Knowledge Area Module VII:

Least Squares Regression Analysis: A New Discovery

Student: Terri Bittner

Faculty Mentor: Dr. Kimberly L. Ross

Faculty Assessor: Dr. Thomas Spencer

Walden University

October 22, 2005

BREADTH ABSTRACT

Least squares regression goes back to Gauss, who first used the method. This section covers the general theory of least squares regression including development of the theory behind the method, various types of problems that can be solved using least squares, and important ways of testing the adequacy of a model. This is illustrated with real data from the scientific literature, including graphics. Various cautions about least squares use and misuse are also given, as well as a brief discussion of alternative performance norms.

DEPTH ABSTRACT

Though estimation via least squares is a very mature area of statistics, a phenomenon that occurs under certain conditions may have escaped attention until now. Some data is not used when making predictions using ordinary least squares. This paper derives the mathematics behind this phenomenon and describes when data drops out for straight line models. A theorem is derived and proven that describes the relationship between the data point being estimated and the data point that drops out of that estimate. A physical connection to the statistical result is also explored.

APPLICATION ABSTRACT

Data from a company called Decision Insite Corporation, a firm that specializes in school enrollment projections, is analyzed using two different models that are linear in the unknown coefficients. The least squares regression method is compared with enrollment predictions made using the previous year percentage change, the method that has been used by the company for the past few years. Several cautions are discussed, mostly having to do with the lack of an adequate amount of data to analyze, and the danger of extrapolating far into the future using least squares.

Learning Agreement

Specialized Knowledge Area Module VII:

Custom Designed KAM—Least Squares Predictions

Student: Terri Bittner

Faculty Mentor: Dr. Kimberly L. Ross

Faculty Assessor: Tom Spencer

Walden University

August 10, 2005

Overview

This learning agreement applies to my doctoral study in AMDS/Operations Research. I will use KAM VII to familiarize myself with classic and current research concerning least squares prediction and to describe and demonstrate a discovery I have developed concerning special cases where some data drops out of calculations in least squares predictions.

I will critically examine least squares theory and methods throughout this study and demonstrate an understanding of both the techniques and the underlying assumptions associated with making predictions using least squares techniques. The concept of modeling data in order to make predictions will be explored in depth. I will also briefly explore the absolute value (L1) norm and the minimax (L∞) norm and describe difficulties in using them for applications.

During the Depth portion of the KAM, a specific problem with the current method will be described. The discovery I have made concerning data that drops out of prediction equations under specific circumstances will be developed mathematically and described in detail. During the Application portion of the KAM a specific study will be made of a recent application problem that uses least squares prediction. This is the application that inspired the result from the Depth section. The application problem will be analyzed in light of the results from the Breadth and the Depth section and specific recommendations will be made about modeling and prediction for this application.

Breadth Objectives

Develop the theory behind least squares prediction for univariate and multivariate problems, including the mathematical theory necessary to understand the underlying assumptions behind the methods. Illustrate the theory using real or simulated data, including some problems that will be applicable to the Application section of the KAM. In addition, I will examine two alternative performance norms, the L1 norm, (also called the absolute value norm), where the goal is to minimize the sum of the absolute values of the errors, and the L∞ norm (also called the minimax norm), where the goal is to minimize the maximum error. I will discuss advantages and disadvantages of these alternatives and explain why least squares is used almost exclusively in modeling despite some advantages to the other performance norms.

Preliminary Breadth References

Farebrother, R. W. (1988). *Linear Least Squares Computations,* New York: Marcel Dekker.

Hoel, P. G., Port, S. C., & Stone, C. J. (1971). *Introduction to Probability Theory.* Boston: Houghton Mifflin.

Kariya, T., & Kurata, H. (2004). *Generalized Least Squares,* Chichester, West Sussex, England; Huboken, NJ: Wiley.

Lawson, C. L., & Hanson, R. J. (1995). *Solving Least Squares Problems,* Philadelphia: Society for Industrial and Applied Mathematics.

Montgomery, D., & Peck, E. (2001). *Introduction to Linear Regression Analysis,* New York: Wiley.

Rao, C. R. (1999). *Linear Models: Least Squares and Alternatives,* New York: Springer.

Van Huffel, S. (1997). *Recent Advances in Total Least Squares Techniques & Errors in Variables Modeling,* Philadelphia: Society for Industrial and Applied Mathematics.

Breadth Demonstration

In a scholarly paper of about 30 pages, I will develop the major theories of least squares prediction including the development of mathematics for models that are linear in their unknown coefficients, including linear models, polynomial models, and linearizable models. Applications necessary to illustrate the development of the theory will also be included, including some examples that will later be used in the Application section of the KAM. The paper will lay the foundation for the detailed Depth examination of my recent discovery about data dropping out under certain conditions when making predictions using least squares models that are linear in their unknown coefficients.

