Walden University

SCHOOL OF MANAGEMENT

This is to certify that the doctoral dissertation by

Teresa Bittner

has been found to be complete and satisfactory in all respects,
and that any and all revisions required by
the review committee have been made.

Review Committee
Dr. Kimberly Ross, Committee Chairperson,
Applied Management and Decision Sciences Faculty

Dr. Thomas Spencer, Committee Member,
Applied Management and Decision Sciences Faculty

Dr. Len Haff, Committee Member,
Applied Management and Decision Sciences Faculty

President and Provost

Paula E. Peinovich, Ph.D.

Walden University

2006

Abstract

The Noncontribution of Some Data in Least Squares Regression Predictions

by

T. L. Bittner
M.S., University of California, San Diego, 1985
B.A., University of California, San Diego, 1984

Dissertation submitted in Partial Fulfillment
Of the Requirements for the Degree of
Doctor of Philosophy
AMDS/Operations Research

Walden University
May 2006

Abstract

Estimation via least squares is a mature area of statistics, but a phenomenon that occurs under certain conditions has escaped attention for hundreds of years, and is the focus of this dissertation. This new discovery demonstrates, both graphically and mathematically, the fact that certain conditions cause data points to have no influence on predictions made using ordinary least squares models. Least squares predictions are widely used in many disciplines to make decisions or to determine what may happen in the future. The loss of data when predicting y-values in a linear model 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. Since noncontributory data can be identified before the dependent variable data is even collected, this research can be used as a tool to help statisticians structure their input data more efficiently and analyze existing data with better understanding.

In this dissertation, the mathematical relationships between predictions and data points that are independent of those predictions have been developed and proven for least squares straight-line models, general polynomial models, and general univariate models that are linear in the unknown coefficients. The effect of noncontributory data were analyzed and shown graphically via numerous examples and mathematically in the general form. The important concept of *data wells* was introduced, defined, and examined to demonstrate the far reaching effect of this new discovery on least squares estimation. Data wells show that the phenomenon of noncontributory data is a continuous

rather than a discrete phenomenon, a fact that extends the impact of this discovery dramatically. Finally, recommendations were made regarding future research in least squares sensitivity analysis, including work that will ultimately find a remedy for the phenomenon discussed in this dissertation.

This dissertation provides a foundation for future work in sensitivity analysis, and will help researchers better understand their data both before and after collection. Future research in this area should ultimately result in better predictions, and will have the effect of saving researchers both time and money in their work.

The Noncontribution of Some Data in Least Squares Regression Predictions

by

T. L. Bittner
M.S., University of California, San Diego, 1985
B.A., University of California, San Diego, 1984

Dissertation submitted in Partial Fulfillment
Of the Requirements for the Degree of
Doctor of Philosophy
AMDS/Operations Research

Walden University
May 2006

Acknowledgements

I am indebted to a number of people who have supported me with my doctoral work over the last year. First, I want to thank Dr. Kim Ross, my advisor, mentor, and committee chair. She carefully checked through all of my doctoral work and helped me to make many improvements. Further, I consider her a friend, and hope to have the opportunity to work with her in the future.

I would also like to thank my doctoral committee, Dr. Len Haff and Dr. Tom Spencer for their helpful comments. Dr. Haff was my mentor when I started my graduate program at University of California, San Diego in 1984, and was my inspiration to study mathematics and statistics at the graduate level. Dr. Jim Stahley has also been extremely supportive in helping me reach my goals.

A special acknowledgment goes to Dr. Kurt Norlin for bringing me his “mystery” statistics problem in the winter of 2005 that got me started on this dissertation. His insightful comments and patience throughout my research process has been invaluable.

Jeff Rector deserves thanks for his help formatting this document with hyperlinks, and for his patience in helping me through the many changes I had to make.

Finally, I would like to thank my husband Martin for being there when I had no more energy to go on, my father Larry Becker for his unwavering support and encouragement, and my three children, Rachel, Carolyn, and Eric for their infinite patience and support while I worked through my degree.

Table of Contents

List of Tables

List of Figures

Chapter 1:
Introduction to the Research

Introduction

Least squares modeling is a mature area of statistics. However, sensitivity analysis, specifically the identification of the influence of particular data points on various aspects of the model, has largely been neglected, and is not nearly as well developed as other areas of least squares modeling (Belsley, Kuh, & Welsch, 2004). The identification of influential data points or subsets of data are important because this information can be used to identify sources of collinearity among regression variates. However, an important phenomenon in sensitivity analysis in least squares appears to have been missed altogether. Surprisingly, when making predictions using linear least squares models, some data points have no influence at all on the predictions. Further, this phenomenon occurs in an infinite number of cases, and mathematical relationships can be derived that allow the calculation of exactly which data points do not contribute and which predictions will be affected by the noncontribution of data.

The fact that not all data are necessarily contributing to least squares predictions is important because it is usually assumed that all data are used in such predictions, and predictions may be biased if a data point is independent of prediction calculations, especially for small data sets. Further, the loss of a data point when predicting *y*-values in a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$ indicates 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. In addition, this phenomenon extends beyond straight-line models to other models that are linear in the unknown coefficients.

Background

The amount of literature on sensitivity analysis is very small, especially when compared to the literature concerning least squares in general. Authors such as Belsley, Kuh, and Welsch (2004), Chatterjee and Hadi (1988), Fox (1991), have produced some of the best known works on sensitivity analysis in least squares regression. However, the vast majority of this work is aimed at determining how different data points and subsets of data affect the unknown coefficients of linear models. A standard equation does exist for *hat values*, which are the contribution that the *i*th data point makes to the *j*th prediction (Belsley, et al., 2001; Chatterjee & Hadi, 1988). Cases when a data point makes no contribution can theoretically be computed by setting the equations for hat values equal to zero. In practice, however, hat value equations exist in closed form only for certain simple models, and the small amount of research aimed at determining how the *i*th data point affects the *j*th prediction from a model has been done for the purpose of finding the most influential data points rather than finding ones that have no influence at all on prediction equations. In fact, no references could be found that address the issue of noncontribution of data in prediction calculations, but the research that does exist on sensitivity analysis applied to regression models will be covered thoroughly in chapter 2. In addition, a brief introduction to least squares will be given at the beginning of chapter 2.

Developmental Theory

Consider the linear models of the form $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ and $y_i = \beta_k x_i^k + \beta_{k-1} x_i^{k-1} + \mathcal{L} + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i$, where *k* is an integer greater than one. In these models, $\hat{y}_i$ represents the *i*th predicted value based on one of these models. This research will use mathematical derivation and proof to define the specific conditions under which predictions are independent of some collected data for certain statistical models, and develop the mathematical relationships between the data point(s) that do not contribute to predictions and the predicted *y*-values for which this happens. For this purpose, define $\hat{y}_p$ as the prediction that is independent of at least one data point, and define y_d as the data point that $\hat{y}_p$ does not depend on. x_p and x_d are defined as the *x*-values corresponding to $\hat{y}_p$ and y_d respectively. The *x*-values are the independent variables and the *y*-values are the dependent variables.

In addition to a statistical application that relates $\hat{y}_p$ and y_d, the two values also have exactly the same relationship in an application in physics. The physical application helped to inspire the statistical relationship for the straight-line model. The physical application and relationship between $\hat{y}_p$ and y_d will be described as it relates to the straight-line model, followed by the statistical relationship for the same model. The statistical relationship will also be described for the polynomial model in the *x* data of degree greater than one and the general univariate model. There is no known physical meaning for polynomial models of degree greater than one or any other models. All of the normal assumptions that pertain to ordinary least squares regression and prediction will be assumed in this work.

While this research will show that some data (y_d) contribute nothing to certain predictions ($\hat{y}_p$) made from certain ordinary least squares models, it is also the case that the contribution of those data points (y_d) quickly *approach* zero in the neighborhood of $(\hat{y}_p)$. The noncontributory data points can therefore better be called *data wells* rather than single noncontributory data points. The mathematics and meaning of this phenomenon will be discussed fully in chapter 4.

It is by no means a new discovery that some data points affect least squares models and predictions more than others. However, since the discovery that some data have no influence at all on least squares prediction calculations is new, no previous work exists to build upon. Hence, the first task to be done is to derive relationships between the variables and to prove mathematically that under certain conditions, $\hat{y}_p$ is indeed independent of y_d. It is also necessary to discuss the relevance of these relationships. This derivation and validation for the linear models, polynomial models, and general univariate models linear in the unknown coefficients is the objective of the current research. The derivation and validation of the relationship for the linear model will be developed in chapter 3, and the relationships for the polynomial and general univariate statistical models will be developed in chapter 4.

Summary

This research is not intended to give remedies for problems that might arise as a result of noncontributory data. It is instead meant to give a basis for finding such data. However, it is important to note that there is an error commonly made when making predictions using least squares models. It is common to assume that outliers are a problem for predictions, and so outliers are often deleted from data sets before modeling. In fact, outliers may or may not have much influence over predictions (Belsley, et al., 2004). Instead of the measured data (the *y*-values) determining which data are most or least influential in predictions, it is the number of data points collected (*n*) and the values of the independent variables (*x*-values) that actually determine how much weight any data point or set of points will have for any specific prediction. This fact will be made clear in this dissertation, and it is hoped that it will lead to better techniques in data collection and analysis. In fact, judicious choices about how many data points are collected and which independent variable values are used can allow a researcher to determine which data points will be most and least influential to predictions before data are even collected.

Any theoretical result that shows a potential problem with predictions made through least squares modeling has significance for researchers in many fields. These include but are not limited to business, economics, psychology, sociology, engineering, and astronomy, to name a few. The purpose of this work is to accurately describe the circumstances in which some data will not affect predictions for linear and polynomial models. These are some of the most widely used models used to make predictions, particularly so the linear model, because linear relationships accurately describe the behavior of many real life situations. Examples will also be shown where one or more prediction is independent of a data point. The ramifications of this will be discussed in light of applications. There are many cases where least squares predictions could be adversely affected by this phenomenon. For example, predictions made using small data sets will be adversely affected by the non-influence of a data point, as will applications where data collection is expensive or difficult. In that case, a data point that is not used would be a literal waste of time and money.

Prediction is not the only reason that least squares models are developed, but it is one of the major reasons. When prediction is the goal of modeling, data are often collected that describes what has happened between an independent variable (x) and a dependent variable (y). The resulting model is then used to predict what will happen at some x-value that is not included in the collected data. However, least squares can only appropriately be used to predict data points either between the collected values, or in the neighborhood of collected the known data (Montgomery, et al., 2001). Therefore it will be of particular interest to determine which predictions in that category are independent of some data point that has been collected. For example, given data points from $x = 1$ to n, it will be of interest to see when data points do not contribute when making predictions for $x = n + 1$.

While most past research about hat values has concentrated on influential data, data that do not contribute are important as well. Influential data can tell a researcher that some data are “skewing” the model and/or the predictions when it is not apparent that this is happening. On the other hand, data that lend low or no contribution to predictions can appear to be helping pull a prediction towards a central value, while actually it has little or no effect at all. Either way, it is important for researchers to have this information. Remedies for this could range from fixing the problem before data are collected to adjusting something afterward. However, the first necessity is to know which data are influential and non-influential to predictions.

Integer values of x_p are of special significance because predictions are so often made using integer values of the independent variable. Therefore, special consideration will be given to finding relationships between y_d and y_p where the *x*-value corresponding to y_d is whole number. One particular value of interest is the prediction $\hat{y}_{n+1}$ when *n* data values are collected and modeled. *x*-values in collected data often represent time, and the (*n* + 1)st prediction would then represent what is predicted to happen at the next incremental future time period. For example, if the *x*-values 1 through *n* represent what has happened for the last *n* years, then the (*n* + 1)st prediction represents what is predicted to happen next year.

Interestingly, data that do not contribute to prediction calculations are not limited to integer values of x_p. The derivations of these relationships will be developed in chapter 4 for integer and real values of x_p for the straight-line model, and for real values of x_p for the polynomial and general univariate models.

This dissertation will include the derivations regarding relationships between $\hat{y}_p$ and y_d for straight-line model, higher order polynomial models, and the general case of the univariate linear model in the unknown coefficients. Though a straight line is a special case of a polynomial model where the degree is one, the case for straight-line models will be handled separately from higher order polynomial models. The straight-line model is the most widely used in applications, and therefore has special significance. The other reason for the separate handling of lines and higher order polynomial models is due to the fact that the mathematics and the relationship that exists between $\hat{y}_p$ and y_d is elegant for straight-line models, and can be expressed easily in closed form. In other words, a mathematical formula relating the two values can be written. Further, this formula depends only on the *x*-data and the number of data points (*n*). On the other hand, the relationship for polynomial models cannot be expressed easily in closed form and therefore must be described as a solution *process* rather than a set of simple equations.

While this research will include the development of relationships for straight-line models, higher order polynomial models, and the general univariate model that is linear in the unknown coefficients, it will not include derivations of the relationship between $\hat{y}_p$ and y_d for multivariate models that are linear in the unknown coefficients, or any models that are nonlinear in the unknown coefficients. These additional models will be discussed in chapter 5 as opportunities for future research.

A process will be derived in chapter 4 that describes the relationship regarding noncontributory data and their corresponding predictions for the general case of the univariate linear model, but exhaustive analysis of the implications of these relationships is beyond the scope of this research. However, the process will allow these relationships to be derived for any specific univariate linear model and corresponding data set. Suggestions for future research to be done will be described in chapter 5, both for describing further relationships between predictions and noncontributory data, and for examining the implications of these relationships.

The relationship between $\hat{y}_p$ and y_d for the model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ can be readily found by setting the “hat value” equation for this model equal to zero. However, the derivations for the polynomial and general univariate linear models are much more complex, and the straight-line relationship will be derived from the beginning in order to establish the methodology for the derivation of the relationship for the more general models.

It is hoped that this work will be the beginning of a study of some of the hidden limitations of least squares prediction. As a result, researchers will be able to determine in advance, at least to some extent, whether or not a given data point will actually contribute to predictions that need to be made. Depending on the specific situation, this may influence what data are collected or determine that a prediction technique other that least squares should be used. This should eventually lead to better overall predictions and a better knowledge of how collected data are used in prediction calculations.

This research is the beginning of work studying data that do not contribute to predictions made with ordinary least squares. This is especially important when making predictions using small data sets or in cases where data collection is expensive or difficult. Chapter 2 of this dissertation gives a brief introduction to least squares and reviews the limited literature that is available concerning the contribution that individual data points make to predictions using ordinary least squares modeling. Chapter 3 examines the physics application that helped to motivate the statistical problem, and derives and proves the statistical relationship between predictions and noncontributory data for the straight-line model. Chapter 4 extends this work to general polynomial models in the *x*-data as well as to the general univariate model that is linear in the unknown coefficients**.** The concept of *data wells* that was introduced earlier in this chapter will also be examined more fully in chapter 4. Chapter 5 will discuss conclusions, a call for action, and recommendations for future work on the topic.

Chapter 2:
Literature Review

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. A description of the technique was first published in 1805 by the French mathematician Legendre, but the German mathematician Gauss later claimed that he had been using it for years before Legendre's publication (Plackett, 1972). Like the dispute between Leibniz and Newton over the invention of The Calculus, an argument between Gauss and Legendre followed Legrendre's 1805 publication. (Plackett, 1972). Despite the problems surrounding its invention, least squares modeling has now been used for over 200 years and has proven to be one of the most useful and well known techniques in statistics (Plackett, 1972).

