

The Noncontribution of Some Data in Least Squares Regression Predictions

Dr. Teresa L. Bittner, Walden University
Dr. Kurt W. Norlin, Claremont Graduate University

Teresa L. Bittner, Ph.D.
Part time faculty, Walden University
School of Management–AMDS/Operations Research
957 Wilmington Way
Redwood City, CA 94062
(650) 599-9188 (Voice)
(650) 599-9422 (Fax)

Kurt W. Norlin, Ph.D.
Adjunct Professor, Claremont Graduate University
School of Educational Studies
Harper Hall 202
150 East Tenth Street
Claremont, CA 91711
(949) 262-9363

Abstract

In least-squares estimation, a phenomenon can occur which has escaped attention thus far. Under certain conditions, one or more points in a data set have no influence on predictions made using ordinary least-squares models. This effect, which amounts to a discarding of y -data values, leads to predictions that may be suboptimal compared with predictions that use all y -data values.

In this paper, the relationship between data points and least squares predictions independent of those data points is demonstrated for straight-line models. When straight-line prediction is used, potentially noncontributory data can now be identified before the dependent variable data is even collected. This result can help statisticians gather their input data more efficiently and analyze existing data with better understanding.

Future research will deal with the same issue for polynomial models as well as general univariate models that are linear in the unknown coefficients.

Keywords: Least Squares; Prediction; Linear Models; Sensitivity Analysis; Noncontributory data

The Noncontribution of Some Data in Least Squares Regression Predictions

Introduction

The method of least squares is probably the most popular technique today to fit data to functions, estimate parameters, and determine the statistical properties of those estimates. However, a phenomenon that occurs only under certain conditions has escaped attention until now due to the fact that theoretical work generally concentrates on models and their properties rather than on individual predictions. The problem is that under certain conditions, some data fails to contribute to predictions made using least squares models. Sensitivity analysis in regression would have been the logical place for this discovery, but it is generally looked at from the perspective of changes in the parameter estimates rather than individual predictions (Chatterjee & Hadi, 1988; Cook, 1977). Even when the effect of the i th observation on a predicted value (\hat{y}) is mentioned, it is generally used to identify collinearities among regression variates, or to find data that is overly influential to the model. (Belsley, Kuh, & Welsch, 2004; Chatterjee & Hadi, 1988, p. 120–121).

Under certain conditions, the i th observation has no influence at all on particular \hat{y}_j predictions. This finding is important because it is usually assumed that all data is being used in such predictions, and results can be skewed if data points are in fact dropping out of the prediction equations, especially for small values of n . It seems clear that the loss of a data point when predicting y -values in a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$ is a loss of information, and such a prediction may be suboptimal in comparison to some other prediction technique that

uses all the y -data points in its calculation. This phenomenon is also important in cases where data collection is expensive. In these cases, cost efficiency can be improved by using “dummy” data to replace real data points that would drop out of the prediction equation anyway. Further, this finding extends beyond straight line models to other models that are linear in the unknown coefficients. In fact, the number of predicted values affected increases as the number of unknown parameters increase, both in univariate and multivariate regression.

This paper describes the specific conditions under which observations drop out of prediction calculations in straight line models, develops relationships between the data point that drops out and the predicted y -value for which this happens, and proves the existence and accuracy of these relationships. A physical application of this phenomenon is also discussed, as are suggestions for further work on this problem.

The Phenomenon of Data Dropping Out in Least Squares Predictions

Linear modeling using the L2 (least squares) norm is a mature field of statistics. The mathematics associated with it is elegant, and the technique lends itself to many applications using closed form solutions that are efficient and convenient. However, some data has no influence on predictions made using least squares. This happens in a predictable way, and occurs for a range of models that are linear in the unknown coefficients, and in multivariate linear modeling as well. This paper will cover only the simplest linear case, while the more complex cases will be covered in future papers.

The analysis of this phenomenon begins with a simple example, followed by a short general analysis of a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$. The implications of this analysis are then explored at length in two phases. First, we look at the special case where the x -values are evenly spaced in order to derive the integer cases of the phenomenon. This is followed by a full analysis of the general case. Finally, a physical application to this phenomenon is presented, and ideas for future work are suggested.

Example 1.

The enrollment in Kindergarten at Allen Elementary School in San Jose, California for the last four years is as follows: (2001, 62), (2002, 43), (2003, 78), (2004, 82) (CA Department of Education, 2005). The data is plotted in Figure 1 below along with the least squares regression line.