Depth Objectives

Develop my new discovery concerning least squares predictions that involves some data unexpectedly dropping out of prediction calculations. Determine exactly what data drops out in univariate straight line modeling situations, and under what specific circumstances this happens. Where possible, develop formulas and/or theorems to describe this behavior. This study will build upon foundation laid in the Breadth section of the KAM, and use the mathematics studied in this section. The Depth section will also include an annotated bibliography of at least 15 current papers from the last 5 years on least squares and applications of least squares theory.

Preliminary Depth References

Asparouhov, T., (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12(3),* 411–434.

Bagarinao, E., Matsuo, K., Nakai, T., & Sato, S. (2003). Estimation of general linear model coefficients for real-time application. *NeuroImage, 19(2),* 422–429.

Bekker, P., & Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica, 59(3),* 239–257.

Burr, T., & Fry, H. (2005). Biased Regression: The case for cautious application. *Technometrics, 47(3),* 284–296.

Chatterji, P., & Markowitz, S. (2005). Does the length of maternity leave affect maternal health? *Southern Economic Journal, 72(1),* 16–41.

Eye, A., Brandtst, J., & Rovine, J. (1993). Models for prediction analysis. *Journal of Mathematical Sociology, 18(1),* 65–80.

Goldenshluger, A., & Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis, 84(1),* 40–60.

Horváth, L., Husková, M., Kokoszka, P., & Steinebach, J. (2004). Monitoring changes in linear models. *Journal of Statistical Planning & Inference, 126(1),* 225–251.

Jia, X., Rao, B.. & Zhang, H. (2003). On weak consistency in linear models with equi-correlated random errors. *Statistics, 37(6),* 463–473.

Kaplow, J., Dodge, K., Amaya-Jackson, L., & Saxe, G. (2005). Pathways to PTSD, Part II: Sexually Abused Children. *American Journal of Psychiatry. 162(7),* 1305–1310.

Manning, W., Basu, A., & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes. *Journal of Health Economics, 24(3),* 465–488.

Olive, D., & Hawkins, D. (2005). Variable Selection for I D Regression Models. *Technometrics, 47(1),* 43–52.

Onatski, A., & Williams, N. (2003). Modeling model uncertainty. *Journal of the European Economic Association, 1(5),* 1087–1122.

Shi, L., & Ojeda, M. (2004). Local influence in multilevel regression for growth curves. *Journal Of Multivariate Analysis, 91(2),* 282–304.

Sievers, G., & Abebe, A. (2004). Rank estimation of regression coefficients using iterated reweighted least squares. *Journal of Statistical Computation & Simulation, 74(11),* 821–831.

Tappeiner, G., Tappeiner, U., Tasser, E., & Holub, H. W. (2004). Statistical aspects of multilayer perceptrons under data limitations. *Computational Statistics & Data Analysis, 46(1),* 173–188.

Trinkoff, A., Johantgen, M., Muntaner, C., & Le, R. (2005). Staffing and Worker Injury in Nursing Homes. *American Journal of Public Health, 95(7),* 1220–1225.

Turlach, B., Venables, W., & Wright, S. (2005). Simultaneous Variable Selection. *Technometrics, 47(3),* 349–363.

Vorobyov, S., Yue, R., Sidiropoulos, N., & Gershman, A. (2005). Robust iterative fitting of multilinear models. *IEEE Transactions on Signal Processing, 53(8),* 2678–2699.

Warwick, J., & Jones, M. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation & Simulation, 75(7),* 581–588.

Wilcox, R. (2005). Estimating the conditional variance of Y, given X, in a simple regression model. *Journal of Applied Statistics, 32(5),* 495–502.

Depth Demonstration

In a scholarly paper of approximately 20 pages, explain the new statistical discovery described above, including underlying assumptions, mathematical theorems and/or formulas, and any special cases. Illustrate this finding with specific examples within the paper. The Depth demonstration will also include an annotated bibliography of at least 15 articles that will be no more than 5 years old.

Application Objectives

Apply the new finding described in the Depth section of the KAM to the application problem that inspired it. This involves prediction of the number of students that will enroll in future years in one or more of Newport Mesa Unified, Saddleback Valley Unified, Las Virgenes Unified, and Oak Park Unified School Districts in CA, given demographic and previous enrollment data. Make the enrollment predictions using a classical least squares model, and then describe the problems that arise in the cases when data drops out of key predictions. An examination will be made of possible alternatives for modeling under conditions where data drops out, including a brief look at the absolute value and minimax norms. The application will be tied in with the work done in the Breadth and Depth sections of the KAM.