Gauss's technique began as a method that is now well known to solve *k* linear equations when there were *k* unknown variables (Farebrother, 1988). Regression analysis was eventually developed by Gauss to solve for *k* variables when there were more than *k* equations (or data points) available (Farebrother, 1988). Solving for *k* variables when there are more than *k* data points is called solving an *overdetermined system of equations.* Gauss's regression analysis is a statistical technique used to model the relationship between variables. This is useful in many areas of social science, business, physical science, engineering, and many other fields. A few examples of fields that use least squares modeling are psychology, finance, biology, and systems engineering. However, the ways in which regression analysis can be used are virtually limitless. Accurate predictions have become a vital need in our society in order to maintain and increase the efficiency that businesses and other organizations have come to expect and depend upon as part of their daily operations.

Least squares regression is the use of the least squares method to fit a model to data. Most often the data are measured or observed from some real world phenomenon. Least squares is one of many possible *norms* in that it is a particular way to measure optimality of a model. The method of least squares is the technique of minimizing the sum of the squares of the difference between the model and the measured data points. Unless otherwise specified, it should be assumed in this research that regression means least squares regression.

Uses of Regression

Regression has multiple uses. These include data description, parameter estimation, prediction and estimation of dependent (response) variables, and the control of one variable by varying another (Montgomery, et al., 2001). All of these uses include the development of a model to describe the relationship between two or more variables. For example, equations are often used to describe relationships between variables. Once available data are fitted to a function, the function is a convenient and efficient way to describe the relationship between the variables. Parameters can also sometimes be estimated using regression. For example, the well-known equation $s(t) = s_0 + v_0 t + 4.9t^2$ describes the position of a falling object at time *t* given the initial position s_0 and the initial velocity v_0. If the initial position and velocity are unknown, random measurement error is assumed, and several measurements of time and position are taken, the parameters s_0 and v_0 can be estimated using regression. Prediction is one of the most common uses for regression. Many researchers and others regularly collect data using two or more variables, fit it to a model believed to be accurate, and then predict values of one of the variables for some future time or point for which the response variable is unknown. Regression can also be used to control one variable by manipulating another. For example, Montgomery (2001) uses an example of a chemical engineer that wants to control the tensile strength of paper using the hardwood concentration in the pulp. The engineer could develop a model relating the two variables and then use the resulting function to change the hardwood concentration until the desired tensile strength is reached.

It is important to note that a cause and effect relationship is not necessarily required if a model is developed using regression and is only going to be used for prediction purposes (Montgomery, et al., 2001). The only requirement in this case is that the relationship between the variables that existed when the data were collected still exists when the predictions are made using the model. In this case, causation is not required and cannot be assumed. It also must be stressed that a model is only as good as the data that created it. If the data have large or nonrandom measurement errors or are otherwise invalid, the model generated from the data will likewise be invalid. Further, if the relationship between the variables is not the same as the model fitted to the data, then the model will do a poor job predicting and estimating other data points or parameters. It is therefore imperative to verify that the data are likely to be related in the way that it is being modeled.

Regression is also often misused by attempting to predict values far outside the range of the independent variable(s) (regressor variables). Regression models are best used to interpolate values in between the range of the independent variables. The further outside the range of the regressor variables an extrapolation, the more careful one must be in using the resulting prediction (Montgomery, et al., 2001). It will also become clear later in this dissertation that some data affect models more than others. For example, very different models are obtained if the data at the ends of the data set are eliminated or changed. In fact, it will be shown in chapters 3 and 4 that there are data points that don't affect certain predictions at all.

Measured data that falls far outside the pattern shown by the rest of the data are called *outliers.* Outliers can cause serious problems with regression models because the outliers generally have a much stronger effect on the model than the statistician would desire. Outliers must be examined carefully before a decision is made whether or not to include them in a model, and they are often deleted from the data set before fitting the data to any model.

While exhaustive coverage of the cautions and limitations of regression would require hundreds of pages and therefore cannot be thoroughly covered in this work, it should also be noted that particular models should be fitted to data with caution. For example, linear models often do not properly describe the relationship between variables. Before selecting a particular model, data must be plotted and visually examined, and certain simple tests should be run to insure a minimum level of fit. The computation of a correlation coefficient between two variables is a minimum standard that should be used before a linear model is applied to data to be used for prediction or any other purpose.

However, even if computations indicate a strong correlation between variables, it cannot be assumed that the one variable actually causes the other. While causation necessitates high correlation, the reverse is not necessarily true. Therefore, regression is not intended as a tool to determine causation.

The following section describes the most widely used of regression models, the simple linear model. It is important to keep all the aforementioned cautions in mind when reading the rest of this dissertation, as the cautions apply to all of the models discussed and are an important part of good modeling.

The Simple Linear Model

A model with a single independent variable (regressor), that has a linear relationship with the dependent variable (the response variable) can be modeled as a straight line where the slope and intercept are "fitted" to the data to minimize the sum of the squared errors. The *errors* in this case are the perpendicular (shortest) distance between the model and the data point corresponding to it.

The simple linear model is generally denoted as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } i = 1, 2, \ldots, n$$

where the x_i's are the independent variables, the y_i's are the measured values corresponding to the x_i's, β_1 is the slope of the model, β_0 is the *y*-intercept of the model, and the ε_i's are random errors. The errors are assumed to be random measurement errors with a mean of 0 and an unknown variance of σ^2. Further, the errors are assumed to be independent of one another in that the value of any one error does not depend on any of the others. The errors are assumed to be errors on the response variable y. The parameters β_0 and β_1 are generally called the *regression coefficients* and are unknown. The sample data values $(x_i,\ y_i)$ are used to estimate the regression coefficients. The calculations used to do this are what is generally called a *regression* (Farebrother, 1988; Montgomery, et al., 2001; Younger, 1979).

The basis for finding the best β_0 and β_1 in the least squares sense is to minimize the sum of the squared errors. The error is defined as the difference between the measured data point and the model evaluated at that data point, or $y_i - \beta_0 - \beta_1 x_i$. Define $L(\beta_0, \beta_1)$ as the function denoting the sum of the squared errors. Then for the simple linear model this can be written as

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

To minimize this function simple calculus can be used by taking the partial derivatives of L with respect to β_0 and β_1, and then setting them equal to 0. The regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\ \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad (1)$$

and

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\ \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 . \qquad (2)$$

Simplification of equation (1) yields

$$-2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\hat{\beta}_0$$

$$\bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0$$

Similarly, equation (2) can be simplified as follows:

$$-2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - n\hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

Then, when $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ is substituted into the equation above, the equation can be simplified to yield

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i x_i - \frac{\overline{xy}}{n}}{\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2} \\ &= \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \\ &= \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Therefore, the fitted model is then

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad (3)$$

which allows an estimate of y for any particular value of x (Farebrother, 1988; Montgomery, et al., 2001; Younger, 1979).

Multiple Linear Regression

Often data are not linearly related, or there is more than one regressor variable. In these cases, there is a generalized theory for least squares fit. While there are methods available to deal with many different types of models, this dissertation will only deal with models that are linear in the unknown coefficients. This does not eliminate models that are nonlinear in the x-data, and some models that are nonlinear in the unknown coefficients can also be transformed to be linear models. To begin, start with the model $Y = X\beta + \varepsilon$, where in this case the X matrix is an $n \times p$ matrix which helps form the model. For example, to form a model of the form $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i{}^2 + \varepsilon_i$ that is linear in β but quadratic in the x_i's, the X $(n \times 3)$ matrix would take the form

$$X = \begin{bmatrix} 1 & x_1 & x_1{}^2 \\ 1 & x_2 & x_2{}^2 \\ \mathcal{M} & \mathcal{M} & \mathcal{M} \\ 1 & x_n & x_n{}^2 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

A model of the form $y_i = \beta_0 + \beta_1 x_{i1} + \mathcal{L} + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ is represented in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \mathcal{M} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \mathcal{L} & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \mathcal{L} & x_{2,p-1} \\ \mathcal{M} & \mathcal{M} & \mathcal{M} & \mathcal{M} & \mathcal{M} \\ 1 & x_{n1} & x_{n2} & \mathcal{L} & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \mathcal{M} \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathcal{M} \\ \varepsilon_n \end{bmatrix}$$

Solving this equation for the unknown coefficients once again requires writing the function that represents the sum of the squared errors. In this case it is a matrix computation, and can be represented as

$$L(\beta) = \sum_{i=1}^{n} \varepsilon_i{}^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

Expansion of the function yields

$$\begin{aligned} L(\beta) &= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \end{aligned}$$

since $\beta'X'Y$ is a scalar value. Just as in the case of the simple linear model, calculus is used to find the minimum value of $L(\beta)$.

$$\left.\frac{\partial L}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

When solved, this becomes

$$X'X\hat{\beta} = X'Y$$

which can be solved for $\hat{\beta}$ by multiplying both sides of the equation by $(X'X)^{-1}$ on the left to obtain

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

Therefore, the least squares solution for any model that is linear in its unknown coefficients can be obtained with the above matrix computation.

It is sometimes possible to transform a model that is nonlinear in its unknown coefficients into a model that is linear in its unknown coefficients and therefore can be solved in closed form. For example, consider the model

$$y_i = \log(\beta_1)x_i + \beta_0 + \varepsilon_i$$

This model is clearly nonlinear in the unknown coefficient β_1. However, a simple transformation $\beta_1^* = \log(\beta_1)$ will make this model into a model linear in all of the unknown coefficients. Namely,

$$y_i = \beta_1^* x_i + \beta_0 + \varepsilon_i$$

which can easily be solved using the standard least squares solution for $\hat{\beta}$. Models that are nonlinear in the unknown coefficients and nonlinearizable must be solved using nonlinear techniques. The methods for solving nonlinear problems are generally iterative.

Alternative Norms

Two alternative methods to least squares involve minimizing the sum of the absolute values of the errors or minimizing the maximum error. Both of these methods precede Gauss's discovery of least squares regression in 1794 or 1795 (Plackett, 1972). The method of minimizing the sum of the absolute values of the errors goes back to Boscovich in the eighteenth century. This is now known as the L_1, or *absolute value norm*, and predates Laplace's 1783 discovery of a method to minimize the maximum error, known now as the *minimax norm* (Plackett, 1972). This method is based upon minimizing a function of the residuals that takes the form

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \rho(e_i) = \underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \rho(y_i - x_i'\beta)$$

where x_i' is the *i*th row of the *X* matrix.

Unfortunately, even if the errors (ε_i) are multiplied by a constant, the optimal answer may be different from the original when using the L_1 norm. This nonscalability has to be corrected for by finding a robust scaling constant. However, the biggest problem with the L_1 norm is the fact that it must be found by iterative means since no closed form solution exists. In general, one takes the first partial derivatives of ρ with respect to β_j $(j = 0,\ 1,\ \mathcal{L}\ ,\ k)$ of the equation to be minimized and sets them each equal to zero. This yields a system of $p = k + 1$ equations, which are often nonlinear and is either solved using nonlinear iterative techniques or by using iteratively reweighted least squares (Beaton & Tukey, 1974).

Another norm that is sometimes used in lieu of the least squares norm is the *Minimax Norm.* The implementation of the minimax norm, sometimes called the L_∞ norm, consists of fitting data to a particular function such that the maximum error between the model and the data is minimized. This technique would be appropriate when, for example, the concern is that no error be greater than a certain threshold. Minimizing the maximum error will by definition result in the sum of the squared errors (the least squares norm) being higher than it would be if least squares regression were used to fit the data to the model. It also normally results in all the errors being fairly close to the maximum error. Therefore, it is not appropriate to compare techniques using different norms in the sense of which is the “best” answer. By definition, each norm provides the best answer to the specific function it is minimizing. One final note about the minimax norm is that the techniques used to fit data using this norm are iterative like the L_1 norm techniques. The lack of closed form solutions makes the L_1 and L_∞ norms less convenient and practical to use than least squares. Further, the least squares norm is mathematically and practically more elegant than the alternative norms. Two reasons are that the linear least squares technique provides a closed form solution for the unknown coefficients and these solutions are unbiased estimates of the true parameters. These and many other convenient mathematical properties make least squares the most used norm for prediction and modeling.

Sensitivity analysis in Least Squares Regression

This research concerns the fact that some least squares predictions are independent of some data point(s) in linear models, polynomial models, and general univariate models. The literature that relates to this research is in the general field of sensitivity analysis in regression, and the specific topic is the determination of how the *i*th data point affects the *j*th prediction. The amount of literature on sensitivity analysis overall is small, particularly when compared to the literature on least squares and regression in general. Belsley, et al. (2004) commented that sensitivity analysis has largely been ignored as an area of research in statistics. This is not because the field is unimportant. To the contrary, good sensitivity analysis techniques can help analysts to find data points and subsets of data that are most and least influential to both the model parameters and to predictions. This is an efficient way to find sources of collinearity in data and ultimately to remedy other potential problems in data sets (Belsley, et al. (2004); Chatterjee & Hadi, 1988).

A typical method of analysis when computing predictions using least squares models is to simply delete outliers from the data set and then model without them. This is often done because it is assumed that observations that fall far from the pattern that the rest of the data imply are somehow "invalid" data values, and also because it is assumed that they will have a profound effect on predictions if incorporated into the model (Belsley, et al., 2004). Both of these assumptions are often wrong. While outliers can be measurement errors or some other kind of anomaly that does not properly describe the relationship between independent and dependent variables, sometimes outliers are valid data points that lend valuable information to the model (Belsley, et al., 2004; Montgomery, Peck, & Vining, 2001). Further, no matter what the cause, outliers may not always have a large effect on predictions in any case (Belsley, et al., 2004). Sensitivity analysis has not provided all the remedies to deal with issues such as outliers and collinearity, but relationships are continuously being developed that will eventually lead to such remedies (Belsley, et al., 2004). There are many examples of such work (Beckman, R. J., 1990; Chave & Thompson, 2003; Cook, 1977; Cook & Weisberg, 1980; Johnson, 1985).

The relationships that have generally been developed to date have involved the detection of influential observations in a model (Cook, 1977; Johnson, 1985; Johnson & Geisser, 1983; Thomas & Cook, 1990). There are various measures of influence that are used, but perhaps one of the best known is a measure based on confidence ellipsoids that was developed by R. Dennis Cook (1977).

To measure the influence of an observation using *Cook's Distance* for a multivariate linear model, a joint confidence region is formed for the unknown coefficients β. Under the assumption of normality, these $100(1-\alpha)\%$ joint regions are ellipsoids centered at $\hat{\beta}$, and given by

$$\frac{(\beta-\hat{\beta})(X'X)(\beta-\hat{\beta})}{k\hat{\sigma}^2} \le F_{(\alpha;k,n-k)},$$

where $F_{(\alpha;k,n-k)}$ is the upper α point of the central *F* distribution with *k* and (*n* – *k*) degrees of freedom. Then the influence of the *i*th observation is measured by computing the change in the center of the confidence ellipsoid when the *i*th observation is removed. This distance can be thought of as the scaled distance between the least squares estimate of the unknown coefficient $\hat{\beta}$ and $\hat{\beta}_{(i)}$, the same estimate with the *i*th observation removed (Chatterjee & Hadi, 1988).