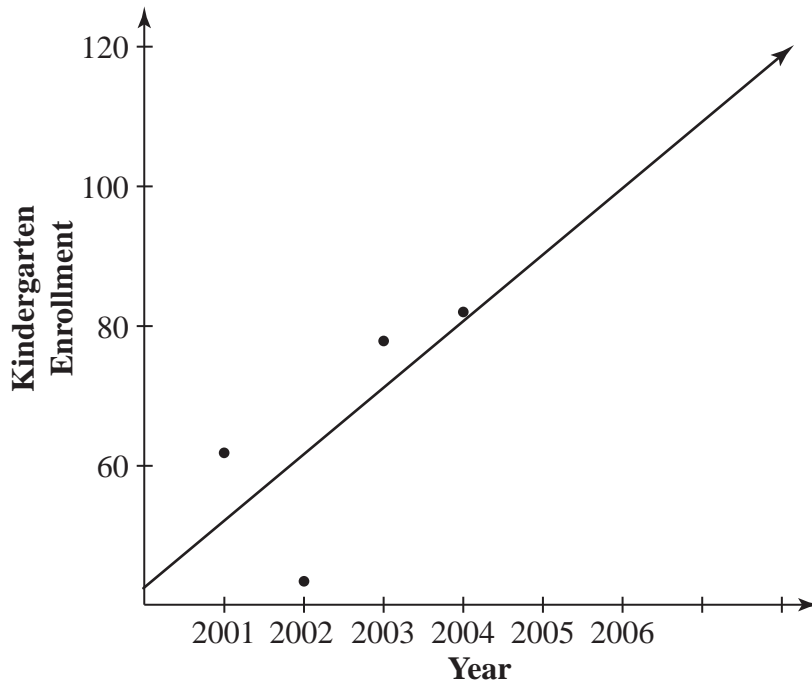


Figure 1. Allen Elementary School Kindergarten Enrollment from 2001 to 2004

Suppose the school wishes to estimate the enrollment for 2005. Assume that the enrollment

behaves according to a linear function $Y = X\beta + \varepsilon$, with $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$, where the first row

corresponds to 2001, the second row to 2002, etc., $Y = \begin{bmatrix} 62 \\ 43 \\ 78 \\ 82 \end{bmatrix}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$. When the least

squares solution is computed, the solution is $\hat{y}_i = X\hat{\beta} = X(X'X)^{-1}X'Y = 9.5x_i + 42.5$. To

estimate enrollment for 2005, $x_5 = 5$ is substituted into the model to obtain

$\hat{y}_5 = 42.5 + 9.5(5) = 90$. Now, suppose the second y-value is changed so that $y_2 = 20$ instead of

the original value of 43. This changes the regression line so that $\hat{y}_i = 31 + 11.8x_i$, but $\hat{y}_5 = 31 + 11.8(5) = 90$ as before. Similarly, the second y -value can be changed again so that $y_2 = 120$. Now the regression yields $\hat{y}_i = 81 + 1.8x_i$. Since the value of y_2 is so large, it seems reasonable to expect that the new estimate for 2005 enrollment would be much higher than before. Yet the computation for \hat{y}_5 is $\hat{y}_5 = 81 + 1.8(5) = 90$ just as before. This is the case even though the regression line itself has certainly shifted. In short, it appears that the value of y_2 has no effect at all on the estimate for the 2005 Kindergarten enrollment.

It is helpful to look at this phenomenon graphically. The three regression lines, $\hat{y}_i = 42.5 + 9.5x_i$, $\hat{y}_i = 31 + 11.8x_i$, and $\hat{y}_i = 81 + 1.8x_i$ are graphed on the same set of axes along with the original data.

