Pivot Points in Bivariate Linear Regression

Carl V. Lutzer, David L. Farnsworth

School of Mathematical Sciences, Rochester Institute of Technology, Rochester, New York, USA
Email: cvlsma@rit.edu, dlfsma@rit,edu

Abstract

There are little-noticed points in the plane, which are artifacts of linear regression. The points, which are called pivot points, are the intersections of sets of regression lines. We derive the coordinates of the pivot point and explain its sources. We show how a pivot point arises in a certain notable data set, which has been analyzed often for points of high leverage. We obtain the application of pivot points that shortens calculations when updating a set of bivariate observations by adding a new point.

Keywords

Augmented Data Set, Bilinear Regression, Influence, Leverage, Pivot Point, Updating a Regression Line

1. Introduction

It is common to produce many lines to fit bivariate data as the observations are being altered in some way. For example, in order to determine a particular data point's influence on the best fit, the point may be moved by changing its y-coordinate and a new line created. Some diagnostic tests are based on this. A point, which is called the pivot point, is the intersection of certain lines that are often used for examining influence.

An example of a pivot point is presented in Section 2. In Section 3, we derive the coordinates of the pivot point. We show that a pivot point can be created in two ways. One way is augmenting an original set of bivariate observations with an additional point, which can have arbitrary multiplicity. Another way is altering an existing observation's y-coordinate as described above. Section 4 presents the benefit of the pivot point in that it can be useful to shorten calculations when adding a new observation.

2. Illustrative Example

Consider the data in Table 1 [1]. The predictor variable (x) is the age in months

at which a child says their first word, and the response variable (y) is the child's Gesell Adaptive Score from an aptitude test. These data have been analyzed many times for influential and outlying observations [2]-[7]. Using various criteria, Cases 2, 18, and 19 have been identified as significant. For illustrative purposes, we focus on Case 18.

When examining an individual observation's influence on a bivariate least-squares linear regression, it is common to generate a sequence of regression lines. These lines fit the same set of observations, except that the y-coordinate is made to vary while its x-coordinate is unchanged for the specified data point of interest. The influence of Case 18 on the least-squares regression line is examined by keeping its x-coordinate of 42 and giving its y-coordinate the values 57, 77, 97, 117, and 137. This produces the five regression lines in **Figure 1**. Clearly, Case 18 could have a large influence on the regression line. Some authors have illustrated and evaluated leverage in this way [8] [9] [10] [11]. All these regression lines pass though a common point, called the pivot point [12]. In **Figure 1**, the pivot point (12.3, 96.1) is shared by the five lines, and its location is indicated by the symbol Δ.

3. Derivation of the Pivot Point

We derive the formula for the coordinates of the pivot point. The pivot point can be created by augmenting an original set of bivariate observations with an additional point, which can have arbitrary multiplicity, which is another method to diagnose influence on the line [5] [13] [14] [15]. We show that formulation to be equivalent to varying the location of a single point, while keeping the same first coordinate, as is done in **Figure 1**.

Consider the bivariate data set $S_0 = \left\{ (x_i, y_i) : i = 1, 2, \cdots, n \right\}$. For simplicity, assume that coordinates are selected so that $\left(\sum x/n, \sum y/n \right) = (0, 0)$. Unindexed summations are over the elements of S_0. Define $V = \sum x^2/n$. Introduce m copies of the new point $R(u, v)$. If R is a point in S_0, these are additional copies. The aggregate of S_0 and $m > 0$ copies of R is denoted S_m.

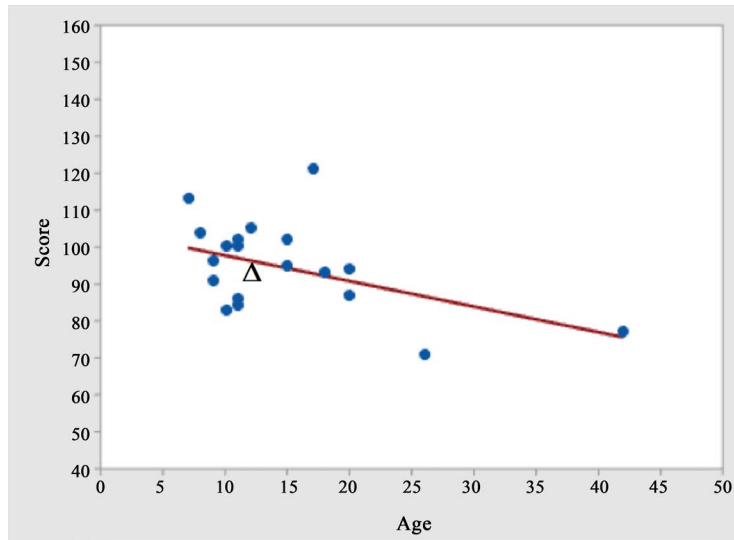For $m = 0$, the least-squares regression line of S_0 is