Influential observations can potentially yield a lot of information about a model. Besides helping to detect collinearity, sometimes a small subset of data exerts a disproportionate amount of influence and a regression model can change greatly based on just a few points (Montgomery, et al., 2001). It can therefore be very important to identify such data or subsets of data. Once identified, a variety of measures can be taken to deal with the problems. Currently, there are two specific types of corrective measures that are most often used. The first type are Bayes-like methods that use prior information to correct for collinearities, and the second employes the collection and introduction of additional data to provide the independent variation necessary to correct for collinearities.

For the model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$, a standard formula has been developed that describes the contribution that the *i*th data point has on the *j*th predicted value. These relationships are known as *hat values*, and are described by the notation h_{ij}. This term is attributed to John W. Tukey, and the mathematics were apparently derived in or about 1967 (Hoaglin & Welsh, 1977).

The relationship between the *i*th data point and the *j*th predicted value for the model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ is

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2},$$

where *n* is the number of data points in the sample, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the well known sample mean). This formula is generally used to help in finding influential data points in a data set (Hoaglin & Welsch, 1977). Since h_{ij} indicates the contribution of the *i*th data point to the *j*th prediction, high hat values often indicate influential data. However, to find precisely which data points are highly influential and the degree of influence, hat values are generally examined along with residuals and variance of the data.

In general, hat values are part of a matrix representation of the influence of the *i*th data point on the *j*th prediction, where h_{ij} are the elements of a matrix *H* that is given by $H = X(X'X)^{-1}X'$ (Belsley, et al., 2004). Though these values can theoretically be expressed in closed form as in the formula for h_{ij} related to the straight-line model, this becomes impractical in practice because the relationships are extremely complex. Therefore hat values have traditionally been computed numerically for specific models and data sets (Belsley, et al., 2004; Chatterjee & Hadi, 1988). This is quite possibly one of the reasons that the phenomenon of noncontributory data was not discovered earlier.

In this work, however, hat values will be used to find data that have no influence on the model. For the linear model described above, the case when a data point makes no contribution can easily be computed by setting the equation for the individual components of the hat matrix equal to zero. Note that the entire hat matrix *H* cannot be set equal to zero. In general, however, each model has its own hat value relationships, and most are not easily expressed in closed form in terms of the *x* values and *n*. Hence the problem at hand cannot be solved by setting a closed form expression equal to zero and solving it to find the desired relationships.

While the literature does address the contribution of the *i*th data point on the *j*th predicted value, no references could be found that address the issue of noncontribution of data in prediction calculations. Further, only one reference could be found regarding small hat values. Belsley, et al. (2004) said that a large coefficient change in a model in the presence of small values of h_i may be more important to the structure of a model than to predictions from that model. However, this comment misses the point that data that do not contribute to a predicted value may appear to a researcher to be helping to pull the prediction towards a central value, while in fact it is having no effect at all. This is a particular problem for small data sets, and when a particular subset of data points is deemed to be more important or more accurate than other data. If some of this data are not contributing to a predicted value, it is important that the researcher be aware of this fact. The issue also arises that there may be no reason to collect a particular data value if it is not going to be used anyway. A dummy point will do just as well.

With this past literature in mind, chapter 3 will proceed to develop the relationship between y_d and $\hat{y}_p$ for the linear model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. This will be accomplished by briefly examining an application from physics that has a relationship that is analogous to the statistical phenomenon to be examined. The theoretical background in least squares that leads to this relationship will also be examined, ultimately leading to a description between the key variables under study.

Chapter 3:
The Linear Case

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. In chapters 1 and 2 it was discussed that most previous work in the field of sensitivity analysis has concentrated on finding data that have a large influence on statistical models or predictions. In this section a new finding regarding least squares is developed. This involves the fact that certain conditions cause data points to have no influence at all on predictions for particular $\hat{y}_i$ values. This finding is important because it is usually assumed that all data are being used in such predictions, and results can be skewed if data points are not contributing to certain predictions, especially for small values of *n* or in cases where data collection is expensive or difficult. It seems clear that the loss of a data point when predicting *y*-values in a linear model 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. While the effect of noncontributory data on predictions must someday be determined systematically, the first step is to define the relationships between data that have little influence and the predictions those points affect.

The finding regarding noncontributory data applies to straight-line models and extends to other models that are linear in the unknown coefficients. This chapter describes the specific conditions under which the *d*th observation has no influence at all on particular $\hat{y}_p$ predictions for straight-line models. This chapter derives the relationships between the data points that have no influence and the predicted *y*-values for which this happens in straight-line models. Higher order polynomial models and general univariate models that are linear in the unknown coefficients will be covered in chapter 4. A physical application of this phenomenon is also discussed for the straight-line model, and examples are shown to illustrate the phenomenon and its importance in statistical modeling.

Notation and Assumptions

Consider the linear model $\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$, where the ε_i values are assumed to have a Normal distribution with mean 0 and unknown variance σ^2. Additionally, the errors are assumed to be uncorrelated. *X* is an $n \times p$ “design” matrix. The *x*-values in this model are the independent variables, while the *Y* vector contains the observed values. Because the errors are uncorrelated, the observations are also uncorrelated. The $p \times 1$ β vector contains the regression coefficient, or unknown values. It is assumed in this work that all models are univariate and linear in the unknown coefficients.

An example of a linear model of the form $Y = X\beta + \varepsilon$ is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. In this example, the $n \times 3$ *X* matrix is

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \mathcal{M} & \mathcal{M} & \mathcal{M} \\ 1 & x_{n-1} & x_{n-1}^2 \\ 1 & x_n & x_n^2 \end{bmatrix},$$

the β vector and *Y* vector are

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \\ \mathscr{M} \\ y_{n-1} \\ y_n \end{bmatrix}$$

respectively.

As seen in the example above, a model that is linear in the unknown coefficients is not necessarily linear in the independent variable. Another example of a model linear in the unknown coefficients is $y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$. In this case the design matrix is

$$X = \begin{bmatrix} 1 & \log x_1 \\ 1 & \log x_2 \\ \mathscr{M} & \mathscr{M} \\ 1 & \log x_n \end{bmatrix}.$$

The individual data points 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.

It will be shown that the measure of influence that an observed value has on a prediction is determined by the independent variables (*x*-values) and the value of *n* rather than the dependent variables (*y*-values). For this reason, the relationship between a data point that has no influence on a prediction, and the prediction for which this happens will be given as a relationship between x_d and x_p.

For linear models with the assumptions given, the method of least squares can be used to estimate the unknown parameters. The famous solution to this problem is

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

where the symbol $\hat{\beta}$ denotes the best fit for β using the least squares norm. In other words, the sum of the squared errors between the observed data and the fitted values of the model is minimized. The predicted values using the least squares estimators is given by

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

For the purposes of derivations performed in this chapter, define I_d as the indicator function at *d*. In other words,

$$I_d = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where the vector contains zeros in every position and a “1” in the *d*th position. Let I_{x_d} denote the same indicator function, except the “1” in the *d*th position is replaced by an x_d in the same position.

This chapter will describe the mathematics behind the phenomenon of non-influential data points as they relate to predictions. Since there is an application from physics that helped inspire the discovery of the statistical phenomenon being studied, the physical relationship between x_d and x_p that is analogous to the statistical relationship between the variables will be discussed and derived before the statistical development is presented. Following this, the statistical relationships between x_d and x_p will be derived for the special case of the straight-line models where the *x*-values are evenly spaced and the value of x_p corresponding to $\hat{y}_p$ is an integer. Then the relationship between x_d and x_p will be derived for the general linear model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the *x*-values are not necessarily equally spaced or integer valued. A theorem regarding this relationship will also be stated and proven. Next, the statistical relationship between x_d and x_p will be derived for the general polynomial model of the form $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_{k-1} x_i^{k-1} + \beta_k x_i^k + \varepsilon_i$ in chapter 4. Finally, the case of the general univariate linear model will be addressed at the end of chapter 4. While the relationship between x_d and x_p can be elegantly expressed in closed form for the straight-line model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the relationship cannot be easily expressed in closed form for polynomials of degree greater than one or for the general univariate linear model. Hence, the relationship will be derived in the form of processes rather than equations, and examples will be shown to illustrate the methods. The implications of this work and recommendations for action and future work will be discussed in chapter 5.

A Physical Application

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 1 below.

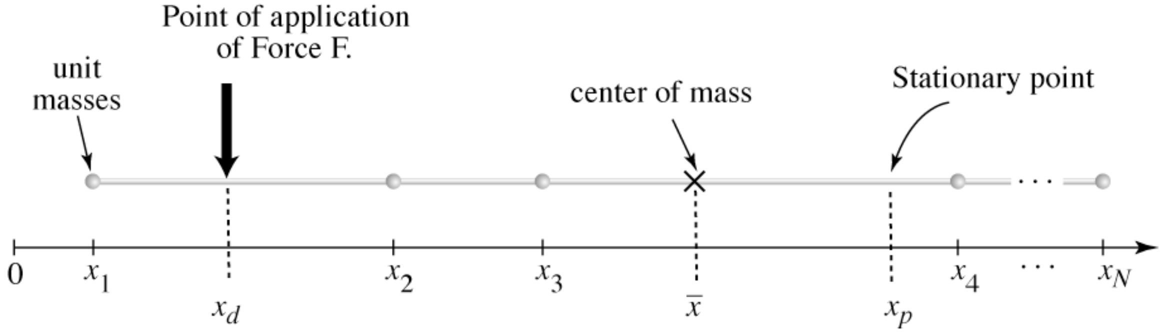

Figure 1. Point masses joined by a massless rod to form a single rigid body.

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 2 below). The relationship between x_d and x_p can be derived from simple physical relationships.

We know from basic physics that

$F = ma$, where *F* is force, *m* is mass, and *a* 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 massless rod,

$ma = Na_{\overline{x}}$ ($a_{\overline{x}}$ = acceleration of the center of mass, *N* 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 2), so

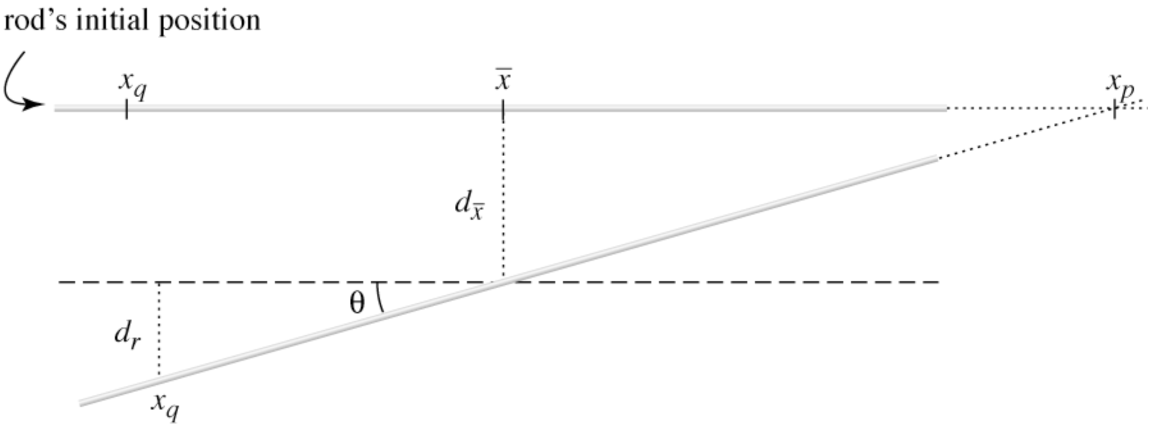

Figure 2. Pivoting massless rod.

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

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

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,

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

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

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

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

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 $\overline{x}$ 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 $\hat{y}_p$.

This application from physics helped inspire the work that follows and develops an analogous statistical relationship to the physical application just described. To look briefly ahead at the analogous statistical phenomenon, consider that the point at which the force is applied is the analogue of the *x*-coordinate of the point that would not affect $\hat{y}_p$, the point called x_d. The point that remains stationary is the analogue of the *x*-coordinate corresponding to the point $\hat{y}_p$, the point called x_p. In fact, it will be found later in this chapter that the relationship between x_p and x_d is exactly the same in the statistical realm as it is in the physical realm. The relationship between the physics application and the statistical one stems from the fact that the rod seeks a position in which the total kinetic energy is minimized. The kinetic energy is directly proportional to the square of the displacement so that minimizing the kinetic energy is the same as minimizing the sum of squares.

However, in statistics the application is that some data are not contributing to particular predictions when modeling with least squares. The fact that these two different disciplines yield an identical relationship is interesting, but the focus throughout the rest of this dissertation will be on the statistical relationships between predictions and data points that make no contribution to those predictions.

The Phenomenon of Noninfluential Data Points in Least Squares Predictions

The analysis of this statistical 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, beginning with the special case where the *x*-values are evenly spaced, and followed by a full analysis of the general case for the straight-line model. The case of the second order polynomial model is then analyzed in chapter 4, and this is followed by a derivation of the relationship between x_d and x_p for the general polynomial model of degree *k*. Finally, the derivation of the relationship is performed for the general univariate model that is linear in the unknown coefficients. Some of the impact of this discovery will also be discussed in chapter 4. This includes the concept of *data wells*, first described in chapter 1. The examination begins with a simple example that shows how the second data point has no effect on the fifth prediction in a straight-line model with four observations.

Example 1.

The enrollment in Kindergarten at Allen Elementary School in San Jose, California for the last 4 years is as follows: (2001, 62), (2002, 43), (2003, 78), (2004, 82) (CA Department of Education, 2005). The data are plotted in Figure 3 below along with the least squares regression line.

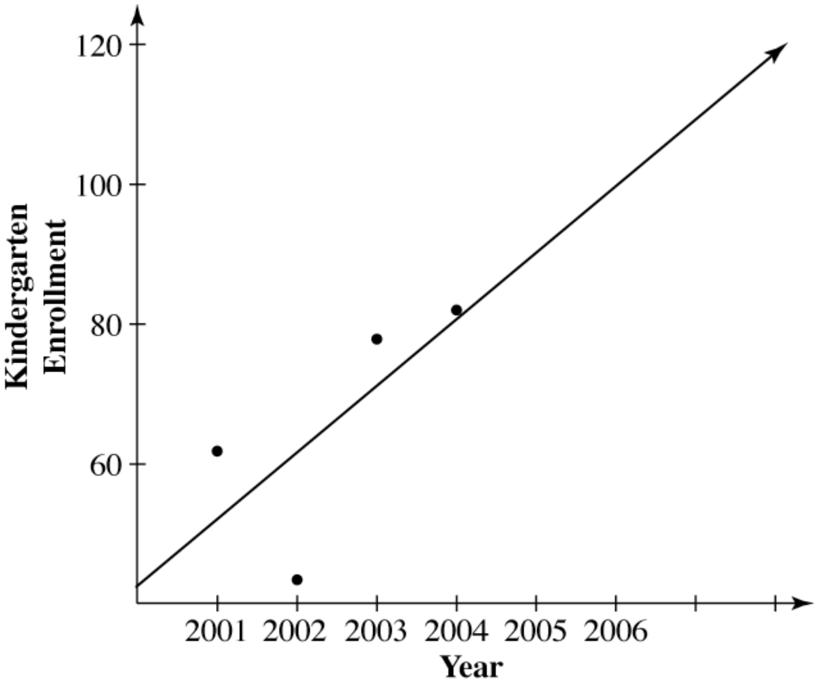

Figure 3. Allen Elementary School kindergarten enrollment from 2001 to 2004.