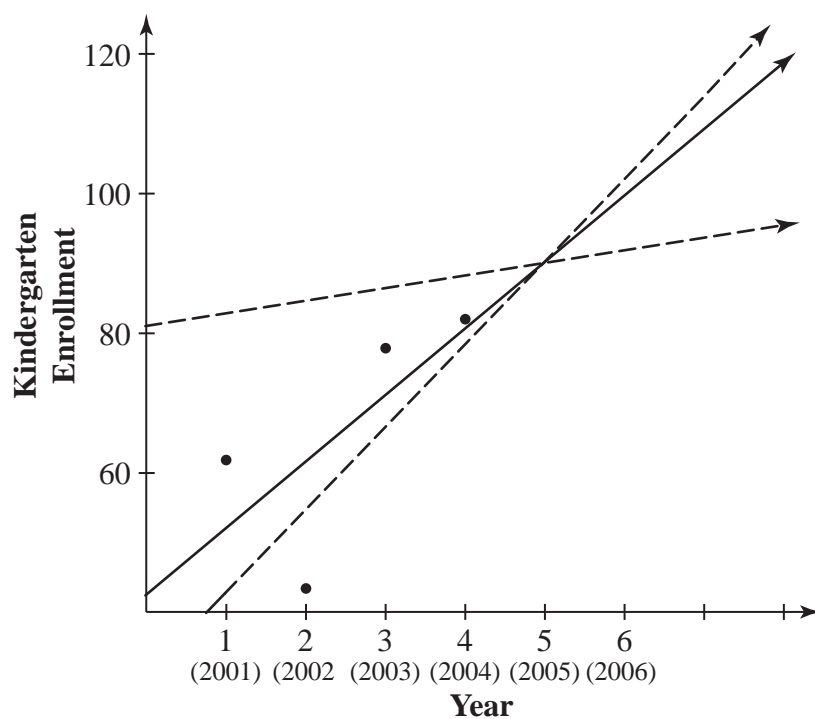


Figure 2. Allen Elementary School Kindergarten Enrollment Prediction Lines

It is easy to see in Figure 2 that the three lines intersect at (5, 90), which corresponds to the prediction that there will be 90 students in the 2005 Kindergarten class. In fact, it will later be shown that y_2 can be changed arbitrarily, and all of the regression lines will intersect at (5, 90). Notice, however, that the lines only intersect at this one point, and that for all other values of x the predictions for enrollment will be different when y_2 is changed. This phenomenon occurs for other values of n and \hat{y} as well, and it will be useful to derive a mathematical relationship between noncontributory data points and the predicted values for which these points are noncontributory.

The Theoretical Basis for the Phenomenon

To set up the basis for the analytical exploration of this phenomenon, assume that the x -data and y -data values are unrestricted real numbers. In other words, assume

x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n where the x -values and y -values are unrestricted real values.

Assume there is a linear model in x such that

$$\begin{matrix} Y & = & X & \beta & + & \varepsilon \\ n \times 1 & & n \times 2 & 2 \times 1 & & n \times 1 \end{matrix}$$

or, in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

It is well known that the ordinary least squares solution for this model is

$$\hat{\beta} = (X'X)^{-1} X'Y.$$

The well known matrix form for $\hat{\beta}$ for this model is:

$$\hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad (1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3)$$

and

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}). \end{aligned}$$

Now consider the special case where the x_i 's are equally spaced. The simplest case is to let $x_i = i$, for $i = 1, 2, \dots, \mathcal{L}$. However, this simple case can be generalized to any equally spaced x -values by taking a linear transformation on $x_i = i$ so that $x_i = ai + b$, where a and b are scalar constants.

Note that if $x_i - \bar{x}$ is multiplied by the scalar constant b , then the new predicted value for y , named $\hat{y}_{i \text{ new}}$, can be expressed as

$$\begin{aligned} \hat{y}_{i \text{ new}} &= \bar{y} + \frac{\sum (y_i - \bar{y})(b)(x_i - \bar{x})}{\sum b^2 (x_i - \bar{x})^2} \bullet b(x_i - \bar{x}) \\ &= \bar{y} + \frac{b^2 (\sum (y_i - \bar{y})(x_i - \bar{x})) (x_i - \bar{x})}{\sum b^2 (x_i - \bar{x})^2} \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ &= \hat{y}_i \end{aligned}$$

In other words, the scalar multiple on $x_i - \bar{x}$ does not affect \hat{y}_i .

Now, instead of $x_i = i$, use the more general case where $x_i = ai + b$ and examine the change, if any, in \hat{y}_i . Then since the only part of the equation for \hat{y}_i that is affected by the transformation is

the quantity $(x_i - \bar{x})$, it is sufficient to look at the effect of the transformation on this quantity. It is easily shown that

$$\begin{aligned} x_i - \bar{x} &= (a + bx_i) - \frac{1}{n} \sum (a + bx_i) \\ &= b \left(x_i - \frac{1}{n} \sum x_i \right) \\ &= b(x_i - \bar{x}) \end{aligned}$$

Since the transformation only results in a scalar transformation of $x_i - \bar{x}$, and it was previously shown that a scalar multiple of $x_i - \bar{x}$ does not affect \hat{y}_i , this result shows that $x_i = i$ can be used without loss of generality to represent any evenly spaced x -values so long as the only concern is \hat{y}_i . The following analysis makes the assumption that $x_i = i$, but the results are valid for any equally spaced x -values.

