Internat. J. Math. & Math. Sci.
VOL. 13 NO. 4 (1990) 799-806

THE EFFECT OF A SINGLE POINT ON CORRELATION AND SLOPE

DAVID L. FARNSWORTH

Department of Mathematics
Rochester Institute of Technology
Rochester, New York 14623 U.S.A.

(Received December 8, 1989 and in revised form April 27, 1990)

ABSTRACT. By augmenting a bivariate data set with one point, the correlation coefficient and/or the slope of the regression line can be changed to any prescribed values. For the target value of the correlation coefficient or the slope, the coordinates of the new point are found as a function of certain statistics of the original data. The location of this new point with respect to the original data is investigated.

KEY WORDS AND PHRASES. Correlation coefficient, deletion technique, influence of data, least squares line, regression diagnostic, sample influence curve.
1980 AMS SUBJECT CLASSIFICATION CODES. 62J05, 62F35.

1. INTRODUCTION.

The correlation coefficient and slope of a least squares regression line are known to be sensitive to a few outliers in the data. For this reason, these estimators are called nonrobust or nonresistant. It is not unexpected that changing one data point or introducing a new data point will greatly perturb the regression line and various statistics associated with it. Actually, we can judiciously introduce a new data point and create a line of our choosing. The effect of adding one new data point is the subject of this paper.

In situations where we have a choice, this addition of a point is a way to lie with statistics. For example, a fit of payroll data, which we fear could reveal our bias, might be changed in this manner by appropriately hiring one more employee. Or, more benignly, in composing an example

for a classroom exercise or a made-up case study, we may want to produce certain outcomes. These might be obtained with the introduction of just one additional point.

The idea of adding one special point in order to force a regression line through the origin was studied by Casella [1]. The location of the point was shown to be related to various statistics for the line.

The robustness of an estimator is in part gauged by the influence function. The influence curve or function for a given estimator is defined pointwise as a limit as a vanishing weight is placed on the data point. These curves are derivatives and measure infinitesimal asymptotic (large sample) influence on the given estimator. The earliest work on this concept was by Hampel [2] and [3]. Also, see Belsley et al. [4], Cook and Weisberg [5], and Huber [6].

Finite sample influence curves are difference quotients, not limits. For sample size n, the numerator consists of the difference between a statistic of interest with a given point included in its formulation and the same statistic without that point. The denominator is $1/n$. This sample influence curve is used to measure the influence of a single data point within a sample. The deletion of an influential data point would lead to large values of the influence curve. The diagnostic tests for influence presented in Section 5 are standardized versions of the sample influence curve.

In Sections 2, 3 and 4 for any target value of the slope or the correlation coefficient, the coordinates of one additional point are found as a function of certain statistics of the original data. The locations of these points are presented both analytically and graphically. They are realizations of curves of constant influence where the original n-1 data are parameters. The trouble is that we may be caught tampering with the data in this way if conventional diagnostics flag the new point. In Section 5, four of the customary measures of leverage and influence are briefly discussed. In Section 6, an illustrative numerical example is given.

2. PRESCRIBING THE CORRELATION COEFFICIENT.

We are given the original data $\{(x_i, y_i) : i = 1, 2, \ldots, n-1\}$. For convenience take the center $(\overline{x}, \overline{y}) = (0, 0)$ and variances $s_x^2 = \sum x^2/(n-1) = 1$ and $s_y^2 = \sum y^2/(n-1) = 1$, that is, x_i and y_i are in standard units. The denominators of s_x^2 and s_y^2 are the full sample size since that choice leads to considerable simplification of subsequent formulas. Thus, the regression line is $y = bx$, and the correlation coefficient is $r = b$. The additional point (x_n, y_n) is expressed in standard deviation units of the original n-1 data as $(u, v) = (x_n/s_x, y_n/s_y) = (x_n, y_n)$.