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 $\hat{y}_i = X\hat{\beta} = X(X'X)^{-1}X'Y = 42.5 + 9.5x_i$. 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 $y_2 = 43$. 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 below on the same set of axes along with the original data.

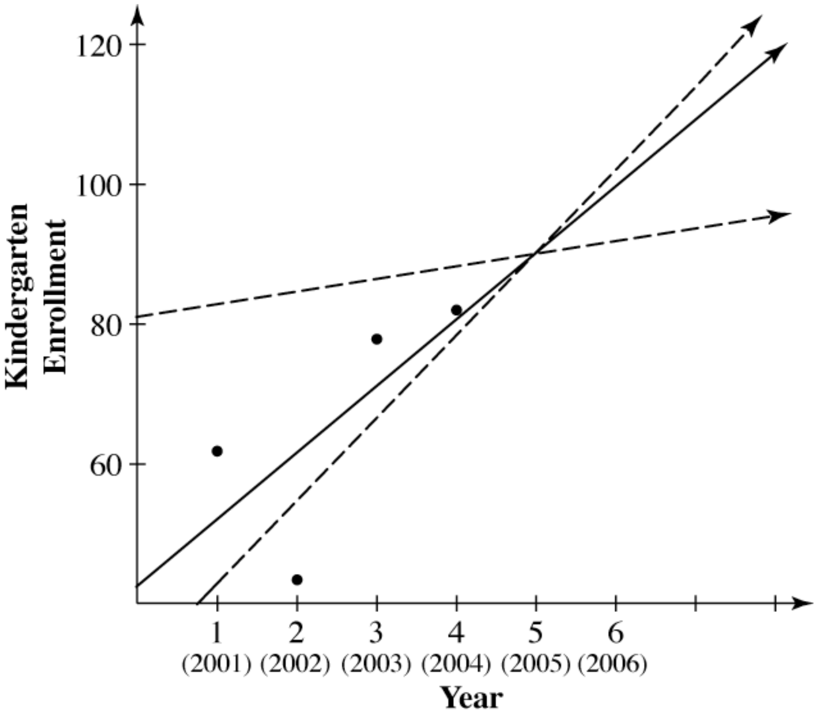

Figure 4. Allen Elementary School kindergarten enrollment prediction lines.

It is easy to see in Figure 4 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 the data point that has no influence on a prediction and the predicted value for which the data point is independent.

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

$x_1,\ x_2, \ldots\ x_n$ and $y_1,\ y_2, \ldots\ y_n$ where the *x*-values and *y*-values are unrestricted real values. Then assume there is a linear model in *x* such that $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. This is often represented as

$$\underset{n \times 1}{Y} = \underset{n \times 2}{X} \underset{2 \times 1}{\beta} + \underset{n \times 1}{\varepsilon},$$

or, in matrix form

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

The well established least squares solution for this model is

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

As noted in chapter 2, an expression for hat values has been derived for this model. The contribution that the *d*th data point makes to the *p*th prediction is given by

$$h_{dp} = \frac{1}{n} + \frac{(x_d - \bar{x})(x_p - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(Belsley, et al., 2004). Note that this expression is dependent only on the independent variables (*x*-values), and the number of observed values collected (*n*). The observations (*y*-values) do not affect the amount of influence that any given data point has on predictions.

When $h_{dp} = 0$, this means that the *d*th observation y_d has no effect on the *p*th predicted value $\hat{y}_p$. Therefore, If the expression for h_{dp} is set equal to zero and solved for x_p, the result is the value of x_d corresponding to observed value y_d that has no influence at all on the predicted value $\hat{y}_p$.

$$\frac{1}{n} + \frac{(x_d - \bar{x})(x_p - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$$

$$\frac{(x_d - \bar{x})(x_p - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = -\frac{1}{n}$$

$$x_p - \bar{x} = -\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(x_d - \bar{x})}$$

$$x_p = -\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(x_d - \bar{x})} + \bar{x}$$

Though it is clear that the relationship between x_d and x_p for the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ can be easily computed in closed form as was done above, the derivation of this relationship is important in order to develop the relationships between these variables for other models. Therefore, the derivation of the above relationship will be derived from its theoretical beginning in this section, and then the technique used also be employed to find a process for finding the relationship for other univariate linear models.

The formal derivation of the relationship between x_d and x_p begins with the matrix representation of the estimated regressor variables for the linear model of the form

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ as described above. To clarify, the well-known matrix form of the solution for $\hat{\beta}$ is developed below:

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \mathcal{M} & \mathcal{M} \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \mathcal{M} \\ y_n \end{bmatrix}$$

$$= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i{}^2 \end{pmatrix}^{-1} \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \mathcal{M} \\ y_n \end{bmatrix}$$

$$= \frac{1}{n\sum x_i{}^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i{}^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\sum x_i{}^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i{}^2 \sum y_i - \sum x_i y_i \sum x_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix} \tag{1}$$

This result can then be simplified to obtain a more convenient form.

First recall that $\bar{x} = \frac{1}{n}\sum x_i$.

Therefore,

$$X'X = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i{}^2 \end{bmatrix}, \quad (X'X)^{-1} = \frac{1}{\sum (x_i - \bar{x})^2} \begin{bmatrix} \frac{1}{n}\sum x_i{}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}, \text{ and } \quad X'Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Some simplification yields the more well-known form

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

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 $i = 1,\ 2, \mathcal{L}$ 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

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

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

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

Since the transformation only results in a scalar transformation of $x_i - \overline{x}$, and it was previously shown that a scalar multiple of $x_i - \overline{x}$ 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

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

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 yields

$$\begin{aligned}
\hat{y}_5 &= \frac{3}{2}\sum y_i - \frac{1}{2}\sum x_i y_i + 5\left[\frac{1}{5}\sum x_i y_i - \frac{1}{2}\sum y_i\right] \\
&= \frac{3}{2}\sum y_i - \frac{1}{2}\sum x_i y_i + \sum x_i y_i - \frac{5}{2}\sum y_i \\
&= -\sum y_i + \frac{1}{2}\sum x_i y_i \\
&= -(y_1 + y_2 + y_3 + y_4) + \frac{1}{2}(y_1 + 2y_2 + 3y_3 + 4y_4) \\
&= -\frac{1}{2}y_1 + \frac{1}{2}y_3 + y_4
\end{aligned}$$

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 when y_2 is varied intersect at one point.

The Development of the General Case for the Model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$

Now the result for the simple linear model shown above is generalized for *n*, and one of the simplest cases to consider is a model with *n* observations with the goal of predicting y_{n+1}. On the way to the general case, a brief look is taken at the case when *n* = 7 in order to help find the pattern for general *n*.

If the *x*-values are evenly spaced as before, then

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

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

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

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

$$\hat{\beta}=\frac{1}{n\sum x_i{}^2-\left(\sum x_i\right)^2}\begin{bmatrix}\sum x_i{}^2\sum y_i-\sum x_iy_i\sum x_i\\ n\sum x_iy_i-\sum x_i\sum y_i\end{bmatrix}$$

$$=\frac{1}{\frac{n^2(n+1)(2n+1)}{6}-\frac{n^2(n+1)^2}{4}}\begin{bmatrix}\frac{n(n+1)(2n+1)}{6}\sum y_i-\frac{n(n+1)}{2}\sum x_iy_i\\ n\sum x_iy_i-\frac{n(n+1)}{2}\sum y_i\end{bmatrix}$$

$$=\frac{12}{2n^2(n+1)(2n+1)-3n^2(n+1)^2}\begin{bmatrix}\frac{n(n+1)(2n+1)}{6}\sum y_i-\frac{n(n+1)}{2}\sum x_iy_i\\ n\sum x_iy_i-\frac{n(n+1)}{2}\sum y_i\end{bmatrix}$$

$$=\frac{12}{n^2(n+1)(4n+2-3n-3)}\begin{bmatrix}\frac{n(n+1)(2n+1)}{6}\sum y_i-\frac{n(n+1)}{2}\sum x_iy_i\\ n\sum x_iy_i-\frac{n(n+1)}{2}\sum y_i\end{bmatrix}$$

$$=\frac{12}{n^2(n+1)(n-1)}\begin{bmatrix}\frac{n(n+1)(2n+1)}{6}\sum y_i-\frac{n(n+1)}{2}\sum x_iy_i\\ n\sum x_iy_i-\frac{n(n+1)}{2}\sum y_i\end{bmatrix}$$

Therefore,

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

and

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

Now, for what values of n is $\hat{y}_{n+1}$ independent of some y_i?

When $n = 4$, $\hat{y}_5$ is independent of y_2, as was seen in Example 1.

The following shows the result when $n = 7$.

When $n = 7$, (with the same other assumptions as before), equation (5) yields

$$
\begin{aligned}
y_8 &= \frac{2(2(8)+1)}{8(8-1)}\sum y_i - \frac{6}{8(8-1)}\sum x_i y_i + 8\left(\frac{12}{8(8+1)(8-1)}\sum x_i y_i - \frac{6}{8(8-1)}\sum y_i\right) \\
&= \frac{17}{4(7)}\sum y_i - \frac{3}{4(7)}\sum x_i y_i + \frac{12}{9(7)}\sum x_i y_i - \frac{6}{7}\sum y_i \\
&= -\frac{7}{28}\sum y_i + \frac{21}{252}\sum x_i y_i \\
&= -\frac{1}{4}\sum y_i + \frac{1}{12}\sum x_i y_i
\end{aligned}
$$

From this result it can be seen that when $n = 7$, $\hat{y}_8$ is independent of y_3 because the term including y_3 is

$$-\frac{1}{4}y_3 + \frac{1}{12}(3y_3) = 0.$$

An example when $n = 7$ is now briefly explored on the way to generalizing the result for n and $\hat{y}_{n+1}$.

Example 2.

According to AGI (2004), the numbers of legal abortions in the United States for women aged 18 and 19 by year are as follows:

Table 1

Legal U.S. Teen Abortions

Year	Number of Legal Abortions in the United States for Women Aged 18 and 19
1994	164,560
1995	156,960
1996	159,000
1997	157,180
1998	153,870
1999	152,520
2000	150,700

A plot of the data points is shown below in Figure 5 below.

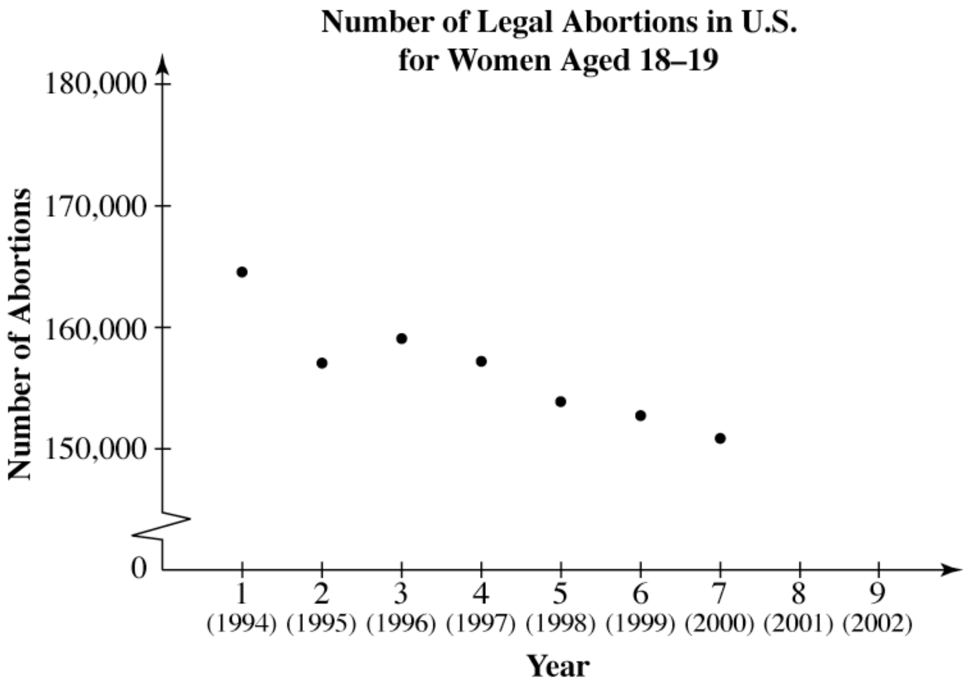

Figure 5. Legal abortions in the U.S. from 1994 to 2000 for women aged 18 and 19.

By representing 1994 by 1, 1995 by 2 and so on, and running a linear regression on the data, the equation $\hat{y}_i = 164,340 - 1985.36x_i$ is obtained. This leads to the prediction $\hat{y}_8 = 148,457$. Now, change the value of y_3 arbitrarily and rerun the regression. For example, if $y_3 = 170,000$, the regression equation changes to $\hat{y}_i = 167,483 - 2378.21x_i$, but the value of $\hat{y}_8 = 148,457$ remains unchanged. In fact, the predictor for y_8 does not depend on y_3 at all. This is illustrated graphically in Figure 6 below, using several different values for y_3 and showing that they all intersect at $(8, 148,457)$.

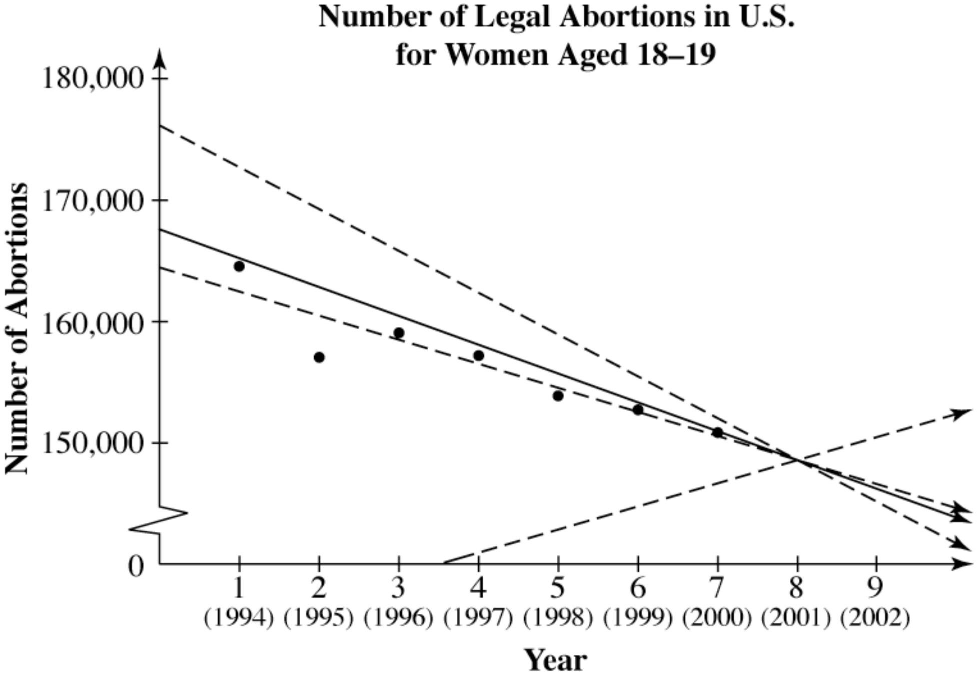

Figure 6. Regression lines for U.S. abortions with various values of y_3.