The Case when $n = 4$

Now suppose that $n = 4$. Then $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ is the simplest case for the x -values if they are evenly spaced. Using (1), the calculations yield

$$\sum x_i = 10, \sum x_i^2 = 30, \left(\sum x_i \right)^2 = 100, \text{ and } \hat{\beta} = \frac{1}{20} \left[\begin{array}{l} 30 \sum y_i - 10 \sum x_i y_i \\ 4 \sum x_i y_i - 10 \sum y_i \end{array} \right].$$

In order to estimate \hat{y}_5 , the value for x_5 is substituted into the model to yield $\hat{y}_5 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_5$, where $x_5 = 5$.

Then simplification gives

$$\begin{aligned} \hat{y}_5 &= \frac{3}{2} \sum y_i - \frac{1}{2} \sum x_i y_i + 5 \left[\frac{1}{5} \sum x_i y_i - \frac{1}{2} \sum y_i \right] \\ &= -\frac{1}{2} y_1 + \frac{1}{2} y_3 + y_4 \end{aligned}$$

Interestingly, this least squares estimator for \hat{y}_5 is completely independent of y_2 , illustrating the theory behind Example 1. Since the estimate of \hat{y}_5 is independent of y_2 , it is clear why the various graphs of regression equations intersect at one point, even when y_2 is varied.

The Development of the General Case

Now the result for the simple linear model shown above is generalized for n , and the simplest case is a prediction of y_{n+1} . If the x -values are as before, then

$$\sum_{i=1}^n x_i = \sum_{i=1}^n i = \frac{n(n+1)}{2},$$

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}, \text{ and}$$

$$\left(\sum x_i\right)^2 = \frac{n^2(n+1)^2}{4}.$$

The result is now generalized for $\hat{\beta}$. Using (1) gives

$$\begin{aligned} \hat{\beta} &= \frac{1}{n \sum x_i^2 - \left(\sum x_i\right)^2} \left[\begin{array}{l} \sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{array} \right] \\ &= \frac{1}{\frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}} \left[\begin{array}{l} \frac{n(n+1)(2n+1)}{6} \sum y_i - \frac{n(n+1)}{2} \sum x_i y_i \\ n \sum x_i y_i - \frac{n(n+1)}{2} \sum y_i \end{array} \right] \\ &= \frac{12}{n^2(n+1)(n-1)} \left[\begin{array}{l} \frac{n(n+1)(2n+1)}{6} \sum y_i - \frac{n(n+1)}{2} \sum x_i y_i \\ n \sum x_i y_i - \frac{n(n+1)}{2} \sum y_i \end{array} \right] \end{aligned}$$

Therefore,

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{2(2n+1)}{n(n-1)} \sum y_i - \frac{6}{n(n-1)} \sum x_i y_i \\ \frac{12}{n(n+1)(n-1)} \sum x_i y_i - \frac{6}{n(n-1)} \sum y_i \end{bmatrix}, \quad (4)$$

and

$$\hat{y}_i = \frac{2(2n+1)}{n(n-1)} \sum y_i - \frac{6}{n(n-1)} \sum x_i y_i + x_i \left(\frac{12}{n(n+1)(n-1)} \sum x_i y_i - \frac{6}{n(n-1)} \sum y_i \right) \quad (5)$$

Now a general result is sought for arbitrary n . In general, (4) and (5) give

$$\begin{aligned} \hat{y}_{n+1} &= \frac{2(2n+1)}{n(n-1)} \sum y_i - \frac{6}{n(n-1)} \sum x_i y_i + (n+1) \left(\frac{12}{n(n+1)(n-1)} \sum x_i y_i - \frac{6}{n(n-1)} \sum y_i \right) \\ &= \sum \left(\frac{-2(n+2)}{n(n-1)} + \frac{6}{n(n-1)} x_i \right) y_i \end{aligned} \quad (6)$$

It can be seen from equation (6) above that the integer cases where y_i drops out can be found when the coefficient of y_i is 0. So,

$$\frac{-2(n+2)}{n(n-1)} + \frac{6}{n(n-1)} x_i = 0$$

Therefore,

$$\begin{aligned} 2n + 4 &= 6x_i \\ n &= 3x_i - 2, \quad x_i = 2, 3, \mathcal{L} \end{aligned}$$

Note that the case where $n = 1$ is eliminated because it is a trivial case.