Preliminary Application References Materials

Baglivo, J. A. (2005). *Mathematica Laboratories for Mathematical Statistics: Emphasizing Simulation and Computer Intensive Methods.* Philadelphia, PA: Society for Industrial and Applied Mathematics, American Statistical Association.

Application Demonstration

In a scholarly paper of approximately 20 pages, including graphical material, I will thoroughly analyze the application problem that inspired the Depth section of this KAM. The specific problem involves predicting the number of fourth graders that will be enrolling in future years in one or more of the school districts named in the Application objectives above. Demographic and previous enrollment data will be used in the modeling, The work will describe the source of the data, underlying assumptions as well as analysis and recommendations for predictions using the data.

Walden University

Self-Evaluation: Knowledge Area Modules (KAMs)

Students: Attach this completed form to each KAM prior to forwarding to the Faculty Assessor
(Please use reverse side as necessary)

Student Name__Teresa (Terri) Bittner_______**Date**______10/21/05________

KAM number ___7__ **Title** _____Least Squares Regression Analysis: A New Discovery______

1. What knowledge/experience did you bring to this KAM? How did you capitalize/expand on this base?

I had a deep background in least squares modeling and data analysis from my master's degree and previous work experience. I made a new discovery about least squares earlier this year, and used my previous experience to develop the theory behind the discovery.

2. Describe the quality of the **Breadth** section in the light of the intellectual and communication skills demonstrated in this KAM.

I wrote a 30 page essay about least squares modeling. I briefly discussed the history and origin of least squares regression, developed the theory behind it, and illustrated the theory all the way through with examples using real data.

3. In the **Depth** section, what key ideas/concepts most engaged your thinking and imagination relative to your area of study?

My new discovery was fascinating to develop. Plowing new ground in mathematics was a new experience for me, and I found it engaging enough that I plan to write my dissertation about this discovery. Of course it will have to be developed further than what I did in this paper, but it will be the same kind of work.

4. Expound on the most meaningful theoretical construct studied and applied to your professional setting in the **Application** section. What can you do differently/better as a result of this KAM?

Developing the tools to state and prove a new theorem is a huge leap in critical thinking. There is no book or reference to help because the work has never been done before. It is this kind of thinking that most helps me in my teaching, and I hope to continue to develop it. Though I have always been a critical thinker, I believe that the continued development of critical thinking will help me to be a better teacher.

5. Briefly describe the most important **Social Issue** covered in this KAM.

Any theoretical result that shows a potential problem in one of the most widely used prediction techniques has potential significance for researchers in many fields. These include but are not limited to business and economics, psychology, sociology, engineering, and astronomy to name a few. The goal for this dissertation is to accurately describe the circumstances in which some data will not affect predictions in increasingly general cases of models that are linear in the unknown coefficients and to show examples from published research in the social sciences where predictions may be flawed due to this phenomenon. Published or ongoing research that fit these criteria could potentially be examined and/or modified in light of the fact that some data is inadvertently not being used in predictions.

BREADTH DEMONSTRATION

BREADTH TABLE OF CONTENTS

BREADTH DEMONSTRATION

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 is 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. The majority of this paper will involve a description of the technique of least squares, its assumptions and limitations, and the various results that come out of the technique. The uses of regression will be described, followed by a description and example of a simple linear model, a discussion on testing the adequacy of models, hypothesis testing, a discussion of other models that are nonlinear in the measured data but linear in the unknown parameters, and weighted regression. Finally, there will be a brief description and discussion of difficulties in using two alternative norms. These include the absolute value norm and the minimax norm. Unless otherwise specified, it should be assumed in this paper 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, 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 is 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, 2001). The only requirement in this case is that the relationship between the variables that existed when the data was 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 has large or nonrandom measurement errors or is 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 is 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, 2001). It will also become clear later in this section of the paper and in the depth section of the paper that some data affects 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 the depth section of this paper 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*. The point *A* is an example of an outlier shown in Figure 1 below. 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.