It is clear that the prediction $\hat{y}_8$, representing the predicted number of abortions in 2001, is independent of the third data value, the number of abortions that occurred in 1996.

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

$$
\begin{aligned}
\hat{y}_{n+1} &= \frac{2(2n+1)}{n(n-1)}\sum y_i - \frac{6}{n(n-1)}\sum x_i y_i + (n+1)\left(\frac{12}{n(n+1)(n-1)}\sum x_i y_i - \frac{6}{n(n-1)}\sum y_i\right) \\
&= \frac{4n+2}{n(n-1)}\sum y_i - \frac{6n+6}{n(n-1)}\sum y_i + \frac{12}{n(n-1)}\sum x_i y_i - \frac{6}{n(n-1)}\sum x_i y_i \\
&= \frac{-2n-4}{n(n-1)}\sum y_i + \frac{6}{n(n-1)}\sum x_i y_i \\
&= \frac{-2(n+2)}{n(n-1)}\sum y_i + \frac{6}{n(n-1)}\sum x_i y_i \\
&= \sum\left(\frac{-2(n+2)}{n(n-1)} + \frac{6}{n(n-1)}x_i\right)y_i \qquad (6)
\end{aligned}
$$

It can be seen from equation (6) above that the integer cases where y_i has no influence on some $\hat{y}_p$ 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,

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

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 in terms of *n* and *k* and beginning with *k* = 0.

Specifically, when $n = 4 + 3k,\ k = 0,\ 1,\ 2,\ 3,\mathcal{L}$ then y_{k+2} has no influence on 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 has no influence when estimating $\hat{y}_{n+1}$, as stated.

The Relationship Between $\hat{y}_p$ and y_d

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.

It is convenient to let all the y-values other than y_d equal 0. 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

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

and

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

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

$$\hat{\beta}_1 = \frac{x_d \cdot n - \sum x_i}{n\sum x_i{}^2 - \left(\sum x_i\right)^2} \tag{7}$$

and

$$\hat{\beta}_0 = \frac{\sum x_i{}^2 - x_d \sum x_i}{n\sum x_i{}^2 - \left(\sum x_i\right)^2}. \tag{8}$$

We wish to find values of x_p that cause any data having to do with the data point (x_d, y_d) to drop out of the prediction equation. Now since the prediction equation as it is currently written only depends on the data point (x_d, y_d), then this is true when $\hat{y}_p = \hat{\beta}_1 x_p + \hat{\beta}_0 = 0$. This yields the equation

$$0 = \hat{\beta}_1 (x_p) + \hat{\beta}_0$$

or

$$x_p = -\frac{\hat{\beta}_0}{\hat{\beta}_1} \tag{9}$$

Substituting equations (7) and (8) into (9) yields,

$$x_p = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{\sum x_i{}^2 - x_d \sum x_i}{\sum x_i - x_d \cdot n}. \tag{10}$$

By solving the equation (10) for different values of n and x_p, the same integer results as before are obtained. Namely, when $n = 4 + 3k,\ k = 0,\ 1,\ 2,\ldots$, y_{k+2} has no influence 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 $\hat{y}_p$ with corresponding x_p (x_p real), be a prediction based upon $\hat{\beta} = (X'X)^{-1} X'Y$. Then there exists an observed data value y_d with corresponding x_d (x_d real, d an integer, $1 \le d \le n$), and $d \ne p$, such that $\hat{y}_p$ does not depend on y_d. The relationship between x_p and x_d is specified by $x_p = \dfrac{\sum x_i{}^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}$.

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 $$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \qquad (11)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (12)$$

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

Note that with some simplification,

$$x_p = \frac{\sum x_i{}^2 - x_d \sum x_i}{\sum x_i - n \cdot x_d}$$ can be written as

$$x_p = \frac{\sum(\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x} \qquad (13)$$

Now, by substituting equation (13) for x_p into the equation $\hat{y}_p = \bar{y} + \hat{\beta}_1(x_p - \bar{x})$ and substituting for (11) 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 the 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

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

Conclusion

This chapter developed the relationship between noncontributory data and their corresponding predictions for a straight-line statistical model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Chapter 4 will extend these results to the second order polynomial model, the general polynomial model in *x*, and then the general univariate model that is linear in the unknown coefficients. The concept of *data wells* will also be explored, which expands this result to predictions in the *neighborhood* of values of $\hat{y}_p$ rather than just at discrete values of $\hat{y}_p$. Chapter 4 will also present and discuss some additional examples using real data.

Chapter 4:
Polynomials and the General Univariate Model

Introduction

Chapter 3 introduced the concept of observations (y_d) that do not contribute to certain predictions $\left(\hat{y}_p\right)$ for a simple model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where the predictions are made using ordinary least squares. This chapter will extend this result to other univariate models, as well as explain how the phenomenon affects predictions in the neighborhood of $\hat{y}_p$. In other words, this phenomenon is a *continuous* phenomenon rather than one that affects only a finite number of points.

The extended analysis begins with an example that uses artificial data in order to introduce the concept of *data wells*. Afterwards, the relationships between predictions and noncontributory data are developed for a general second order polynomial model, a general polynomial model, and finally for the general univariate model that is linear in its unknown coefficients. Examples using real data are shown at appropriate points in the development.

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 the data were recorded in an inconvenient manner, and it will therefore take a considerable amount of effort to retrieve the data for the number of tacos sold each year. 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 needs to calculate be independent of any of the available 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 for x_d four separate times, once for each of years corresponding to $x_p = 12$, 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 observed values 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	5.0
15	5.1

From Table 2, 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 retrieving 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. Note the fact that data point corresponding to $x_d = 4.5$, an observation that does not in fact exist, would theoretically not affect the prediction for $\hat{y}_{12}$. Likewise, the observation corresponding to $x_d = 4.8$ would not affect $\hat{y}_{13}$, and the observation corresponding to $x_d = 5.1$ would not affect $\hat{y}_{15}$. There are actually "degrees of relevance" for various data points that would strengthen the case for omitting the fifth data point. This is true because the hat values (what can be thought of as the *influence function*) are linear in x_p. This implies that the function is continuous in the neighborhood of x_d, and the influence function concerning y_d is continuous in the neighborhood of x_p. This means that y_d will have *almost* no influence on $\hat{y}_p$ in the neighborhood of points y_d that actually make no contribution. The graph of the function that shows the influence of the fourth data point $\left(\text{where } x_d = 5\right)$ for predictions between $x = 0$ and $x = 20$ is shown below.

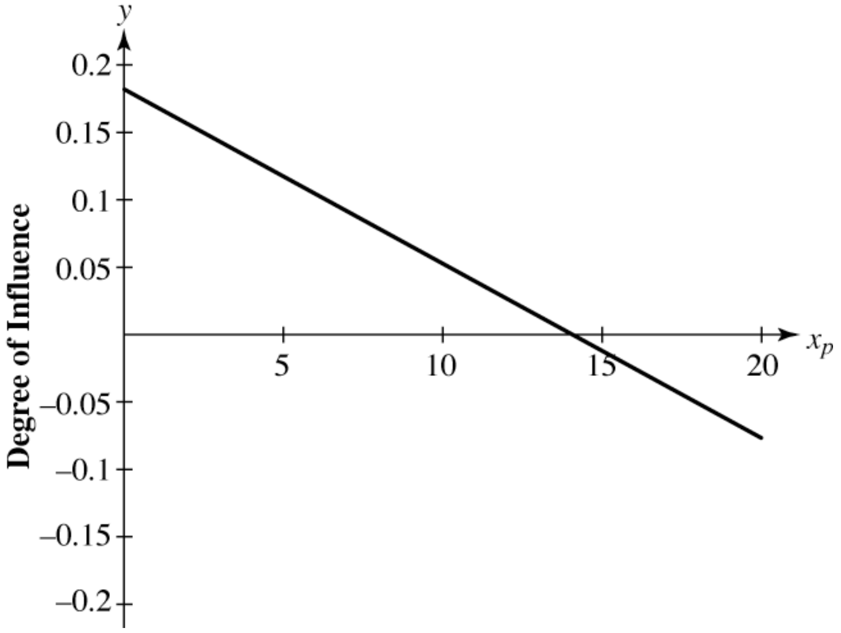

Figure 7. Degree of influence vs. x_p from Example 3.

The graph is of the hat value function for the linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ using the *x*-values from Example 3. In this case that function is given by

$$h_{dp} = \frac{1}{n} + \frac{(x_d - \bar{x})(x_p - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where h_{dp} is the influence that y_d has on $\hat{y}_p$.

In this case $x_d = 5$, which is the *x*-value of the fourth observation. It was shown in Example 3 that the observation corresponding to this *x*-value will have no influence on $\hat{y}_{14}$. The value of *n* is 10 because of the 10 observations used in Example 3. The graph shows the degree of influence (between –1 and 1) that the fourth observation will have on each prediction corresponding to x_p. Notice that the graph shows that the fourth data point has no influence on the prediction $\hat{y}_{14}$, and very little influence anywhere around this prediction. Similar graphs could be shown for each of the 10 possible values of x_d, and all would show linear relationships of influence versus x_p.

Therefore, as x_p in our example nears 15, the point corresponding to $x_d = 5$ will make little contribution to any prediction made in that neighborhood. For this reason, there is a strong case for omitting the fourth data point $\left(\text{when } x_d = 5\right)$ when predicting points around $\hat{y}_{15}$ if the collection of that data point is difficult or expensive. After all, if will have almost no influence on the prediction anyway.

The Second Order Polynomial Case

Consider the quadratic model of the form $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, where the ε_i values have a Normal distribution. Then the β_j values can be found using least squares. This model can be written in matrix form as $\underset{(n\times1)}{Y} = \underset{(n\times3)}{X}\ \underset{(3\times1)}{\beta} + \underset{(n\times1)}{\varepsilon}$, where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\quad X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix},\quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},\ \text{and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The solution for the β vector is $\beta = \left(X'X\right)^{-1} X'Y$. In order to derive the relationship between any predictions $\hat{y}_p$ and collected data value y_d that does not contribute to $\hat{y}_p$, look at the matrix algebra of the solution for the β vector.

$$X'X = \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \\ x_1^2 & x_2^2 & \mathcal{L} & x_n^2 \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \mathcal{M} & \mathcal{M} & \mathcal{M} \\ 1 & x_n & x_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}$$

Therefore,

$$\hat{Y} = X\hat{\beta} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \mathcal{M} & \mathcal{M} & \mathcal{M} \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \\ x_1^2 & x_2^2 & \mathcal{L} & x_n^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \mathcal{M} \\ y_n \end{bmatrix}$$

Recall that if 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 once again by taking the derivative of $\hat{y}_p$ with respect to y_d $\left(\frac{d\hat{y}_p}{y_d}\right)$. 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 $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p + \hat{\beta}_2 x_p^2$. 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. Just as in the case of the linear model, it can be seen 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. This can be seen even without stating the expression for $\hat{y}_p$ explicitly. 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 models corresponding to different values of y_d intersect when the other y-values are held constant. For this purpose the 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.) Therefore let the Y vector consist of zeros except let $y_d = 1$. Recall this is indicated by I_d.

In other words, let $Y = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, where the 1 is located in the dth position.

Then

$$X'Y = \begin{bmatrix} 1 & 1 & \mathcal{L} & 1 \\ x_1 & x_2 & \mathcal{L} & x_n \\ x_1^2 & x_2^2 & \mathcal{L} & x_n^2 \end{bmatrix} \begin{bmatrix} 0 \\ \mathcal{M} \\ 0 \\ 1 \\ 0 \\ \mathcal{M} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_d \\ x_d^2 \end{bmatrix}$$

and

$$\hat{\beta} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_d \\ x_d^2 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

By taking the inverse and multiplying the matrices, the $\hat{\beta}_j$ values can be expressed in terms of x_d and x_d^2. Now, the *y*-value corresponding to x_p is 0. This yields the equation

$$\hat{\beta}_2 x_p^2 + \hat{\beta}_1 x_p + \hat{\beta}_0 = 0.$$

This equation can be solved using the quadratic formula

$$x_p = \frac{\hat{\beta}_1 \pm \sqrt{\hat{\beta}_1^2 - 4\hat{\beta}_0\hat{\beta}_2}}{2\hat{\beta}_2} \qquad (14)$$

But $\hat{\beta} = (X'X)^{-1} X'Y = (X'X)^{-1} \begin{bmatrix} 1 \\ x_d \\ x_d^2 \end{bmatrix}$, and this depends on x_d.

Substituting for the $\hat{\beta}_j$ values in (14), an explicit equation for x_p can be written in terms of x_d. Unfortunately, the closed form solution is rather complex and inelegant, so the solution is illustrated as a process rather than a closed form equation.

From equation (14) it is clear that for each value of x_d there will be exactly two corresponding values of x_p. Therefore, two predictions are independent of each collected data value. In practice, however, some of these values will not be practical in use. This will become clear from the results in Example 4.

Example 4.

The following table shows the strength of kraft paper and the percentage of hardwood in the batch of pulp from which the paper was manufactured (Montgomery, Peck, & Vining, 2001).

Table 3

Hardwood Concentrations in Pulp vs. Tensile Strength of Paper

x_i Hardwood Concentration (%)	y_i Tensile Strength of Paper (psi)
1	6.3
1.5	11.1
2	20.0
3	24.0
4	26.1
4.5	30.0
5	33.8
5.5	34.0
6	38.1
6.5	39.9
7	42.0
8	46.1
9	53.1
10	52.0
11	52.5
12	48.0
13	42.8
14	27.8
15	21.9

A scatterplot of the data is shown below in Figure 8 below.

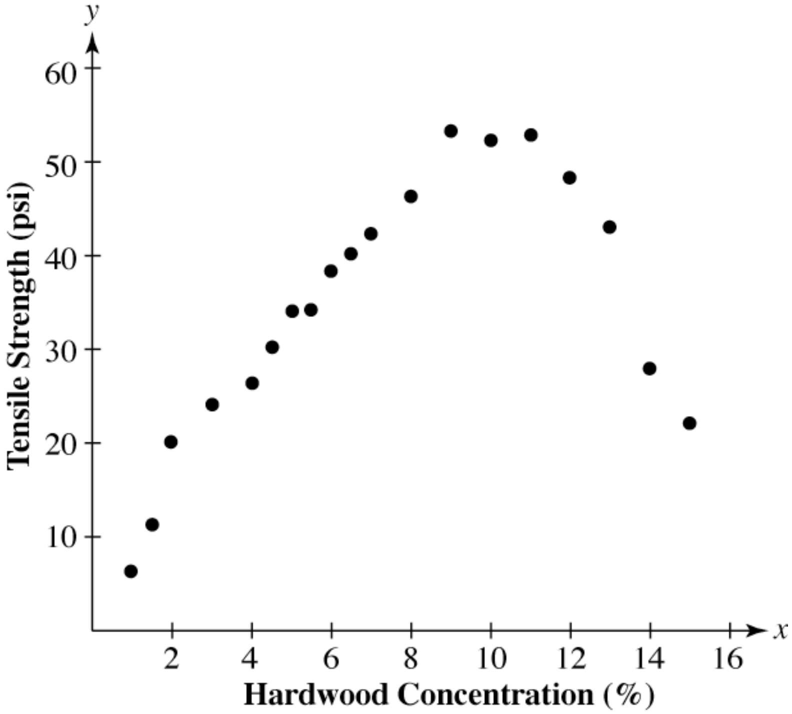

Figure 8. Scatterplot of hardwood concentration vs. tensile strength of paper.