By manipulating the above equation, a more convenient form of the sequence can be written.

When $n = 4 + 3k$, $k = 0, 1, 2, 3, \mathcal{L}$ then y_{k+2} drops out of the prediction for \hat{y}_{n+1} .

In particular, substituting $n = 4 + 3k$ in the above equation for \hat{y}_{n+1} and simplifying shows the result that the $(k + 2)$ th y -data point drops out when estimating \hat{y}_{n+1} , as stated.

The Relationship Between \hat{y}_p and y_d

To continue, it will be necessary to define some notation. The individual data points to be fitted to a linear model can be represented as ordered pairs of the form (x_i, y_i) . Data points that have no influence on a prediction will be denoted (x_d, y_d) , where y_d is the observed data point in the d th position. The prediction for which (x_d, y_d) has no influence is denoted by (x_p, \hat{y}_p) , where \hat{y}_p is the predicted value computed by substituting x by x_p in the model. The indices d and p are integer valued, but x_d and x_p are real numbers, unless otherwise specified.

The simplified case where the x_i 's are equally spaced helped to determine which integer values of k cause y_k to have no influence when estimating some \hat{y}_i , and to find the relationship between k and i . While the integer cases have an obvious use, it is also possible to derive a closed form relationship between x_d and x_p . In this case, d is restricted to be a value for which x_d exists as measured data, while x_p may be any real value. Here the x_i values are unrestricted. Supposing such a relationship exists, the derivation of the relationship between x_p and x_d follows.

Recall the solution for $\hat{\beta} = (X'X)^{-1} X'Y$, where X and Y are defined as before. Assuming that there exists a value of \hat{y}_p that is not dependent on y_d , the relationship between x_p and x_d is independent of the actual values of the y -data. This can be seen by taking the derivative of \hat{y}_p with respect to $y_d \left(\frac{d\hat{y}_p}{dy_d} \right)$. This literally shows the change in \hat{y}_p with respect to y_d . If the observation y_d does not affect the prediction \hat{y}_p , then this derivative should be equal to zero.

Now, the \hat{y}_p values are dependent on the value of x_p and the $\hat{\beta}$ -values, and these values are related by the equation $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$. However, the $\hat{\beta}$ -values are linear combinations of the y_i values. Therefore, once the derivative is taken, there is no y in the expression. Even without stating the expression for \hat{y}_p explicitly, we can say that taking the derivative of this value with respect to y_d will eliminate all expressions that have to do with any value of y except for y_d . At that point the only thing left is the coefficient of anything having to do with y_d , since we know based on our model that there would be no expression that is nonlinear in y_d .

Since the change in \hat{y}_p with respect to y_d does not depend on any of the y -values, any values of y can be used in order to derive the relationship between \hat{y}_p and y_d , and their corresponding values x_p and x_d . Therefore, in order to derive the relationship between x_p and x_d , the values of the y -vector can be varied at will without loss of generality. Visually, the desired result is the point x_p at which all the various regression lines corresponding to different values of y_d intersect when the other y -values are held constant. For this purpose any two lines will suffice, and thus y -values can be chosen for maximal convenience. Thus, let all the y -values other than y_d equal 0, and let y_d be either 0 or 1. (Thus Y is either the zero vector or the indicator function on Y at d . The indicator function is the vector that contains all zeros except for a “one” in some position d .) The other y -values do not matter for this purpose anyway, so it is easiest to let them equal 0. Thus Y is the indicator function on Y at d , denoted I_d . Then

$$XY = \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_d & \mathcal{L} & x_n \end{bmatrix} \begin{array}{c} 0 \\ \mathcal{M} \\ 0 \\ 1 \\ 0 \\ \mathcal{M} \\ 0 \end{array} = \begin{bmatrix} 1 \\ x_d \end{bmatrix},$$

and

$$\hat{\beta} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} 1 \\ x_d \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}.$$

Now, by performing the matrix multiplications, the individual components of $\hat{\beta}$ can be expressed in terms of x_d as

$$x_p = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - x_d \cdot n}. \quad (7)$$

By solving the equation (7) for different values of n and x_p , the same integer results as before are obtained. Namely, when $n = 4 + 3k$, $k = 0, 1, 2, \mathcal{L}$, y_{k+2} drops out when estimating \hat{y}_{n+1} . A theorem and the proof of this result follows.