The condition that the correlation coefficient of the augmented data is the prescribed value p is

$$\frac{nr + uv}{\sqrt{(n + u^2)(n + v^2)}} = p. \tag{2.1}$$

The dependence of p upon the original data is only through the values of n and r. Equation (2.1) is expressible as

$$[(1 - p^2)u^2 - np^2]v^2 + [2nru]v + [n^2(r^2 - p^2) - np^2u^2] = 0. \tag{2.2}$$

For each p, symmetries of the solution curves are about the origin and the lines $v = u$ and $v = -u$ since equation (2.1) is invariant under replacement of (u, v) with $(-u, -v)$ or $(\pm v, \pm u)$.

Vertical asymptotes appear at $(1 - p^2)u^2 - np^2 = 0$, that is, $u = \pm\sqrt{n}p/\sqrt{1 - p^2} = \pm k$. By symmetry, horizontal asymptotes are $v = \pm k$. Note that k is a monotonically increasing function of n and p^2. For $p = 0$, the asymptotes are the axes, and equation (2.1) becomes $uv = -nr$.

The discriminant of equation (2.2) is $4np^2[n + u^2][(1 - p^2)u^2 + n(r^2 - p^2)]$ which is nonnegative for $u^2 \geq n(p^2 - r^2)/(1 - p^2)$. Thus, for $p^2 \leq r^2$ there is no restriction upon u. For $p^2 > r^2$ the nonnegativity of the discriminant does restrict u from a strip about the v-axis. If $r = 0$, the $p = r$ solution curve becomes the two axes. If $r = 1$, there is just the $p^2 \leq r^2$ type of solution curves.

Representative solution curves are shown in Figure 1 for $r = b = 0.5$ and $n = 12$. Because of the symmetries, only the right half-plane is displayed. The regression line $v = 0.5u$ and the asymptotes $u = 2$ and $v = 2$ for the $p = r = 0.5$ curve are shown. The $p = -0.6$ and the $p = +0.6$ curves have asymptotes $u = \pm 2.59$ and $v = \pm 2.59$. Only the bulge about $v = u$ for the $p = 0.9$ curve is visible. Its shape is similar to that of the curve for $p = 0.6$. The $p = 0.9$ curve doubles back, changing its concavity and approaching the vertical asymptote $u = 7.15$ from the left and the horizontal asymptote $v = 7.15$ from below. All of the $p = 0$ curve is outside the square $\{-2 \leq u \leq 2, -2 \leq v \leq 2\}$, indicating how unusual the new point (u, v) would have to be to convert from $r = 0.5$ to $p = 0$. Of particular interest are the two branches of the $p = r = 0.5$ curve. The correlation coefficient is unchanged even by points which are very distant along this curve.

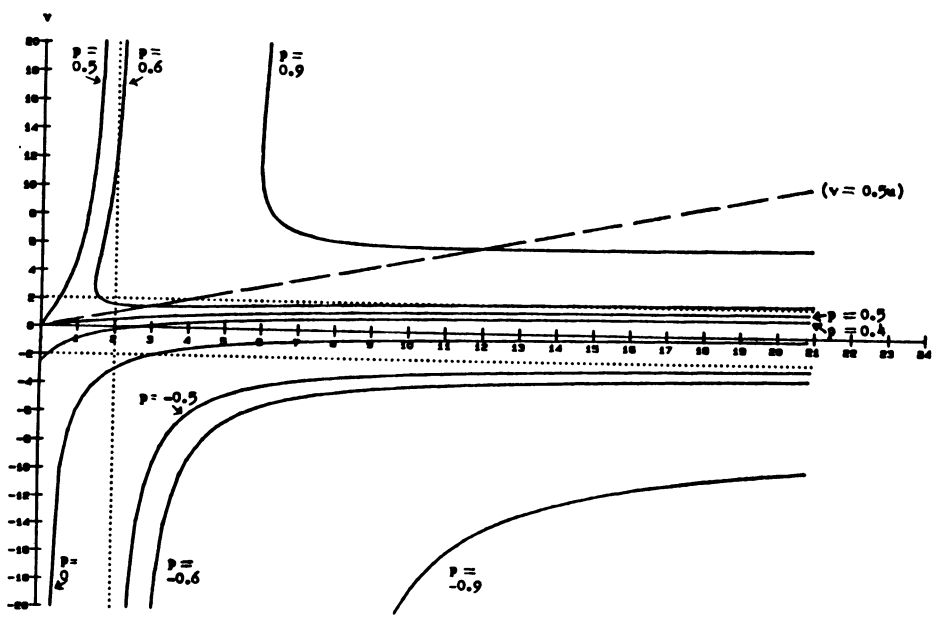

Figure 1: Curves along which (u, v) may be placed to change the value of the correlation coefficient to selected values p for $r = b = 0.5$ and $n = 12$.