Since the scatterplot looks a lot like a second order polynomial function, use the model $y_i = \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i$. When least squares is used to find the β-values, the fitted model is $y_i = -0.635 x_i^2 + 11.764 x_i - 6.674$, where $\hat{\beta}_2 \approx -0.635$, $\hat{\beta}_1 \approx 11.764$, and $\hat{\beta}_0 \approx -6.674$. The fitted model is shown with the scatterplot of the data in Figure 9 below.

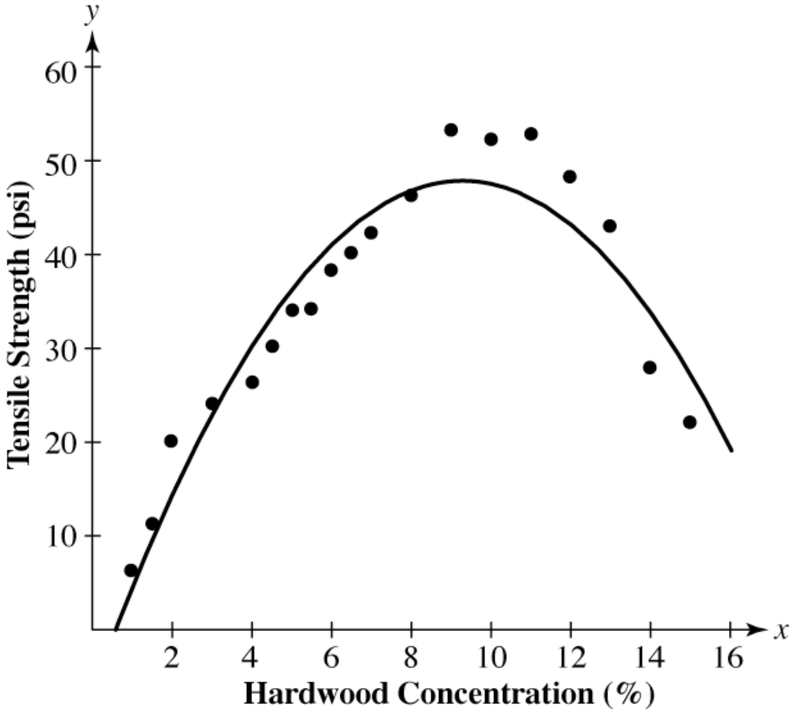

Figure 9. Quadratic model for hardwood concentration vs. tensile strength data.

Since the matrix X in the model $Y = X\beta + \varepsilon$ is a 19×3 matrix where each row is $\left\{1,\ x_i,\ x_i^2\right\}$, $(X'X)^{-1}$ can be computed as

$$(X'X)^{-1} = \begin{bmatrix} 19 & 138 & 1335 \\ 138 & 1335 & 14935.5 \\ 1335 & 14935.5 & 181427 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 0.591508 & -0.157583 & 0.00862011 \\ -0.157583 & 0.0514625 & -0.00307696 \\ 0.00862011 & -0.00307696 & 0.000195385 \end{bmatrix}$$

So,

$$(X'X)^{-1}X'Y = (X'X)^{-1}\begin{bmatrix} 1 \\ x_d \\ x_d^2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.591508 - 0.157583x_d + 0.00862011x_d^2 \\ -0.157583 + 0.0514625x_d - 0.00307696x_d^2 \\ 0.00862011 - 0.00307696x_d + 0.000195385x_d^2 \end{bmatrix}$$

This gives a representation of the $\hat{\beta}$ vector in terms of x_d. Substituting one of the 19 values of the *x*-vector into the expression for $\hat{\beta}$ and then solving the quadratic equation

$$\hat{\beta}_2 x_p^2 + \hat{\beta}_1 x_p + \hat{\beta}_0 = 0$$

gives the numerical relationship between x_p and x_d. By repeating this procedure for each of the 19 values in the *x*-vector, the following table results.

Table 4

Values For Which y_d Does Not Affect $\hat{y}_p$ for Example 4 Data

Observation Number	x_d	x_p Values for which $\hat{y}_p$ is not Affected by y_d
1	1	5.9, 13.2
2	1.5	6.3, 13.3
3	2	7.0, 13.6
4	3	10.3, 16.6
5	4	–14.2, 12.5
6	4.5	–3.5, 12.7
7	5	–0.8, 12.9
8	5.5	0.5, 13.1
9	6	1.2, 13.2
10	6.5	1.6, 13.4
11	7	2, 13.5
12	8	2.4, 14.0
13	9	2.7, 14.7
14	10	2.9, 16.0
15	11	3.2, 20.0
16	12	3.6, 97.0
17	13	0, 5.3
18	14	2.4, 8.0
19	15	2.8, 9.3

From looking at the row corresponding to observation number 8 in Table 4 it is clear that the data point corresponding to $x = 5.5$ has no effect on the predicted values when $x = 0.5$ or for $x \approx 13$. In fact, the table is comprehensive, in that it shows all the relationships between the measured data values and predicted values that are independent of them. However, not all the results are useful or meaningful. For example, the third row of the table indicates that the predictions corresponding to $x = 7$ and $x = 13.6$ will be independent of the data point corresponding to $x = 2$. However, it is unnecessary to predict a value for $x = 7$ because it was a measured point and is already part of the original data set. Also, the table indicates that the prediction corresponding to $x = 97$ is independent of the data point corresponding to $x = 12$. In practice, one would never predict a y-value when $x = 97$ because for this data set such a prediction would be useless. Least squares is not meant to predict values far outside the neighborhood of the collected data.

Even with these limitations, the table indicates useful information about predictions that will be independent of a data point. Note that in this example $x \neq i$, or even a linear transformation of i, so care must be taken to refer to the value of x_d whose corresponding y_d contributes nothing to some $\hat{y}_p$. In this case $x_d \neq d$.

Again in this case, the concept data wells, or "degrees of relevance" of data points applies. This is true because the hat values (what can be thought of as the *influence function*) are continuous in x_p since this function can be represented as the derivative of y_p with respect to y_d. When the model in y_i is a polynomial its derivative is also a polynomial, and polynomials are continuous everywhere. This implies that the function is continuous in the neighborhood of x_d, and the influence function concerning y_d is continuous in the neighborhood of x_p. This means that y_d will have *almost* no influence on $\hat{y}_p$ in the neighborhood of points y_d that actually make no contribution. From this analysis it should be clear that the concept of data wells will apply for any polynomial model as well as any model that has a derivative that is continuous everywhere. When a model has discontinuities in its derivative then the concept of data wells will apply in any area of the model for which the derivative is continuous.

Note that rounding errors and near-singularity problems can arise when taking matrix inverses and doing calculations with them. This means that the values of y_d may sometimes appear to make a slight contribution to values of $\hat{y}_p$. This only happens due to numerical problems with taking inverses or with rounding errors. There are also significant calculation costs to taking matrix inverses. The field of numerical analysis offers much insight and direction on how to find and handle these problems. While this dissertation will not go into the details of numerical computations, the potential is worth mentioning, and some evidence of these problems will even be seen in examples to follow in this research.

The General Polynomial Case

Consider the polynomial model of the form

$$y_i = \beta_k x_i^k + \beta_{k-1} x_i^{k-1} + \mathcal{L} + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i \, .$$

where the ε_i values have a Normal distribution as usual. Then the β_j values can be found using the usual least squares method. This model can be written in matrix form as

$$\underset{(n\times 1)}{Y} = \underset{(n\times(k+1))}{X} \underset{((k+1)\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon} ,$$

where

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \mathscr{M} \\ y_n \end{bmatrix},\ X=\begin{bmatrix} 1 & x_1 & x_1^2 & \mathscr{L} & x_1^k \\ 1 & x_2 & x_2^2 & \mathscr{L} & x_2^k \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{L} & \mathscr{M} \\ 1 & x_{n-1} & x_{n-1}^2 & \mathscr{L} & x_{n-1}^k \\ 1 & x_n & x_n^2 & \mathscr{L} & x_n^k \end{bmatrix},\ \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \mathscr{M} \\ \beta_k \end{bmatrix},\ \text{and } \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathscr{M} \\ \varepsilon_n \end{bmatrix}.$$

The solution for the β vector is $\beta=(X'X)^{-1}X'Y$. In order to derive the relationship between any predictions $\hat{y}_p$ and the collected data value y_d that does not contribute to $\hat{y}_p$, look at the matrix algebra of the solution for the β vector. It will become apparent that there are exactly *k* such values $\hat{y}_p$ that are independent of each collected data value.

The solution for the β vector is $\beta=(X'X)^{-1}X'Y$. In order to derive the relationship between any predictions $\hat{y}_p$ and collected data value y_d that does not contribute to $\hat{y}_p$, we proceed as before and look at the matrix algebra of the solution for the β vector.

$$X'X=\begin{bmatrix} 1 & 1 & \mathscr{L} & 1 & 1 \\ x_1 & x_2 & \mathscr{L} & x_{n-1} & x_n \\ x_1^2 & x_2^2 & \mathscr{L} & x_{n-1}^2 & x_n^2 \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} \\ x_1^k & x_2^k & \mathscr{L} & x_{n-1}^k & x_n^k \end{bmatrix}\begin{bmatrix} 1 & x_1 & x_1^2 & \mathscr{L} & x_1^k \\ 1 & x_2 & x_2^2 & \mathscr{L} & x_2^k \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{L} & \mathscr{M} \\ 1 & x_{n-1} & x_{n-1}^2 & \mathscr{L} & x_{n-1}^k \\ 1 & x_n & x_n^2 & \mathscr{L} & x_n^k \end{bmatrix}$$

$$=\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \mathscr{L} & \sum x_i^k \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \mathscr{L} & \sum x_i^{k+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \mathscr{L} & \sum x_i^{k+2} \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{L} & \mathscr{M} \\ \sum x_i^k & \sum x_i^{k+1} & \sum x_i^{k+2} & \mathscr{L} & \sum x_i^{2k} \end{bmatrix}$$

Therefore,

$$\hat{Y} = X\hat{\beta}$$

$$= \begin{bmatrix} 1 & x_1 & x_1^2 & \mathscr{L} & x_1^k \\ 1 & x_2 & x_2^2 & \mathscr{L} & x_2^k \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{L} & \mathscr{M} \\ 1 & x_{n-1} & x_{n-1}^2 & \mathscr{L} & x_{n-1}^k \\ 1 & x_n & x_n^2 & \mathscr{L} & x_n^k \end{bmatrix} \begin{bmatrix} n & \sum x_i & \sum x_i^2 & \mathscr{L} & \sum x_i^k \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \mathscr{L} & \sum x_i^{k+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \mathscr{L} & \sum x_i^{k+2} \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{L} & \mathscr{M} \\ \sum x_i^k & \sum x_i^{k+1} & \sum x_i^{k+2} & \mathscr{L} & \sum x_i^{2k} \end{bmatrix}^{-1}$$

$$\cdot \begin{bmatrix} 1 & 1 & \mathscr{L} & 1 & 1 \\ x_1 & x_2 & \mathscr{L} & x_{n-1} & x_n \\ x_1^2 & x_2^2 & \mathscr{L} & x_{n-1}^2 & x_n^2 \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} \\ x_1^k & x_2^k & \mathscr{L} & x_{n-1}^k & x_n^k \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \mathscr{M} \\ y_n \end{bmatrix}$$

Using the same arguments as before, let the *Y* vector be I_{x_d}.

In other words, let $Y = \begin{bmatrix} 0 \\ \mathscr{M} \\ 0 \\ 1 \\ 0 \\ \mathscr{M} \\ 0 \end{bmatrix}$, where the 1 is located at x_d.

Then

$$X'Y = \begin{bmatrix} 1 & 1 & \mathscr{L} & 1 & 1 \\ x_1 & x_2 & \mathscr{L} & x_{n-1} & x_n \\ x_1^2 & x_2^2 & \mathscr{L} & x_{n-1}^2 & x_n^2 \\ \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} & \mathscr{M} \\ x_1^k & x_2^k & \mathscr{L} & x_{n-1}^k & x_n^k \end{bmatrix} \begin{bmatrix} 0 \\ \mathscr{M} \\ 0 \\ 1 \\ 0 \\ \mathscr{M} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_d \\ x_d^2 \\ \mathscr{M} \\ x_d^k \end{bmatrix}$$

and

$$\hat{\beta}=\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \mathcal{L} & \sum x_i^k \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \mathcal{L} & \sum x_i^{k+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \mathcal{L} & \sum x_i^{k+2} \\ \mathcal{M} & \mathcal{M} & \mathcal{M} & \mathcal{L} & \mathcal{M} \\ \sum x_i^k & \sum x_i^{k+1} & \sum x_i^{k+2} & \mathcal{L} & \sum x_i^{2k} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_d \\ x_d^2 \\ \mathcal{M} \\ x_d^k \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \mathcal{M} \\ \beta_k \end{bmatrix}$$

Again, we wish to find values of x_p that cause any data having to do with the data point (x_d, y_d) to drop out of the prediction equation. Now since the prediction equation as it is currently written only depends on the data point (x_d, y_d), then this is true when $\hat{y}_i = \hat{\beta}_k x_p^k + \hat{\beta}_{k-1} x_p^{k-1} + \mathcal{L} + \hat{\beta}_2 x_p^2 + \hat{\beta}_1 x_p + \hat{\beta}_0 = 0$. This yields the equation

$$\hat{\beta}_k x_p^k + \hat{\beta}_{k-1} x_p^{k-1} + \mathcal{L} + \hat{\beta}_2 x_p^2 + \hat{\beta}_1 x_p + \hat{\beta}_0 = 0.$$

This is a *k*th degree polynomial, and it has *k* roots. The nature of the *Y* vector that helped determine the β values makes it likely the all the roots of this polynomial are likely to be unique and real valued. (Recall that the *Y* vector used was the indicator function denoted I_d.) The reason for this is that it seems intuitive that maximizing the number of sign changes in the β values, and thus the coefficients of the polynomial, will be needed for a best approximation to the model. However, the actual proof of this seemingly simple concept is likely to be quite difficult and is related to the area of mathematics known as approximation theory. However, if all the roots of the polynomial are real and unique, then there are exactly *k* predictions (where *k* is the degree of the polynomial) that are affected by noncontributory data for each observed value in the *Y* vector.

Now the general process for computing the values of x_p and corresponding x_d values can now be explained. This is the main result for the general polynomial model:

Process for finding data wells for a polynomial model:

Given the polynomial model $Y = X\beta + \varepsilon$ of the form

$$y_i = \beta_k x_i^k + \beta_{k-1} x_i^{k-1} + \mathcal{L} + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i :$$

Set a counting variable (i) to 1: i = 1.