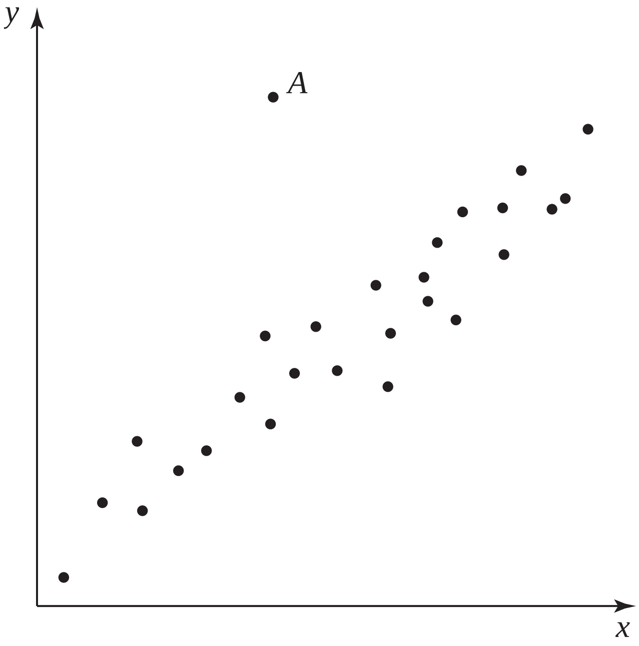

Figure 1. An Example of an Outlier.

While exhaustive coverage of the cautions and limitations of regression would require hundreds of pages and therefore cannot be thoroughly covered in this paper, 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 subsection 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 paper, 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 any one error value 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, 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

$$\begin{aligned}
-2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\
\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i &= n\hat{\beta}_0 \\
\bar{y} - \hat{\beta}_1 \bar{x} &= \hat{\beta}_0
\end{aligned}$$

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, 2001; Younger, 1979).

One of the most important ways to evaluate the adequacy of a model is by computing and looking at the *residuals* of the data. A residual is the difference between the observed value of y and the corresponding estimate $\hat{y}$. Mathematically, the residuals for a data set can be written as

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right), \text{ for } i = 1, 2, \ldots, n$$

The specific use of residuals will be discussed in the next subsection of this paper.

Example 1.

Byers and Williams (1987) studied how temperature impacted the viscosity of toluene–tetralin blends. The following table gives the data for blends with a 0.6 molar fraction of toluene.

Temperature (°C)	Viscosity (mPa • s)
25.0	0.9208
35.0	0.8904
45.0	0.7874
55.0	0.7101
65.0	0.6412
75.0	0.5843
85.0	0.5350
95.0	0.4949

The scatter diagram in Figure 2 below shows a strong linear relationship. As temperature increases, viscosity decreases linearly. Therefore, it is reasonable to fit a linear model to this data.

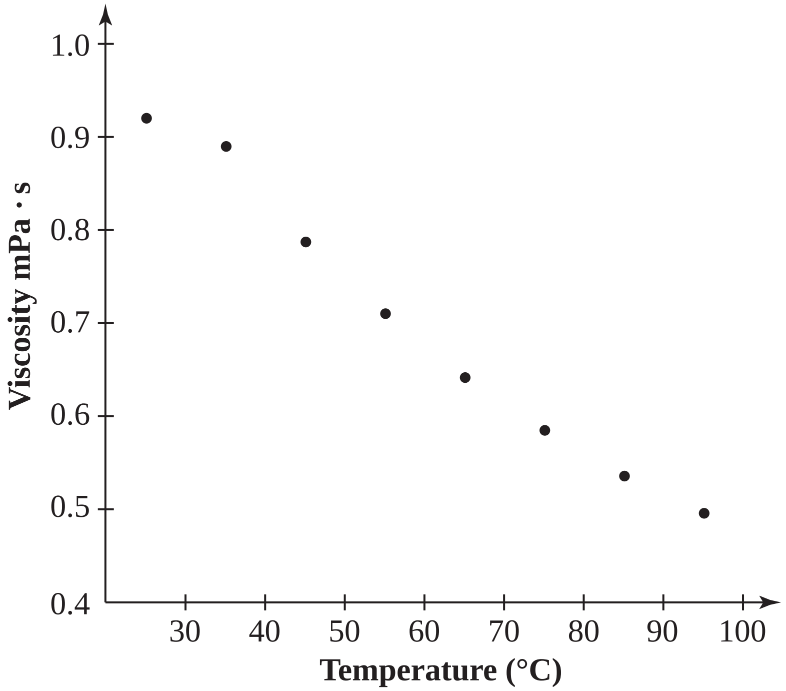

Figure 2. Scatterplot of Temperature vs. Viscosity.