3. PRESCRIBING THE SLOPE.

The condition that the slope of the least squares regression line based on the n points is the prescribed value d is

$$\frac{nb + uv}{n + u^2} = d \tag{3.1}$$

or

$$v = n(d - b)/u + du.$$

For each d the solution curve is symmetric with respect to the origin. Asymptotes are $u = 0$ and $v = du$ for $d \neq b$. Maxima and minima occur at $(\pm\sqrt{n(d-b)/d}, \pm 2d\sqrt{n(d-b)/d})$ for $(d-b)/d > 0$. Intersections with the u-axis occur at $(\pm\sqrt{n(b-d)/d}, 0)$ for $(d-b)/d \leq 0$. Placing $d = b$ in equation (3.1) implies $u = 0$ and $v = bu$.

Figure 2 displays selected solution curves for $b = r = 0.5$ and $n = 12$. Because of symmetry, just the right half-plane is shown. Of course, the $d = 0$ curve in Figure 2 is the same as the $p = 0$ curve in Figure 1.

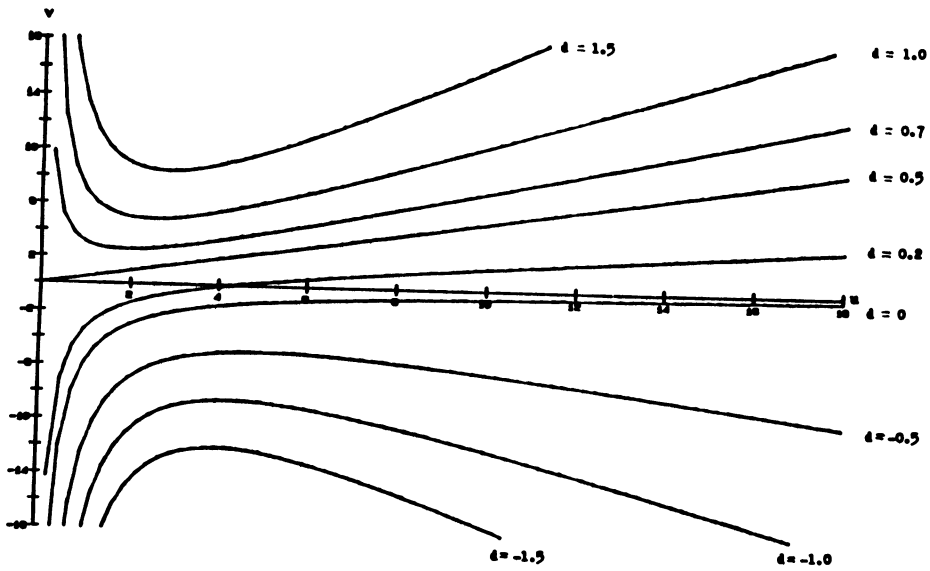

Figure 2: Curves along which (u, v) may be placed to change the value of the slope to selected values d for $r = b = 0.5$ and $n = 12$.

4. PRESCRIBING BOTH THE CORRELATION COEFFICIENT AND THE SLOPE.

From Figures 1 and 2 it is apparent that the slope can be drastically changed while the correlation coefficient is numerically the same. Setting $p = r = b$, equations (2.1) and (3.1) give

$$d = \sqrt{\frac{n + v^2}{n + u^2}}\, b. \tag{4.1}$$

If n is large, then u or v must be very large for d to be much different from b. Within the constraints $p \in (-1, 1)$ and sign p = sign d, the two values p and d can be chosen arbitrarily by specifying

the two numbers u and v. However, the issue of remoteness of (u, v) from $(\overline{x}, \overline{y}) = (0, 0)$ and from $y = bx$ is important and is addressed below.