Step 1: Compute $(X'X)^{-1}\alpha$, where the α vector is the vector $\begin{bmatrix} 1 & x_d & x_d^2 & \mathcal{L} & x_d^k \end{bmatrix}$.

This gives a vector representation of the β-values in terms of x_d.

Step 2: Set $x_d = x_i$

Step 3: Substitute the value for x_d into the expression for $\hat{\beta} = (X'X)^{-1}\alpha$.

Step 4: Solve the equation $\hat{\beta}_k x_p^k + \hat{\beta}_{k-1} x_p^{k-1} + \mathcal{L} + \hat{\beta}_2 x_p^2 + \hat{\beta}_1 x_p + \hat{\beta}_0 = 0$ for x_p, numerically if necessary, where the values for $\hat{\beta}$ were computed in Step 3.

Step 5: Increment the counter variable *i: i* = *i* + 1. Go back to step 2 if *i* ≤ *n*.

In this way, the relationship between x_d and x_p value(s) will be found for all values of x_d. Note that there will be exactly *k* values of x_p corresponding to every value for x_d. This process is now illustrated with an example.

Example 5.

The following table shows the observed voltage drop in a guided missile motor in relation to the time of the missile flight (Montgomery, Peck, & Vining, 2001).

Table 5

Voltage Drop Data vs. Time

Observation number	Time (Seconds)	Voltage Drop
1	0.0	8.33
2	0.5	8.23
3	1.0	7.17
4	1.5	7.14
5	2.0	7.31
6	2.5	7.60
7	3.0	7.94
8	3.5	8.30
9	4.0	8.76
10	4.5	8.71
11	5.0	9.71
12	5.5	10.26
13	6.0	10.91
14	6.5	11.67
15	7.0	11.76
16	7.5	12.81
17	8.0	13.30
18	8.5	13.88
19	9.0	14.59
20	9.5	14.05
21	10.0	14.48

A scatterplot of the data is shown in Figure 10 below.

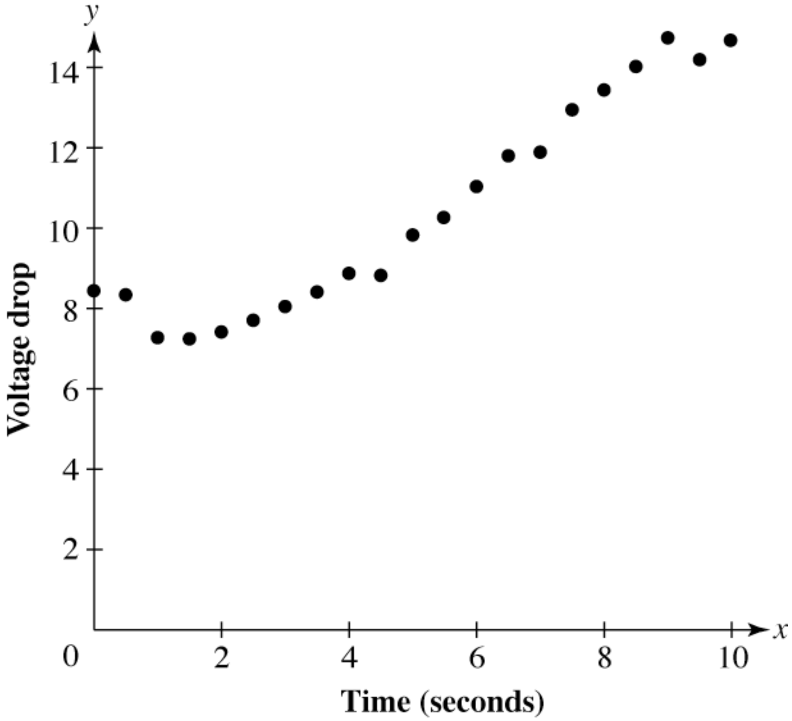

Figure 10. Scatterplot of voltage drop data.

Since the scatterplot looks a lot like a third order polynomial function, use the model $y_i = \beta_3 x_i^3 + \beta_2 x_i^2 + \beta_1 x_i + \beta_0 + \varepsilon_i$. When least squares is used to find the β-values, the fitted model is $y_i = -0.024 x_i^3 + 0.425 x_i^2 - 1.228 x_i + 8.392$, where $\hat{\beta}_3 \approx -0.024,\ \hat{\beta}_2 \approx 0.425,\ \hat{\beta}_1 \approx -1.228,$ and $\hat{\beta}_0 \approx 8.392$. The fitted model is shown with the scatterplot of the data in Figure 11 below.

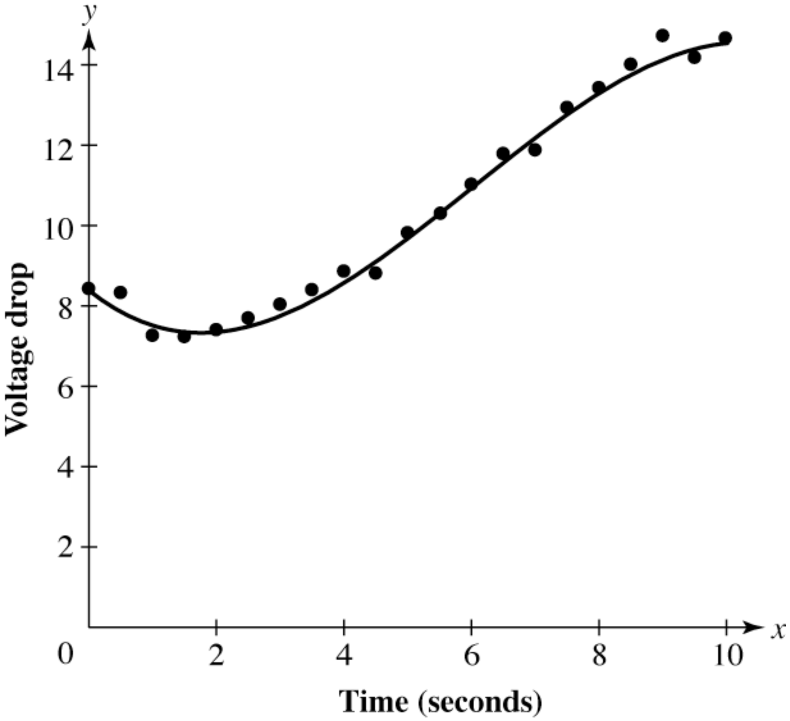

Figure 11. Cubic model for voltage drop data.

Now the process described above is used to find the relationships between the x_p and x_d values.

Step 1: Compute $\left(X'X\right)^{-1}\alpha$.

The design matrix *X* is a 21 by 4 matrix where the elements in each row are $\left\{1, x_i, x_i^2, x_i^3\right\}$. Therefore, $\left(X'X\right)^{-1}$ is a 4 by 4 matrix,

$$\left(X'X\right)^{-1} = \begin{bmatrix} 0.543 & -0.396 & 0.077 & -0.004 \\ -0.396 & 0.429 & -0.098 & 0.006 \\ 0.077 & -0.097 & 0.024 & -0.001 \\ -0.004 & 0.006 & -0.001 & 0.000 \end{bmatrix}$$

The values shown in the square matrix above are rounded. Since rounding errors can be greatly increased when numbers are small, all the calculations were carried out to six decimal places throughout and only rounded at the end.

The vector α was computed by multiplying the transpose of the design matrix by I_{x_d}. This yields

$$\alpha = \begin{bmatrix} 1 \\ x_d \\ x_d^2 \\ x_d^3 \end{bmatrix}.$$

Hence,

$$(X'X)^{-1}\alpha = \begin{bmatrix} -0.004x_d^3 + 0.077x_d^2 - 0.396x_d + 0.543 \\ 0.006x_d^3 - 0.098x_d^2 + 0.429x_d - 0.396 \\ -0.005x_d^3 + 0.024x_d^2 - 0.097x_d + 0.077 \\ -0.001x_d^2 + 0.006x_d - 0.004 \end{bmatrix}$$

Finally, substituting $x_d = x_i = [0, 0.5, 1.0, \mathcal{L}, 9.0, 9.5, 10.0]$ one element at a time yields the following table:

Table 6

Values For Which y_d Does Not Affect $\hat{y}_p$ for Example 5 Data

Observation Number	x_d	x_p Values for which $\hat{y}_p$ is not Affected by y_d
1	0	2.2, 6.1, 9.3
2	0.5	3.2, 6.8, 9.5
3	1.0	–23.5, 5.1, 9.1
4	1.5	–1.3, 5.7, 9.2
5	2.0	–0.2, 5.9, 9.3
6	2.5	0.2, 6.2, 9.4
7	3.0	0.4, 6.5, 9.4
8	3.5	0.6, 7.1, 9.6
9	4.0	0.7, 7.9, 10.1
10	4.5	0.8, 8.7, 13.1
11	5.0	0.9, 9.0, 2058.4
12	5.5	–3.1, 1.3, 9.2
13	6.0	–0.1, 2.1, 9.3
14	6.5	0.4, 2.9, 9.4
15	7.0	0.6, 3.4, 9.6
16	7.5	0.6, 3.8, 9.8
17	8.0	0.7, 4.1, 10.2
18	8.5	0.7, 4.3, 11.4
19	9.0	0.9, 4.9, 35.3
20	9.5	0.5, 3.2, 6.8
21	10	0.7, 3.9, 7.8

Note that all the values in the right column (values of x_p) are approximate. However, this does not present a problem because of the concept of *data wells* that was described earlier. Any prediction in the neighborhood of the actual value of x_p will be essentially independent of the related observation taken at x_d. Because of data wells, the closer *x* is to the actual value of x_p, the closer the corresponding observation at x_d is to a zero contribution to the predicted value. Note that the slope of the derivative will affect the degree to which data wells apply. A full analysis of the sensitivity of data wells to various models is out of the scope of this research, but should be examined in future research.

As in Example 4, the table shows some values that are not meaningful or not useful. An example of a value that is not meaningful is the row corresponding to observation 20. Note that the table indicates that the prediction when $x = 0.5$ does not depend on the observation when $x = 9.5$. An observation already exists for $x = 0.5$, so a reference to a prediction at $x = 0.5$ is not meaningful. An example of a value that is not useful is in the row corresponding to observation 11. This row indicates that the prediction when $x = 2058.4$ is not dependent on the observation when $x = 5.0$. Least squares would not be an appropriate modeling method to predict what would happen at a point so far outside the neighborhood of the observed *x*-data.

General Linear Model

In previous sections, the relationship between x_p and x_d has been derived for straight-line models, quadratic models, and then for the general polynomial model. The same derivation technique can now be used to derive the relationship for any univariate model. Because the relationship between x_p and x_d changes with each unique model, the computation technique must be described as a process, just as it was for the general polynomial model.

Given the general univariate model $Y = X\beta + \varepsilon$ of the form

$y_i = X_{[i]}\beta + \varepsilon_i$, where

$X_{[i]}$ represents the *i*th row of the design matrix *X*.

Then the least squares estimate for the β-vector is

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

and predicted values from the model are given by

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1} X'Y \qquad (15)$$

Process to find the relationship between x_p and x_d follows:

Step 0: Set a counting variable (*i*) to 1: *i* = 1.

Step 1: Compute $(X'X)^{-1}\alpha$, where the $\alpha = X' I_{x_d}$.

This gives a vector representation of the β-values in terms of x_d.

Step 2: Set $x_d = x_i$

Step 3: Substitute the value for x_d into the expression for $\hat{\beta} = (X'X)^{-1}\alpha$.

Step 4: Solve the equation $\hat{Y}(x_p) = 0$ for x_p, numerically if necessary, where the values for $\hat{\beta}$ were computed in Step 3, and $\hat{Y}$ is computed as indicated in (15) above.

Step 5: Increment the counter variable *i: i* = *i* + 1. Go back to step 2 if *i* ≤ *n*.

In this way, the relationship between x_d and x_p value(s) will be found for all values of x_d. Note that the number of values of x_p corresponding to every value for x_d will depend on the number of solutions to the general equation $\hat{Y}(x) = 0$.

As before, the phenomenon is continuous rather than only applying to a finite number of points. Recall the previous discussion about data wells. Predictions in the neighborhood of actual predictions $\hat{y}_p$ that are independent of some observation y_d will be virtually independent of y_d so long as the derivative of the model is continuous in that neighborhood.

Conclusions

The research presented about observations that do not contribute to certain predictions raises a new area of research in least squares sensitivity analysis. This area of sensitivity analysis has been long neglected compared to other areas of research in least squares (Belsley, et al., 2004), but sensitivity analysis may contribute to better prediction techniques in the future. It is hoped that this research will add to the knowledge that will result in a better understanding of how to deal with observations, models, and predictions.

The relationships between predictions and related noncontributory data were derived in this research for the straight-line model, the second order polynomial model, the general polynomial model, and the general univariate model linear in the unknown coefficients. The concept of data wells was also introduced, defined, and discussed. Data wells show that noncontributory data are a continuous, rather than discrete, phenomenon.

Examples were shown throughout that illustrated the effect that noncontributory data might have on data collection decisions as well as data analysis.

Since the phenomenon studied in this research is brand new, it raises a large number of future research possibilities that could not be covered in this dissertation. A number of ideas about future research avenues will be suggested in Chapter 5 along with some preliminary findings about some of the topics.

Chapter 5:
Summary, Conclusion, and Recommendations

Introduction and Brief Overview of Findings

Least squares sensitivity analysis was examined in this research in light of how any particular observation affects a particular prediction made calculated using a least squares modeling technique. Past research has centered almost exclusively on finding very influential observations, while this research concentrated on finding observations that had no influence.

The influence of a data point on predictions was studied for two specific models, and then for two general models. A closed form relationship relating y_d and $\hat{y}_p$ was derived for the straight-line model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and numerical processes were developed to find those relationships for the second order polynomial in *x*, the general polynomial in *x* with degree *k*, and for the general univariate model that is linear in its unknown coefficients.

The lack of influence of certain data points on specific predictions was found to be more than a discrete phenomenon affecting a finite number of cases. It is actually a continuous phenomenon affecting an infinite number of data points. The idea that a prediction $\hat{y}_p$ that is independent of an observation y_d is also virtually independent of the observation y_d anywhere in the neighborhood of $\hat{y}_p$ was developed and defined as *data wells.* The concept of data wells implies that any problem arising out of noncontributory data points cannot simply be fixed by adding some value δ to the value of *p* before calculating the prediction $\hat{y}_p$. Data wells are active in all areas of a model where the derivative of $\hat{y}_p$ with respect to y_d is continuous.

Preliminary results indicate that the phenomenon of noncontributory data extends beyond univariate models to multivariate models as well, and that there are *j* values of $\hat{y}_p$ corresponding to every observation y_d, where *j* is the number of unique variables in the model. The analysis of this concept is beyond the scope of this research, but appears to be a promising area for future work, and is described in the future research section below.