To compute the coefficients for the linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, first note that $\sum_{i=1}^{8} x_i = 480$, $\sum_{i=1}^{8} y_i = 5.5641$, $\sum_{i=1}^{8} x_i y_i = 306.664$, and $\sum_{i=1}^{8} x_i{}^2 = 33{,}000$. $\bar{x} = 60$ and $\bar{y} = 0.695513$ can be easily computed using $\sum x_i$ and $\sum y_i$. Finally, substituting into the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{8} y_i x_i - \frac{\left(\sum_{i=1}^{8} y_i\right)\left(\sum_{i=1}^{8} x_i\right)}{8}}{\sum_{i=1}^{8} x_i{}^2 - \frac{\left(\sum_{i=1}^{8} x_i\right)^2}{8}} \\
&= \frac{306.664 - \frac{(5.5641)(480)}{8}}{33{,}000 - \frac{(480)^2}{8}} \\
&= \frac{-27.182}{4200} \\
&= -0.0064719
\end{aligned}$$

and

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 0.695513 - (-0.0064719)(60) \\ &= 1.083827\end{aligned}$$

Therefore, the linear model is $\hat{y} = 1.083827 - 0.0064719x$. We can, for example, estimate the viscosity when the temperature is 60°C. The estimate is

$$\begin{aligned}\hat{y} &= 1.083827 - 0.0064719(60) \\ &= 0.695513\end{aligned}$$

A plot of the data with the regression line is shown in Figure 3 below.

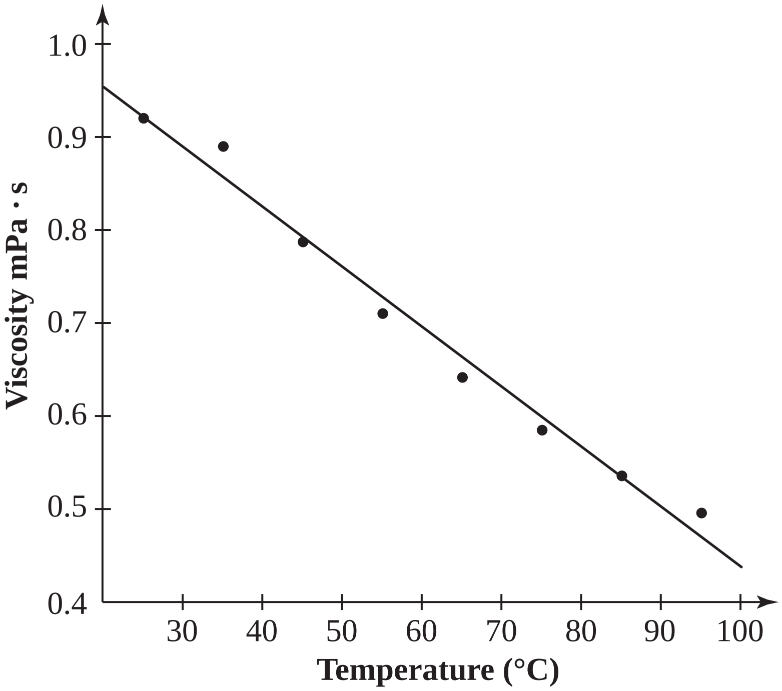

Figure 3. Temperature vs. Viscosity with Regression Line.

The residual plot is shown in Figure 4 below.

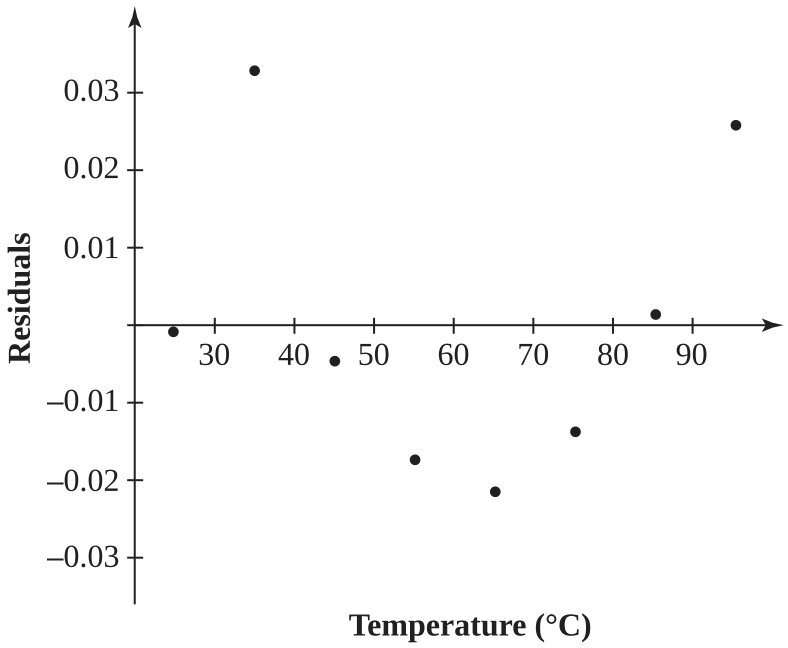

Figure 4. Residuals for Temperature vs. Viscosity Model.