5. DETECTING THE NEW POINT.

Among the available diagnostic tests for gauging whether a data point (x_i, y_i) is an interloper, four are the most common. Each measures a different feature of data. See Neter et al. [7], Atkinson [8], or Cook and Weisberg [5]. The notation and critical values are not uniformly agreed upon in the literature. Those suggested in Chapter 11 of Neter et al. [7] are used below.

First, the point (u, v) could be an x-outlier, that is, it could be far in an horizontal direction from the mean of all n data. Such points influence slope the most for a given change in the y-coordinate and are said to have high leverage. The usual measure of leverage for (x_i, y_i) is the diagonal element h_{ii} of the hat matrix. See Hoaglin and Welsch [9]. Each h_{ii} depends upon all the x-coordinates but does not depend upon the y-coordinates of the data. If $4/n < h_{ii} \leq 1$, then (x_i, y_i) will be said to have high leverage. If $1/n \leq h_{ii} \leq 4/n$, then (x_i, y_i) has low leverage.

The next three diagnostic measures arise from the deletion technique. The impact of a data point is often detected by using this technique: The data point is removed; the statistic of interest is computed; and a standardized difference between that statistic and the corresponding statistic utilizing all the data is analysed as the diagnostic measure. The unstandardized differences divided by $1/n$, are the sample influence curves discussed in the Introduction.

Second, the point (u, v) could be outlying with respect to the y-value. A diagnostic statistic for measuring whether (x_i, y_i) is a y-outlying point is the externally studentized residual d_i^*, which is a standardized vertical distance at x_i between the point (x_i, y_i) and the regression line created without (x_i, y_i). Each d_i^* is t-distributed with df = n-3. Using the prediction limits for a new point (x_i, y_i) based on the other n-1 data's regression line as the criteria for designating a point as outlying is equivalent to using d_i^*, as shown in Neter et al. [7], page 399.

Third, (u, v) could inordinately change predicted values. The statistic $(DFFITS)_i$ is a standardized difference between the predicted values at x_i — one computed with, and one without, the data point (x_i, y_i). The point (x_i, y_i) is said to be influential in this sense if $|(DFFITS)_i| > 2\sqrt{2/n}$ for large n and if $|(DFFITS)_i| > 1$ for small to moderate n. This test is based on the usual confidence bands for regression lines.

Fourth, (u, v) could unduly impact the slope. The statistic $(DFBETA)_i$ is a standardized difference between the slope of the two regression lines — one with (x_i, y_i) and one without (x_i, y_i). If $|(DFBETA)_i| > 1$ for small to moderate n and if $|(DFBETA)_i| > 2/\sqrt{n}$ for large n, then (x_i, y_i) is called influential in this sense.

Generally, in data analysis each of these four measures is routinely computed for all n data sets obtained by deleting each point in turn. Then, the structure of each of the four batches of numbers is examined. But, since all critical values are absolute, we shall use an abridged procedure and find these four statistics for only the possible augmenting point (u, v) in the illustrative example in the

next section.

6. NUMERICAL EXAMPLE.

Consider the following constructed data in standard units: (-1.54, -1.07), (-1.16, -1.26), (-0.77, 1.26), (-0.77, -0.63), (-0.39, -0.60), (-0.39, 0.60), (0.39, -0.63), (0.77, 0), (1.16, 1.89), (1.16, -0.63), and (1.54, 1.07). For these 11 data $r = b = 0.50$. Let us select $d = 0.7$ as our desired slope and points $A(0.39, 6.49)$, $B(0.80, 3.55)$, $C(1.85, 2.59)$, and $D(3.00, 2.90)$ as possible auxiliary points on the $d = 0.7$ curve. Point B is one where the correlation coefficient remains 0.5. See equation (4.1). Point C is the local minimum point of the $d = 0.7$ curve.

Each of the four diagnostic statistics discussed in Section 5 is presented in Table 1 for each of these four points representing (x_n, y_n). The symbol * next to a numerical value means that the point would be deemed unusual, that is, considered of high leverage or influence, by the criteria given in Section 5. The level of significance is 0.05 for the d_n^* column.