Theorem 1. Given $y = \beta_0 + x\beta_1 + \varepsilon$, Y ($n \times 1$), X ($n \times 2$), specified. Let \hat{y}_p (p real), be a prediction based upon $\hat{\beta} = (X'X)^{-1} X'Y$. Then there exists an integer value d , $1 \leq d \leq n$, and $d \neq p$, such that \hat{y}_p does not depend on y_d , and the relationship between p and d is specified by

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}.$$

Proof: It needs to be shown that \hat{y}_p does not change when y_d is varied arbitrarily, and the other y -values are fixed.

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x})\end{aligned}$$

and
$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \quad (8)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (9)$$

Also, $\hat{y}_p = \bar{y} + \hat{\beta}_1 (x_p - \bar{x})$.

Note that with some simplification,

$$\begin{aligned}x_p &= \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d} \text{ can be written as} \\ x_p &= \frac{\sum (\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x}\end{aligned} \quad (10)$$

Now, by substituting equation (10) for x_p into the equation $\hat{y}_p = \bar{y} + \hat{\beta}_1 (x_p - \bar{x})$ and substituting for (8) for $\hat{\beta}_1$ as well, the equation becomes

$$\hat{y}_p = \bar{y} + \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \left(\frac{\sum (\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x} - \bar{x} \right).$$

Since $\sum (x_i - \bar{x})^2 = \sum (\bar{x} - x_i)^2$, the equation above simplifies to

$$\hat{y}_p = \bar{y} + \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{n(\bar{x} - x_d)}.$$

Expanding gives

$$\hat{y}_p = \frac{1}{n} \sum y_i + \frac{\sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \cdot \bar{y})}{n(\bar{x} - x_d)},$$

and multiplying both sides of the equation by $n(\bar{x} - x_d)$ gives

$$n(\bar{x} - x_d) \hat{y}_p = \frac{1}{n} (n)(\bar{x} - x_d) y_d + \sum (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x} \cdot \bar{y})$$

If y_d has no influence on equation for \hat{y}_p , only the parts of \hat{y}_p that involve y_d need to be

calculated. It needs to be proven that these terms equal zero. Therefore, noting that $\bar{x} = \frac{1}{n} \sum x_i$

and $\bar{y} = \frac{1}{n} \sum y_i$, and eliminating all the terms not depending on y_d by writing $\sum y_i = y_d + \sum_{i \neq d} y_i$

yields

$$\begin{aligned} n(\bar{x} - x_d) \hat{y}_p &= \bar{x} y_d - x_d y_d + x_d y_d - \bar{x} y_d - \bar{y} \sum x_i + \sum \bar{x} \cdot \bar{y} \\ &= -n\bar{x} \cdot \bar{y} + n\bar{x} \cdot \bar{y} \\ &= 0. \text{ QED} \end{aligned}$$

Example 3.

The finance manager of a major fast food chain suspects that the gradually increasing number of tacos sold can be usefully modeled by a linear function. She has decided to compile data on the number of tacos sold for several years in order to estimate the number of tacos likely to be sold over the next several years if the pattern continues. She knows she can compile data for the last 11 years, except that the year 3 data was irretrievably lost due to a computer crash several years ago. Therefore, she can compile data for years 1, 2, 4, 5, 6, 7, 8, 9, 10, and 11,

where year 11 corresponds to last year. The finance manager has noted that it will take a considerable amount of effort to retrieve the data for the number of tacos sold each year, due to the way the data was originally recorded. Therefore, she wants to make sure that all the data she collects will be used in her estimates for years 12, 13, 14, and 15. Will any of the predicted values the manager wants be independent of any of the data values?

Using Theorem 1, it is clear that the relationship between a data point (x_d, y_d) and any predicted value (x_p, y_p) that is independent of that data point is

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}.$$

In this case, the above equation needs to be solved four separate times for x_d , once for each of years 12, 13, 14, and 15.

The required calculations are $\sum x_i = 63$, $\sum x_i^2 = 497$, and $n = 10$. The solutions to the equation for each value of x_p are given in the table below. These values of x_p correspond to the collected data value of y_p .

Table 2. Values For Which y_d Does Not Affect \hat{y}_p

x_p	x_d That Does Not Affect y_p
12	4.5
13	4.8
14	5
15	5.1

Therefore, the fourth data value (when $x = 5$) will not affect the predicted value of \hat{y}_{14} . This might be a reason for the finance manager not to bother collating the data for that year.