Though the magnitude of the residuals is clearly small, indicating small errors between the model and the measured data, the residuals show a curved pattern. When a model fits well, a residual plot resembling random noise around zero would be expected. A curved residual plot often indicates a nonlinearity in the data. This hints that simply looking at a scatterplot of the data may not be sufficient in determining the adequacy of a linear model. Methods to determine the adequacy of models will be discussed in the next subsection of this paper, and the data from Example 1 will be more analyzed completely. √

Testing the Adequacy of a Model

According to Montgomery (2001), there are four assumptions that are made when doing the kind of regression analysis outlined in previous sections. These are:

1. There is a linear relationship between the regressor variable (*x* data) and the response variable (*y*), at least an approximate one.

2. In the linear model $y = x\beta + \varepsilon$, the expected value of ε is equal to zero. In other words, $E(\varepsilon) = 0$.

3. In the linear model $y = x\beta + \varepsilon$, the variance (σ^2) of ε is constant.

4. The errors are uncorrelated.

5. The errors behave according to a Normal distribution.

These assumptions are the first things that should be looked at in testing the adequacy of a model. If one or more of the assumptions is known or strongly suspected to be untrue, the model should be used only with great caution. For example, if there is a known bias in the measurements of the *y* data, assumption number 2 is false and any linear regression will be flawed.

Another test of the adequacy of a model begins with some additional notation. To begin, recall that one of the forms of the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ is

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}.$$

The denominator of $\hat{\beta}_1$ is the corrected sum of squares of the *x*-data. This can be written more simply as $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Similarly, the corrected sum of squares of the *y* data can be written as $SS_{\mathrm{T}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. It is also useful to compute some estimate of the variance, also called σ^2, of the *y* data. It is best to compute some estimate of the variance that is independent of the adequacy of the model, but this is only possible when multiple measurements of the response variable are available for each value of *x*. When this is not possible, then the variance is generally measured by computing the sum of squares of the residuals, or $SS_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Note that SS_{T} measures the variability in *y* without considering the effect of the *x* data, and SS_{res} measures the variability in *y* that is left after considering the *x* data. The *coefficient of determination,* widely known as R^2, can be thought of as the proportion of variation that is explained by the *x* data, and is computed as

$$\begin{aligned} R^2 &= 1 - \frac{SS_{\text{res}}}{SS_{\text{T}}} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned}$$

R^2 provides one indication of the adequacy of a model by providing the percentage of the variability in *y* that is explained by *x*. A low value of R^2 therefore indicates that the relationship between *x* and *y* are not explained very well by a linear model (Farebrother, 1988; Montgomery, 2001; Younger, 1979).

The sample correlation coefficient between *x* and *y,* usually called *r*, is simply the square root of the coefficient of determination R^2. While R^2 is dependent upon the regression model because of the term in its formula involving $\hat{y}$, the sample correlation appears to be dependent only on the *x* and *y* data itself. However, the sample correlation coefficient is closely related to $\hat{\beta}_1$, and hence to the regression model, by the equation

$$\hat{\beta}_1 = r\left(\frac{SS_T}{S_{xx}}\right)^{1/2}, \text{ where } r = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}} \; .$$

The sample correlation coefficient *r* gives an estimate of the degree of linear association between the *x* and the *y* data, while $\hat{\beta}_1$ is actually the estimated slope of the linear model and therefore provides an estimate of the change in *y* per unit change in *x*. Recall that this is the definition of the slope of a line.

Example 2.

The coefficient of determination for the data in Example 1 can be computed as follows: Recall from Example 1 that $\bar{x} = 60$ and $\bar{y} = 0.695513$. Then

and $\sum_{i=1}^{8}(y_i - \bar{y})^2 = 0.178719$ and $\sum_{i=1}^{8}(y_i - \hat{y}_i)^2 = 0.00279362$, so

$$R^2 = 1 - \frac{0.00279362}{0.178719} \approx 0.984, \text{ and } r = \sqrt{R^2} \approx 0.992.$$

Clearly, there is a high degree of linearity between the variables. √

An important note is that a value of R^2 or *r* close to 1 or –1 cannot not by itself indicate that a model is adequate. For example, from 1952 to 1976 there seemed to be a perfect correlation between the league that won the World Series and the political party that won the presidential election. In each of those years an American League win in the World Series resulted in a Republican win for the White House. A National League win in the World Series resulted in a Democratic win for the White House (World Almanac, 2005). Of course, this all changed after 1976, and the reasonable conclusion is that this relationship was something that happened by random chance rather than a real relationship between the variables. Likewise, a mathematical relationship between variables should not be the only thing tested to determine the adequacy of a model. While such tests can be helpful, the relationship between the data should make reasonable sense as well, and should not be a random relationship that might have been caused by chance.