Point	h_{nn}	d_n^*	$(DFFITS)_n$	$(DFBETA)_n$
A	0.09	6.27*	2.03*	0.70
B	0.13	3.08*	1.19*	0.71
C	0.29	1.48	0.93	0.79
D	0.48*	1.06	1.01*	0.92

Table 1.

Points on the $d = 0.7$ curve with $x_n = u \geq 3.59$ will be flagged with $(DFBETA)_n > 1$.

Table 1 shows that introducing point C will give us the desired change of slope but will not be detected with these four diagnostic statistics. Other points on the $d = 0.7$ curve may or may not be detected with various test statistics depending upon their locations.

The original eleven data, the regression line based on the original data, and a branch of the $d = 0.7$ curve with the four points labeled A, B, C, and D are shown in Figure 3. For purposes of scaling of d_n^*, the hyperbolic 95% prediction band is displayed as well.

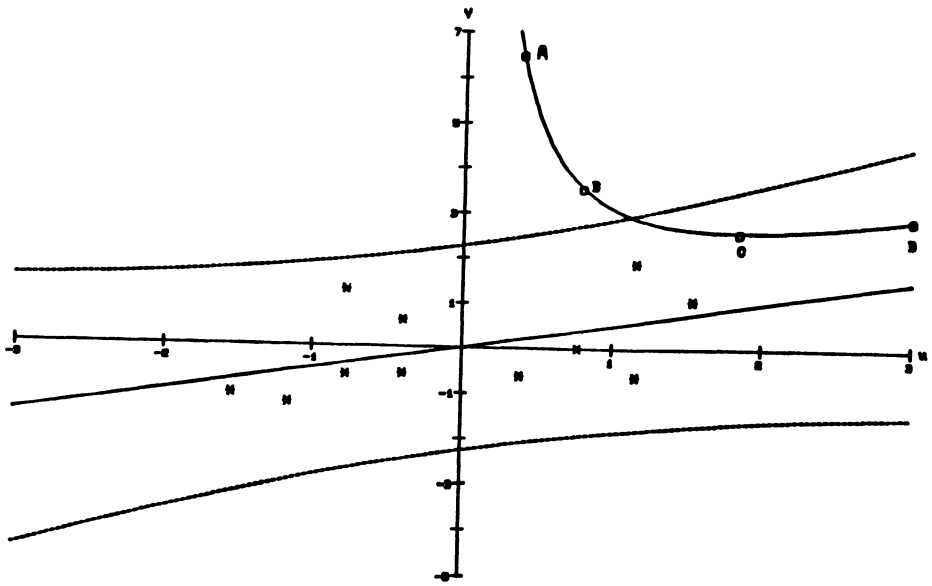

Figure 3: Numerical example described in Section 6.

REFERENCES

1. CASELLA, G. Leverage and Regression Through the Origin, Amer. Statist. 37 (1983), 147-152.

2. HAMPEL, F.R. Contributions to the Theory of Robust Estimation, Ph.D. Thesis, University of California, Berkeley, 1968.

3. HAMPEL, F.R. The Influence Curve and Its Role in Robust Estimation, J. Amer. Statist. Assoc. 69 (1974), 383-393.

4. BELSLEY, D.A., KUH, E. and WELSCH, R.E. Regression Diagnostics, John Wiley and Sons, New York, 1980.

5. COOK, R.D. and WEISBERG, S. Residuals and Influence in Regression, Chapman and Hall, New York, 1982.

6. HUBER, P.J. Robust Statistics , John Wiley and Sons, New York, 1981.

7. NETER, J., WASSERMAN, W. and KUTNER, M.H. Applied Linear Regression Models, Second Edition, Richard D. Irwin, Inc., Homewood, Illinois, 1989.

8. ATKINSON, A.C. Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Oxford University Press, Oxford, 1985.

9. HOAGLIN, D.C. and WELSCH, R.E. The Hat Matrix in Regression and ANOVA, Amer. Statist. 32 (1978), 17-22.