$$y = a_0 + b_0 x = \left(\sum xy / \sum x^2 \right) x.$$

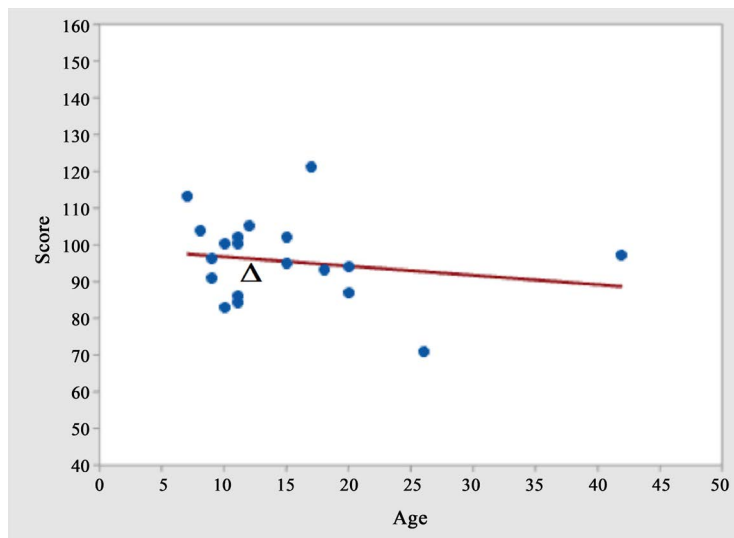**Table 1.** Age at First Word (x) and Gesell Adaptive Score (y).

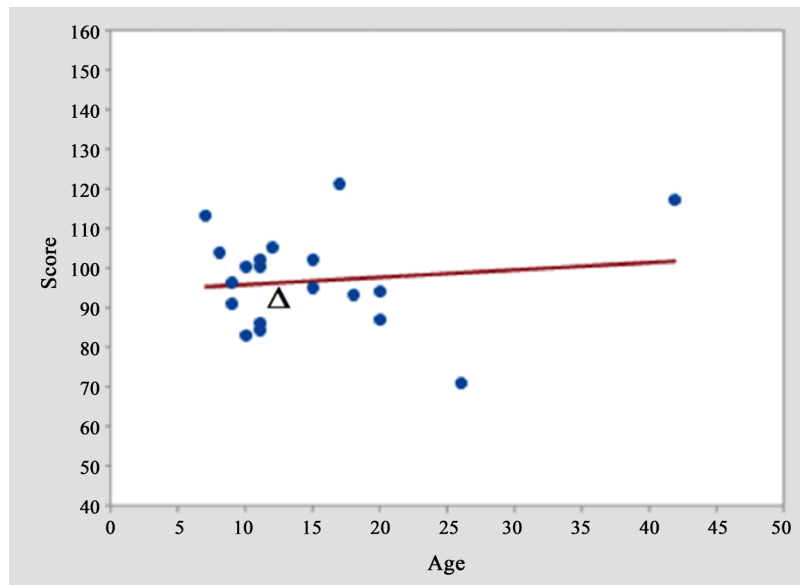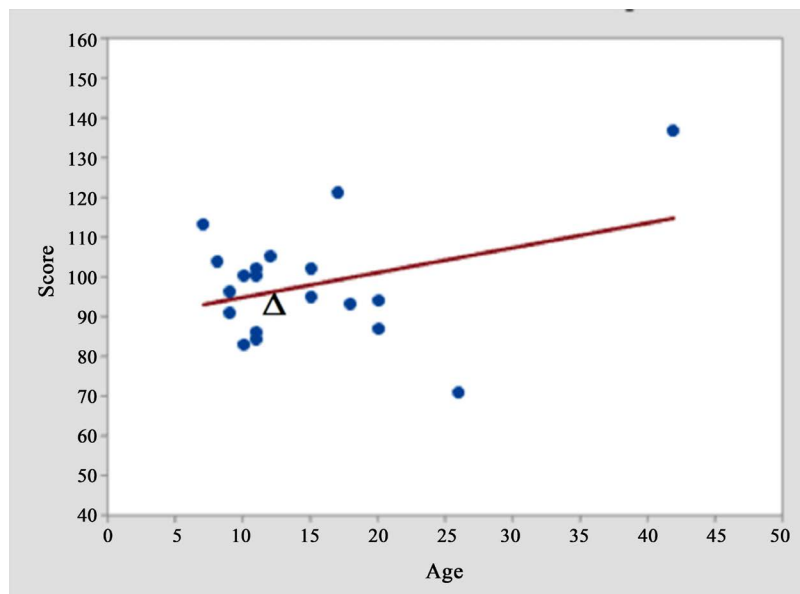| Case	x	y	Case	x	y	Case	x	y
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100
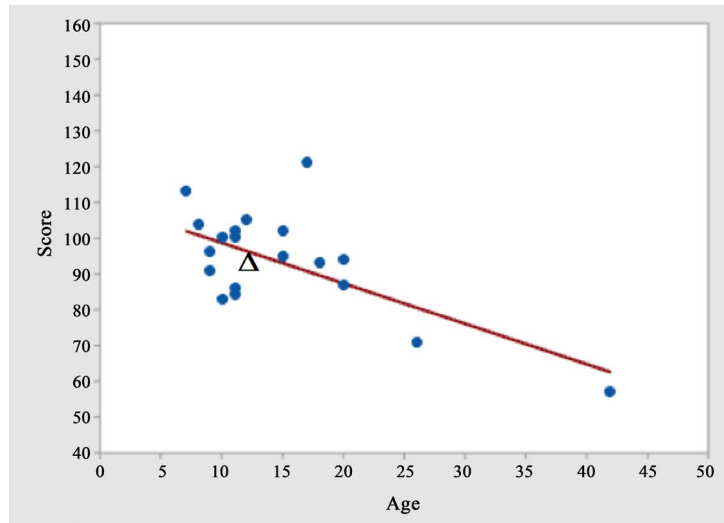
(a)