An analysis of residuals is a second important way to assess the adequacy of a regression model. Recall that the residuals, e_i, are the difference between the *y* data values and the estimates for those *y* data values obtained from the least squares model. They can be thought of as the error between the model and the *y* data. In mathematical terms,

$$e_i = y_i - \hat{y}_i,\ i = 1, 2, ..., n.$$

Plotting each of the residuals $(x_i,\ e_i)$ can be a very useful tool in assessing how well the model fits the data. The shape of these plots can be an extremely helpful way to analyze the adequacy of a model. If a model is adequate, the residuals should look like random noise with a mean of 0. An example is shown in Figure 5. Any pattern shown by residuals is generally an indication that the model is not adequate. Fortunately, methods have been devised to correct for many of the residual patterns that indicate inadequate models. For example, an outward opening funnel pattern like the one in Figure 6 indicates that the variance of the errors increases as *y* increases, while the symmetric "bow" pattern in Figure 7 is often seen when *y* is a proportion (or percentage) between 0 and 1. In this case, the variance will be greater when *y* is near 0.5 than near 0 or 1. Transformations of *x* or *y* are usually used to correct for patterns that indicate a relationship between the data and the errors. The *y*-data can be transformed by taking its square root if the variance is proportional to the expected value of *y* as in the residual patterns from Figures 6 and 7. Another classic example is the quadratic residual pattern shown in Figure 8. Notice that the residual pattern from Example 1, shown in Figure 3, has the same type of quadratic pattern as that shown in Figure 8. This type of pattern indicates that the data may not have a linear relationship. In this case a quadratic term or some other regressor variable may need to be added to the model. Models of this type will be discussed later in this paper. There are also standard transformations for various residual patterns that can be found in textbooks and other sources (Larsen & McCleary, 1972; Montgomery, 2001).

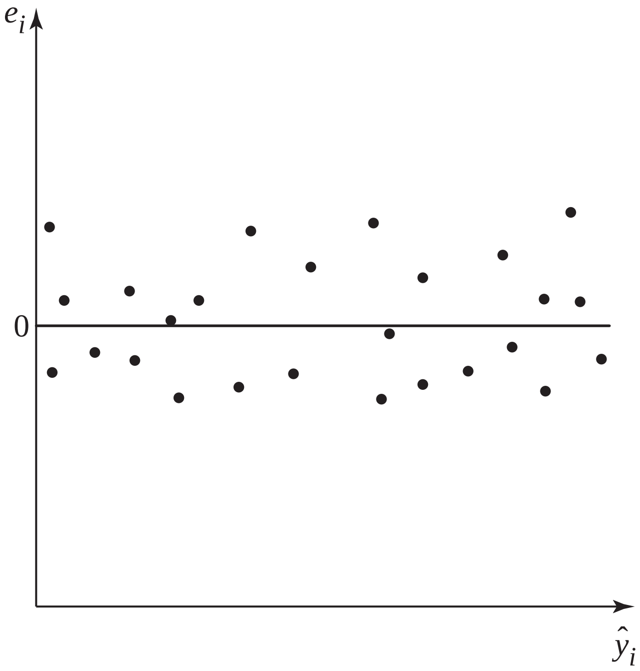

Figure 5. Ideal Residual Plot.

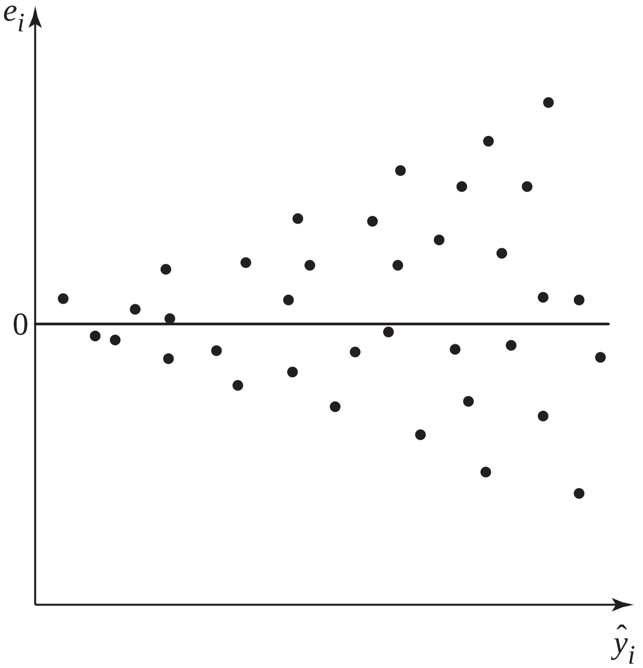

Figure 6. Outward Opening Funnel Residual Plot.