Interpretation of Findings

If all data are not being used when making a prediction, this represents a loss of information; this therefore implies that other prediction techniques that use all the available observations could potentially give more accurate predictions. How much this loss of data actually affects the accuracy of predictions is almost certainly in itself a function that is dependent upon the number of observations and the variance of those observations. It also seems likely that such a function would be dependent on other variables such as the distance the *x*-value of the prediction is from the *x*-value of the unused data point, but future research will have to make such determinations. Minimally, the results given in this research call loudly for more research into predictions that are affected by the phenomenon of noncontributory data.

Further, the phenomenon of noncontributory data will almost certainly be most important when predictions are made using small data sets, where data collection is difficult or expensive, or when a particular data point that does not contribute to the desired predicted value is deemed to be more accurate than other observations. This research suggests that care should be used when making predictions with small data sets, and alternative prediction techniques should be considered in cases where a data point is known to have no contribution to a desired predicted value.

The well-known analogy from physics that was shown in chapter 3 gives further credibility to the concept of noncontributory data. The relationship between the physical example and the statistical example for the straight-line model also makes it interesting that this phenomenon was not discovered earlier. However, this also underlines the need for further research in the general area of sensitivity analysis in least squares.

Implications for Social Change

The major potential for social change arising from this research has to do with the way researchers might approach data collection and prediction computations in the future. The following paragraphs describe how accuracy might be improved as a result of knowledge about noncontributory data. It is also possible researchers might save time and money as a result of changing data collection methods. In fields where human life is at stake or quality of life is influenced, better prediction techniques that may arise out of this research have the potential to make an even bigger difference. Though the calculation of predictions is not the only reason that least squares modeling is employed, it is one of the most common reasons for modeling data. Until now it was generally assumed that all data were used in making predictions derived from a least squares model. Though it was recognized that observations contributed to predictions in an uneven manner, it was a surprise to learn that some data are actually irrelevant for the purpose of certain predictions.

Forecasting is utilized in almost every field in current academic research and practice. Modeling and prediction are commonplace in virtually every area of the social sciences. These include but are not limited to sociology, psychology and business. The hard sciences also routinely use modeling and prediction. Some examples are biology, genetics, chemistry, physics, and medicine. Any phenomenon that has a potential effect on the accuracy of predictions has broad implications for research in every field in which modeling and prediction are used. In fields of medicine, for example, better predictions could literally help doctors to make more accurate diagnoses and patients to make decisions about treatment based on more accurate information. In general, a better understanding of data that we collect may eventually result in better, more accurate predictions.

In cases where data collection is difficult or expensive, it may not be necessary to collect data for values of the independent variable where that observation would not be used in the desired prediction calculation anyway. Therefore, the knowledge gained by this research could potentially save both time and money for researchers or anyone collecting data for the purpose of prediction. In cases where researchers only have access to small data sets, it will be important for them to know that one of their data points may not have any effect on predictions made. This knowledge may lead to better prediction techniques for small data sets, lead to more accurate disclosure of potential problems, or both.

A more accurate understanding of how observations are actually used in the computation of predictions has the potential of helping researchers design experiments better so that the most useful data are collected. It is possible to know in advance of data collection what degree of influence each observation will have. If it is known in advance that a particular observation will be particularly valuable, an experiment should be able to be arranged so that this particular observation will have a reasonable degree of influence on a desired prediction. While this idea certainly has its limits, in that it is not possible to design an experiment so that all observations collected will have the degree of influence desired), the knowledge that one or two data points can be manipulated in this way may be helpful in some cases.

Recommendations for Action

The fact that the fairly simple phenomenon of data that does not contribute to predictions was missed until now underlines the fact that sensitivity analysis in general has not received much attention. The research done in this dissertation raises many new potential areas of research in sensitivity analysis, some of which may lead to better prediction techniques. At the very least, sensitivity analysis allows researchers and data analysts to better understand the data they are working with. Predictions and models will only be as good as the understanding of the processes and methods that are used to analyze the data and make predictions. To this end, sensitivity analysis needs to be given more attention by statisticians who work in least squares, and even with other norms such as the L_1 norm (absolute value norm) and the L_∞ norm (the minimax norm). Future research may show that there are cases that clearly indicate the use of one norm over another when making predictions under certain conditions.

Even before further research is done in sensitivity analysis, researchers who are using or have used least squares for prediction with small data sets should look at their analysis to see if important predictions are affected by the phenomenon of noncontributory data. In cases where noncontributory data has had an effect on predictions, an assessment should be made about whether another prediction technique (like minimax modeling) should be used, at least for comparison purposes. In cases where further analysis is not feasible, the fact of noncontributory data should at least be explicitly disclosed.

Recommendations for Further Research

The research described in this dissertation involved an addition to current knowledge regarding sensitivity analysis in least squares. Because the discovery of data that does not contribute to certain predictions is brand new, it raises many possibilities for future research. Some of these are briefly described below:

1. Now that calculations can be performed to find predictions that are independent from a data point, thorough sensitivity analysis needs to be performed to determine when this noncontributory data actually begins to make a significant difference to the predictions. In other words, in which situations are predictions adversely affected by the loss of a data point? Another way to look at this is that there should be a curve showing the distance one is from a prediction for which a data point contributes nothing vs. the degree of effect on that prediction. The effect of an observation on that prediction would be 0 right at the *x*-value corresponding to the prediction, and would change as one moved away from that point. The effect would seemingly be sensitive to the number of observations, and would be larger when the value of *n* is small than when *n* is large. The

sensitivity may also be affected by variables such as the variance of the data and the actual model used, among others. An ideal result would find some numerical measure of the effect of the “loss” of a data point, and relationships that describe how the key variables affect this measure.

2. The concept of *data wells* was introduced in this dissertation. This dissertation clearly defined predictions that are independent of an observation. Fortunately, a desired predicted value will rarely lie directly on a point where an observation does not contribute to it. However, these predictions are also *essentially* independent of the same observation anywhere in the neighborhood of the actual point where the observation drops out. Hence the predictions are called data wells rather than just individual points. A clear area of inquiry would involve sensitivity analysis about data wells, and how far away from the actual prediction one can go before the independent observation becomes a factor (i.e., does contribute to the prediction). This relationship is almost certainly defined by a continuous function of some sort. Factors that may affect this function are the number of observations, specific model in use, and variance of the data. The definition of a function or set of function dependent on one or more of these variables would be very useful in determining where noncontributory data becomes a factor in a prediction.

3. Matrix inverses are well known to be sensitive to rounding errors and near singularities. It is unknown how sensitive the issue of data wells is to these numerical issues. Near singularities in matrices are well understood in the field of numerical analysis, but a good future study would be one to discover how sensitive noncontributory data are to rounding errors. This itself would make an interesting sensitivity analysis.

4. The hat matrix is documented to be described by the matrix expression $H = X(X'X)^{-1}X'$ (Belsley, et al., 2004). The most efficient method to find noncontributory data, however, required a numerical approach for all but the simplest of linear models. The literature on sensitivity analysis in least squares also suggests the use of numerical methods to find the hat matrix as well (Belsely, et al., 2001; Chatterjee & Hadi, 1988). It is therefore currently believed that no single formulaic relationship between noncontributory data and the predictions corresponding to this data exists. A future study could either find a formulaic relationship or prove that none exists. Since the expression for the hat matrix is valid for multivariate as well as univariate models that are linear in the unknown coefficients, the same question exists for multivariate linear models.

5. A general procedure was found in this dissertation to relate noncontributory data points to the particular predictions that correspond to them. This procedure was developed for general univariate models that are linear in the unknown coefficients. Initial findings indicate that the phenomenon of noncontributory data extends to multivariate models and that there are exactly *k* predictions affected for each observation for which a linear multivariate model has *k* degrees of freedom. A future study could find a procedure to find the relationships for general multivariate models.

6. This research studied noncontributory data only for models that are linear in the unknown coefficients. A study could be made of nonlinear models that are fitted via least squares. Is there a similar phenomenon that occurs in nonlinear models?

7. This research examined only the least squares norm. In other words, the coefficients of all models in this research were fitted using the least squares error norm. The L_2 (least squares) norm finds the coefficients that minimize the sum of the squared errors of the distance between the fitted values and the observations. Other error norms are well known, however. Two of the most well known norms are the L_1 (absolute value) norm and the L_∞ (minimax) norm. The L_1 norm finds the coefficients that minimize the sum of the absolute value of the distance of the fitted values from the observed values, and the L_∞ norm finds the coefficients that minimize the maximum error, or distance between fitted and observed values.

The use of other norms may be a possible way to correct predictions for the loss of information arising from noncontributory data, but it is unknown whether a phenomenon of noncontributory data may also exist when making predictions using a norm other than the least squares norm. Future research could determine whether this phenomenon exists when making predictions using one or more of these other error norms.

8. This research aimed to determine when an observation failed to make any contribution to a prediction or set of predictions. Future studies could determine what corrective action should be taken in such cases, and under what circumstances corrective action is necessary. Presumably the phenomenon will have a greater effect on predictions when *n* is small, but even then it is not known what corrective action to take. The concept of data wells makes it useless to simply make a prediction at a point very close to the

desired point. Whether a different error norm should be used in these cases, or whether there is some corrective estimator is unknown and needs to be studied.

9. Any data point that isn't used in the calculation of a prediction is a loss of information. However, it needs to be determined exactly how much information is really lost, and how this affects predictions. One way to determine this would be to use least squares to make predictions for which the actual observation is already known. Least squares predictions that are independent of a data value could be compared with predictions made with the same number of data points, but in cases where there is no noncontributory data. By repeating this procedure and comparing the results, a determination could be made about how much information is lost because of noncontributory data.

In addition, least squares predictions with noncontributory data could be compared to predictions made using some alternative method. Then the results from each method could be compared to determine whether the least squares predictions are, on the average, suboptimal to predictions made using some other method where all the data points are used.

10. A physical application that is equivalent to noncontributory data was shown in chapter 3. However, this application only applied to straight-line models. Though there are no obvious physical applications to models other than the simple straight-line model, further research could determine whether any similar physical applications exist that correspond to other linear models.

11. The polynomial in x_p from chapter 4 that has the β values as coefficients intuitively will have exactly *k* unique, real roots, where *k* is the degree of the polynomial. A study in approximation theory could possibly prove this result. This would determine exactly how many predictions are affected by noncontributory data for each observed value.

Concluding Statement

The idea that all observations are not necessarily used to make predictions using least squares is a fairly simple phenomenon that has escaped attention until now. This research made the important first step of defining under what conditions this phenomenon occurs. However, it will ultimately be imperative to determine how much of an effect this phenomenon really has on the predictions that are involved, and to find alternative prediction techniques to compensate in cases where predictions are detrimentally affected.

Further, the discovery of a new phenomenon in such a mature area of statistics underlines the need for further research in sensitivity analysis. This will hopefully lead to better understanding of data that is collected and analyzed, and better prediction and analysis techniques for the future.

References

Beaton, A. E. & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band spectroscopic data. *Technometrics, 16,* 147–185.

Beckman, R. (1990). Discussion of 'assessment of local influence' by R. D. Cook. *Journal of the Royal Statistical Society, Ser. B., 48,* 161–162.

Belsley, D., Kuh, E., & Welsch, R. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

California Department of Education, Educational Demographics Unit. district and school enrollment by grade, San Jose Unified School District, 2001-2004 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: Wiley.

Chave, A., & Thompson, D. (2003). A bounded influence regression estimator based on the statistics of the hat matrix. *Applied Statistics, 52(3),* 307–322.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 42(1),* 65–68.

Cook, R. D. & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics, 22(4),* 495–508.

Farebrother, R. W. (1988). *Linear least squares computations.* New York: Marcel Dekker.

Fox, J. (1991). *Regression diagnostics: Quantitative applications in the social sciences.* Newbury Park, CA: Sage Publications.

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].

Hoaglin, D., & Welsch, R. (1977). The hat matrix in regression and ANOVA. *The American Statistician, 32(1),* 17–22.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Technometrics, 32(1),* 59–65.

Johnson, W., & Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Technometrics, 32(1),* 137–144.

Montgomery, D., Peck, E., & Vining, G. (2001). *Introduction to linear regression analysis.* New York: Wiley.

Plackett, R. (1972). Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares. *Technometrika, 59(2),* 239–251.

Thomas, W., & Cook, R. D. (1990). Assessing influence on predictions from generalized linear models. *Technometrics, 32(1),* 59–65.

Younger, M. S. (1979). *A handbook for linear regression.* Belmont, CA: Wadsworth Publishing Company, Duxbury Press.

CV

Teresa L. Bittner
957 Wilmington Way
Redwood City, CA 94062
(650) 599-9188

EDUCATION:

Ph.D. Applied Management Decision Sciences, 2006
Walden University
GPA: 4.0

M.S. Statistics, June, 1985
University of CA, San Diego
GPA: 3.9

B.S. Applied Mathematics, (Minors: Elec. Eng. & Music), June, 1984
University of California, San Diego
GPA: 3.9

EXPERIENCE:

Founder and CEO,
Laurel Technical Services, LLC, (LTS), (1991-2000)

As an entrepreneur, Ms. Bittner began LTS in an industry in which she had no previous work experience. From its inception, LTS was a one of a kind company specializing in math and science education, maintaining contracts with virtually every major U.S. publisher of math and science educational materials. LTS provided services from conception and writing all the way to full production services of print, software, videos, websites and other multimedia materials from Kindergarten through graduate school level math, science, business, and statistics. The company specialized in the most difficult and technical educational materials. Ms. Bittner founded and ran all aspects of LTS from its inception in 1991 through its sale to a private investment group in 1999, leaving a mature and thriving company (of approx. $12 million in sales) in 2000. Bittner began and grew the company without any outside investors, providing for its growth with net profits exceeding 50% of sales. In addition to traditional CEO responsibilities, Ms. Bittner acted as the chief marketing officer and chief financial officer until after the company's sale in 1999. Ms. Bittner also participated in writing various math education materials.

Founder and CEO, Laurel Tutoring Services (1989-1991)
Founded and managed all aspects of a tutoring company, specializing in tutoring junior high and high school students in math and science. Managed up to a dozen adult tutors and tutored many students personally. Began and ran weekly class groups of Advanced Placement Calculus students, resulting in near 100% success of these students on the Advanced Placement Calculus exam.

Teacher, UC San Diego, Walden University, Canada College, Carlmont High School, North Star Academy and Roy Cloud School (Various assignments between 1984 to present)
Miscellaneous teaching assignments in Algebra, Calculus, Finite Math with Probability and Statistics at San Francisco Bay Area Junior High School, High School and Junior College. Was given high degree of responsibility in teaching assignments at UC San Diego, working with tenured professors in classes including calculus, upper division calculus based statistics and probability.

Systems Engineer, Mirage Systems, Inc., and Systems Control Technology, Inc. (1985-1989)

Developed algorithms and performance models for two San Francisco Bay Area defense firms.

Speaker
Ms. Bittner has spoken in several forums, including a talk as keynote speaker to undergraduate and graduate students of mathematics, statistics and engineering at UC San Diego in 2003, teachers of mathematics, and many other student and parent groups.

Consultant, various small technical companies
(1998-present)
Ms. Bittner has provided operations consulting for various small technical companies in the San Francisco Bay Area.

Ms. Bittner has extensive experience giving oral presentations to a wide variety of audiences, including special presentations to high-ranking military officers.

LANGUAGES: Conversational in German