Admittedly, given that the fourth data value is still apparently relevant to the predictions for years 12, 13, and 15, there might still seem to be a reason to go ahead and compile the fifth data

value. However, we see a hint emerging that the fifth data value may not have much effect on the predictions for years 12, 13, and 15 after all. In fact, initial findings indicate that there are “degrees of relevance” for various data points that would strengthen the case for omitting the fifth data point. The analysis of this issue is beyond the scope of this paper and will be the topic of a future paper.

A Physical Application

Interestingly, there is a physical parallel to the statistical result expressed by Theorem 1. Recall the linear model in x where

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Consider the following physical application:

Suppose there are N x -values where the x -values are allowed to be arbitrary, (i.e., not necessarily equally spaced). For each x -value, place a point mass at the corresponding point on the number line, where all the masses are 1 unit in magnitude. Now suppose all these masses are joined by massless rod connectors to form a single rigid body that is floating in space, as seen in Figure 3 below.

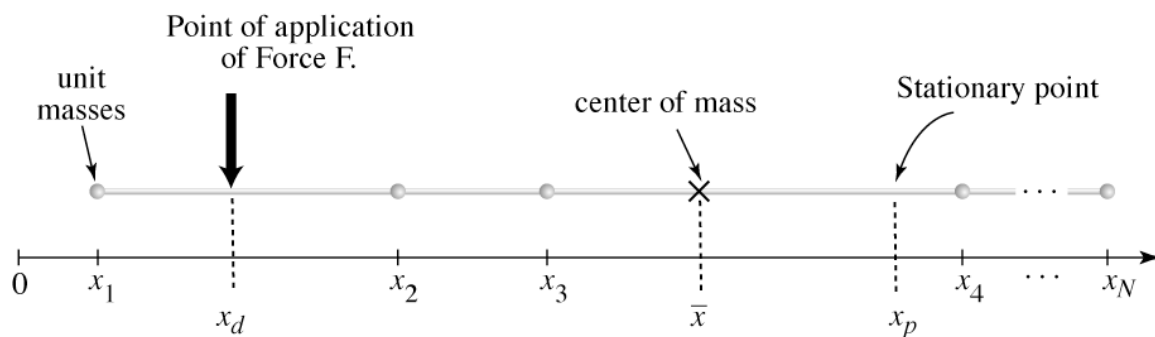


Figure 3. Point masses joined by a massless rod to form a single rigid body.

If one were to press sideways against this linear body at some arbitrary point x_d that is not the center of mass, then the body will now begin to translate and to rotate. There will, however, be a point x_p at which the effects of translation and rotation cancel out, a point that will remain stationary. The linear body will pivot about that point. (See Figure 4 below). The relationship between x_d and x_p can be derived from simple physical relationships.

We know from basic physics that

$$F = ma, \text{ where } F \text{ is force, } m \text{ is mass, and } a \text{ is acceleration}$$

and

$\tau = I\alpha = Fd$, where τ is torque and $I\alpha$ is the moment of inertia multiplied by the angular acceleration, all with reference to rotation about the center of mass.

For our rod-connected unit masses,

$$ma = Na_{\bar{x}} \quad (a_{\bar{x}} = \text{acceleration of the center of mass, } N \text{ is the number of point-masses),}$$

$$Fd = F \cdot (\bar{x} - x_d),$$

and the moment of inertia is $I = \sum_{i=1}^N (\bar{x} - x_i)^2$.

Then $Fd = I\alpha$ becomes

$$F(\bar{x} - x_d) = \sum_{i=1}^N (\bar{x} - x_i)^2 \times \alpha.$$

The displacement, d_{x_q} of an arbitrary point x_q along the rod, is the displacement of the center of mass $d_{\bar{x}}$ plus the displacement due to the rod rotating an angle θ about the center of mass d_r (see Figure 4), so

