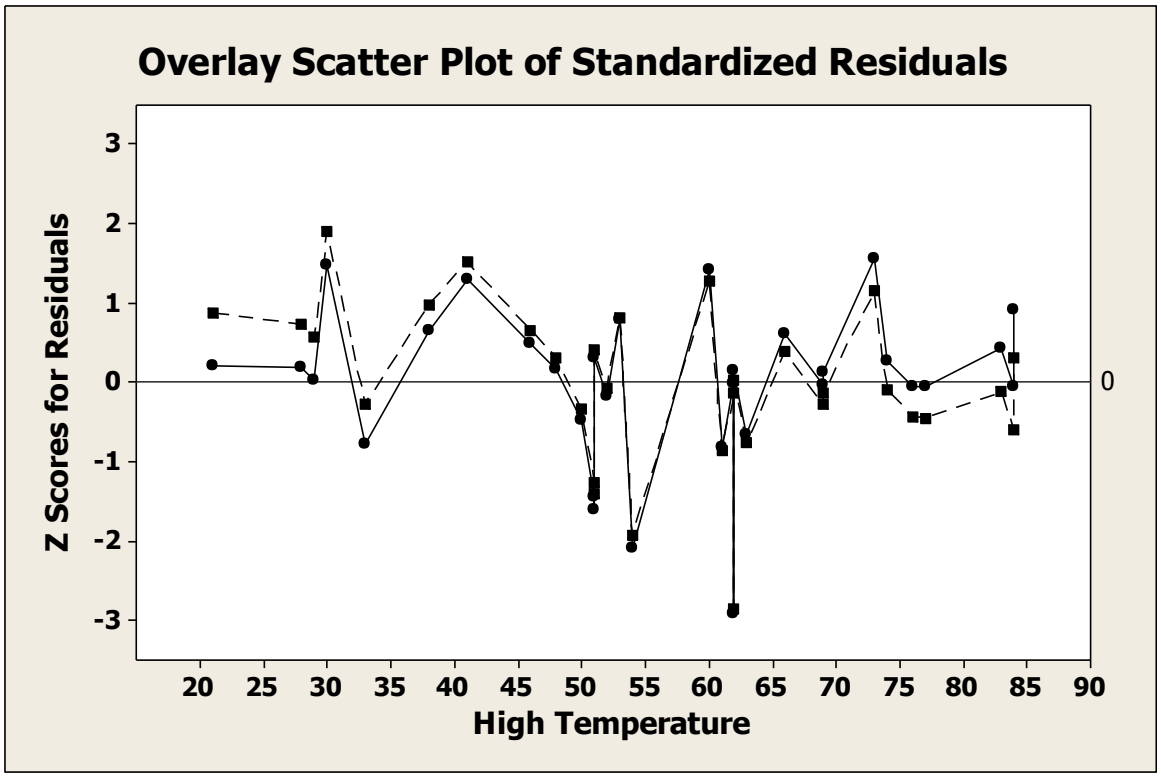


Data values (High, Low) are darkened circles,
Least Squares Regression Predictions of Low on High Temperature are open diamonds,
Inverse Regression Predictions of Low on High Temperature are open squares

Regression and Least Squares Regression of "**x** on **y**" give rise to different equations for estimating a city's low temperature reading based on its corresponding high temperature reading. A moment's reflection shows why. In general, two different criteria are used: Inverse Regression employs inverting the Least Squares Regression of "**y** on **x**", while Least Squares Regression of "**x** on **y**" is used directly with the data points, where the x and y roles are reversed. So, there should be no expectation that the methods would yield the same slopes and/or intercepts, and, in general, they do not.

Finally, Figure 3 presents the standardized, residual plots versus the low temperature readings for both models. The broken, polygonal line connecting solid squares represents the residual plot associated with the Inverse Regression model. The broken, polygonal line connecting solid circles represents the residual plot associated with the Least Squares Regression of low temperature on high temperature.

Figure 3. Residual Plots: ▪--▪ (Inverse Regression) •—• (Inverse Regression)



Tables 4 and 5 along with figures 2 and 3 suggest that there is good reason to believe that Inverse Regression may sometimes be successfully used in applications. However, delving deeper into that realm is the subject of Part II of our bipartite exposition.

4. Conclusion

The topic of Inverse Regression, a natural companion to Least Squares Regression, has been considered here. Two motivating examples where Inverse Regression is useful are first considered. Next, a brief review of Least Squares Regression using low and high temperatures for thirty-two American cities was presented. This data set was used as a foundation for comparing Inverse Regression with Least Squares Regression. It is seen that, in some instances, Inverse Regression is a close competitor of Least Squares regression when the original explanatory and response variables are interchanged. A deeper, comparative investigation is ventured in Part II of this series of articles.

Funding: This research received neither internal nor external funding

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

Author Biographies

Di Gao is an Assistant Professor of Statistics in the Department of Mathematics and Statistics, Sam Houston State University, USA. His areas of interest intersect between Bayesian Inference, Biostatistics, Dimension Reduction, Statistical Learning and Sparsity, Statistical Consulting.

Stephen M. Scariano is a Professor of Statistics in the Department of Mathematics and Statistics, Sam Houston State University, USA. His areas of research interest include: Quality Control, Time Series, Regression, Design of Experiments, Multivariate, Statistics Education.

Authorship and Level of Contribution: The authors equally contributed in the research, writing and preparation towards publishing. They both took part in the revision of till final publication.

References

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

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

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