(b)

(c)

(d)

(e)

Figure 1. Altering Case 18's y-coordinate from 57 to 77, 97, 117, and 137, yielding five lines through a pivot point. (a) $y = 57$; (b) $y = 77$; (c) $y = 97$; (d) $y = 117$; (e) $y = 137$.

For any integer $m \geq 0$, the least-squares regression line of S_m is

$$y = a_m + b_m x = \frac{mV(v - b_0 u)}{(m+n)V + mu^2} + \frac{(m+n)Vb_0 + muv}{(m+n)V + mu^2}x, \tag{1}$$

and the *point of means* is

$$M_m = \left(\frac{m}{m+n}u, \frac{m}{m+n}v\right), \tag{2}$$

which is on line (1) for S_m.

When $m > 0$ and $u \neq 0$, the pivot point

$$P = \left(-\frac{V}{u}, -\frac{Vb_0}{u} \right) \tag{3}$$

is on the least-squares line for all sets S_m. This can be seen by substituting point (3) into the equation of the line (1), that is,

$$a_m + b_m \left(-V/u \right) = -Vb_0/u .$$

Point P on (3) is called the pivot point of R with respect to S_0, because P is on all regression lines for S_m, which have different slopes. Because the y-coordinate v of R is absent from the coordinates of P, it is also called the pivot point of u with respect to S_0. The set of regression lines that is created by adding copies of R, is called a pencil of lines or fan of lines through P.

When $u = 0$, the best-fit line (1) translates in the y-direction as m increases, and the pivot point is said to be at infinity. The pivot point is solely an artifact of the least-squares regression equations. Initially, it was found and explained in a linear-algebraic setting [12].

The regression lines in a fan, which is formed by vertically moving one point in the data set, intersect at the pivot point. In particular, the regression line formed by adding m copies of the point $R(u,v)$ to S_0 is equivalent to the line formed by adding a single point (u, v_m) with

$$v_m = \frac{n(1-m)Vb_0 u}{(m+n)V + mu^2} + \frac{m\left((1+n)V + u^2 \right)}{(m+n)V + mu^2} v,$$

which can be seen algebraically by setting $m = 1$ and $v = v_m$ in line (1), which yields (1).

Pivot points occur when the data are not centered at the origin. All best-fit lines can be rigidly translated, so that the new center is $(\overline{x}, \overline{y})$. The slope of each line can be found from

$$\frac{\sum (x-\overline{x})(y-\overline{y})}{\sum (x-\overline{x})^2},$$

which shows the dependence solely on the differences of each coordinate from its mean. The observations in **Figure 1** are centered at the data set's mean point $(\overline{x}, \overline{y})$.

4. Computational Shortcuts When Augmenting a Bivariate Set

The pivot point offers two shortcuts for computing equations for regression lines. This is analogous to adding the n + 1$^{\text{st}}$ value a to the data set $\{x_i : i = 1, 2, \cdots, n\}$, whose mean is $\overline{x}$. The new mean can be calculated using $(n\overline{x} + a)/(n+1)$, which requires considerably less computation than not using $\overline{x}$ [11].