rod's initial position

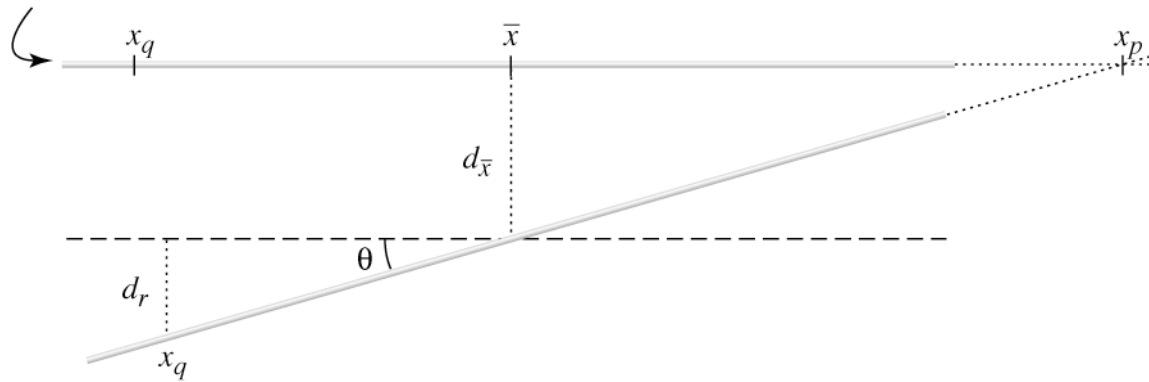


Figure 4. Pivoting massless rod.

$$d_{x_q} = d_r + d_{\bar{x}} \quad \text{and}$$

$$d_r = (\bar{x} - x_q) \sin \theta$$

Using $\theta \approx \sin \theta$ for small values of θ and differentiating twice yields

$$a_{x_q} = (\bar{x} - x_q) \alpha + a_{\bar{x}}.$$

Now, the point where $a_{x_q} = 0$ is the point x_p , which is the point that does not move when force is applied at x_d . At that point,

$$\alpha(x_p - \bar{x}) = a_{\bar{x}}$$

$$x_p = \frac{a_{\bar{x}}}{\alpha} + \bar{x}$$

$$= \frac{F/N}{F(\bar{x} - x_d) / \sum (\bar{x} - x_i)^2} + \bar{x} \quad (\text{by substitution of initial equations})$$

$$= \frac{\sum (\bar{x} - x_i)^2}{N(\bar{x} - x_d)} + \bar{x}$$

which, after some simplification becomes

$$x_p = \frac{\sum x_i^2 - x_d \sum x_i}{\sum x_i - x_d \cdot N},$$

where \bar{x} is the statistical notation for the physical quantity of the center of mass, x_d is the x -value that is independent of the prediction \hat{y}_p , and x_p is the x -value corresponding to \hat{y}_p .

Future Work

Preliminary results show that the result regarding data points that have no influence on predictions holds for models of the form $Y = X\beta + \varepsilon$, which are linear in the unknown coefficients but polynomial equations in X . Exactly k values of \hat{y}_i will be independent of some data point for any linear model that is polynomial in X with power k . This will be the subject of a future paper. Preliminary work has also shown that the result is extendable to the general univariate model that is linear in the unknown coefficients, and for multivariate linear models as well. Future work will extend the results for these cases and compare the estimates of \hat{y}_p for the least squares norm in this special case when data drops out to other methods for estimating \hat{y}_p .

Additionally, future work will explore the “degrees of relevance” as introduced at the end of Example 3. Finally, if we measure distance from \bar{x} in one direction as positive and in the other direction as negative, then the relationship between x_p (an x -data point at which we want to predict y_p) and x_d (the x -data point at which y_d is irrelevant to that prediction) has the

general shape of the function $x_p = -\frac{1}{x_d}$. That is, a data point slightly to the left of \bar{x} will be

irrelevant to a prediction far to the right of \bar{x} , and a data point far to the left of \bar{x} will be

irrelevant for a prediction slightly to the right of \bar{x} . These topics will also be covered in a future paper.

References

- Belsley, D., Kuh, E., & Welsch, R. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.
- California Department of Education, Educational Demographics Unit. District and School Enrollment by Grade, San Jose Unified School District, 2001-2004 [Electronic data retrieval]. Retrieved on September 9, 2005, from <http://www.cde.ca.gov/ds/sd/cb>.
- Chatterjee, S., & Hadi, A. (1988). *Sensitivity Analysis in Linear Regression*. New York: John Wiley and Sons.
- Cook, R.D. (1977). Detection of influential observation in linear regression, *Technometrics*, 42(1), 65–68.
- The Alan Guttmacher Institute. (2004, February 19). U.S. Teenage Pregnancy Statistics With Comparative Statistics For Women Aged 20-24: Notes on Teenage Pregnancy Statistics [Electronic version], Retrieved on September 1, 2005, from http://www.agi-usa.org/pubs/teen_stats.html.