One shortcut is, given set S_0, the regression line for S_m can be computed as the line containing the point of means (2) and the pivot point (3). Recall that in (4), V and b_0 are based only on the unaugmented data set.

The second shortcut involves the line obtained when multiplicity m becomes

very large, then the line (1) approaches the line

$$y = a_\infty + b_\infty x = \frac{V(v - b_0 u)}{V + u^2} + \frac{Vb_0 + uv}{V + u^2} x, \tag{4}$$

which contains the new point R and the pivot point P. The coefficients in (4) provide the tool for rapid computation for the line (1) for any m, including $m = 1$ for a single additional point. In (1), a_m is a weighted average of a_0 and a_∞, and b_m is a weighted average of b_0 and b_∞ with the same weights, in particular,

$$a_m = w a_0 + (1 - w) a_\infty \quad \text{and} \quad b_m = w b_0 + (1 - w) b_\infty, \tag{5}$$

where

$$w = \frac{nV}{(m + n)v + mu^2}, \tag{6}$$

Equations (5) are seen by substituting a_0 and b_0 from (1), a_∞ and b_∞ from (4), and w from (6) into the right-hand sides of (5), which yields a_m and b_m in (1).

5. Conclusion

Pivot points are omnipresent in applications of bivariate linear regression. In particular, they are points through which new lines pass when a data point is altered. One important purpose of altering a point is to determine its influence. We have displayed this phenomenon with the well-known data set of ages at first word versus Gesell scores, which has been analyzed by many authors from many points of view. A pivot point is a handy and efficient tool for shortening calculations when new data arises.

Acknowledgements

We are grateful to many of our colleagues who have frequently and freely shared their knowledge about regression and computational statistics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mickey, R.M., Dunn, O.J. and Clark, V. (1967) Note on the Use of Stepwise Regression in Detecting Outliers. *Computer and Biomedical Research*, **1**, 105-111. https://doi.org/10.1016/0010-4809(67)90009-2

[2] Andrews, D.F. and Pregibon, D. (1978) Finding the Outliers that Matter. *Journal of the Royal Statistical Society, Series B (Methodological)*, **40**, 85-93. https://doi.org/10.1111/j.2517-6161.1978.tb01652.x

[3] Dempster, A.P. and Gasko-Green, M. (1981) New Tools for Residual Analysis. *Annals of Statistics*, **9**, 945-959. https://doi.org/10.1214/aos/1176345575

[4] Draper, N.R. and John, J.A. (1981) Influential Observations and Outliers in Regression. *Technometrics*, **23**, 21-26. https://doi.org/10.1080/00401706.1981.10486232

[5] Moore, D.S., Notz, W.I. and Fligner, M.A. (2017) The Basic Practice of Statistics. 8th Edition, Freeman, New York.

[6] Paul, S.R. (1983) Sequential Detection of Unusual Points in Regression. *Journal of the Royal Statistical Society, Series D* (*The Statistician*), **32**, 417-424. https://doi.org/10.2307/2987543

[7] Rousseeuw, P.J. and Leroy, A.M. (1987) Robust Regression and Outlier Detection. Wiley, New York. https://doi.org/10.1002/0471725382

[8] Chatterjee, S. and Hadi, A.S. (1986) Influential Observations, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, **1**, 379-393. https://doi.org/10.1214/ss/1177013630

[9] Hoaglin, D.C. (1988) Using Leverage and Influence to Introduce Regression Diagnostics. *College Mathematics Journal*, **19**, 387-416. https://doi.org/10.1080/07468342.1988.11973146

[10] Hoaglin, D.C. (1992) Diagnostics. In: Hoaglin, D.C. and Moore, D.S., Eds., *Perspectives on Contemporary Statistics*, Mathematical Association of America, Washington, 123-144.

[11] Montgomery, D.C., Runger, G.C. and Hubele, N.F. (2011) Engineering Statistics. 5th Edition, Wiley, New York.

[12] Lutzer, C.V. (2017) A Curious Feature of Regression. *College Mathematics Journal*, **48**, 189-198. https://doi.org/10.4169/college.math.j.48.3.189

[13] Brase, C.H. and Brase, C.P. (2017) Understandable Statistics: Concepts and Methods. 12th Edition, Cengage Learning, Boston.

[14] Larose, D.T. (2015) Discovering Statistics. 3rd Edition, Freeman, New York.

[15] Triola, M.F. (2017) Elementary Statistics. 13th Edition, Pearson, Boston.