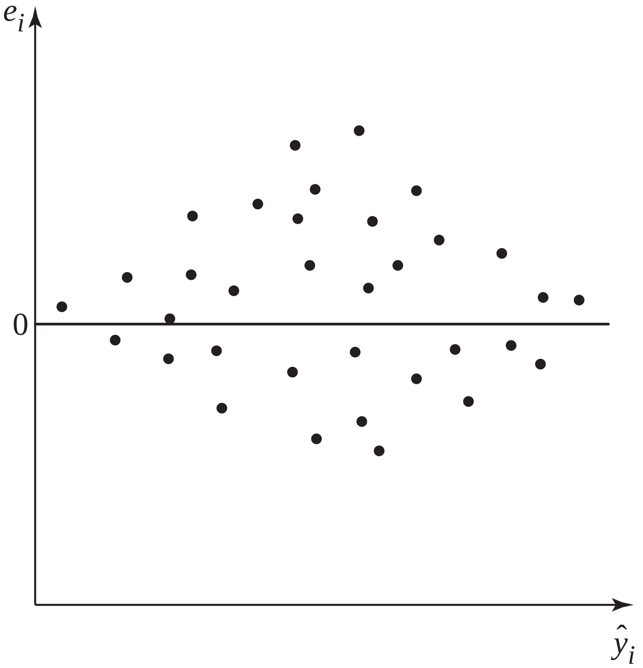

Figure 7. Symmetric “Bow” Pattern Residuals.

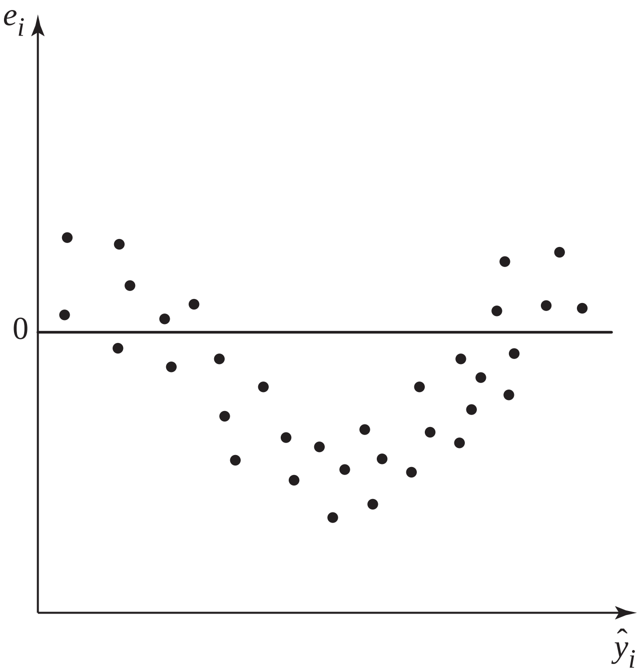

Figure 8. Quadratic shape in Residual Plot.

In the case of Example 1, any nonlinearity may be so subtle that a change in the model may not be necessary. Notice the magnitude of the residuals is very small, the coefficient of determination is very close to 1, and the data covers the entire temperature range that is likely to be of interest. However, given the nearly perfect quadratic pattern of the residuals, the idea of a quadratic term in the model cannot be ignored. The procedure for doing this will be discussed later in the paper. This example demonstrates the importance of looking at multiple indications of the adequacy of a model rather than relying on just one test. The way to make the final determination about the model from Example 1 is to add a quadratic term and examine the resulting coefficients. This will be done later in the paper.

Another important aspect of linear modeling is the detection of outliers. Recall that outliers are data points that are atypical compared to the rest of the data. Outliers can result from explainable events such as faulty equipment, unusual measurement errors, or recording errors. This type of outlier should generally be removed from the data set before modeling. Other outliers, however, may be unusual but legitimate observations. It can be dangerous to simply remove this type of outlier because doing so will falsely suggest a more homogenous model with lower variance than what actually exists. In either case, identifying and analyzing outliers can even result in improvement of the model and in greater knowledge about the relationship between the variables. Stefansky (1971, 1972) suggested a test for identifying outliers based on the maximum normalized residual $\frac{|e_i|}{\sqrt{\sum_{i=1}^{n} e_i{}^2}}$, and it is a fairly simple test to use. Another way to identify possible outliers is to look at the residual plot. Residuals that have a large magnitude compared to the other residuals may indicate outliers. Once identified, the possible cause of the outlier needs to be assessed, and then appropriate action, such as deleting the data, should be taken if necessary (Montgomery, 2001).

Yet another method of testing the adequacy of a model is by hypothesis testing. This can be done in several ways, and will be discussed along with confidence intervals in the next subsection.

Hypothesis Testing and Confidence Intervals

It is sometimes useful to test hypotheses about the coefficients of a model or to create confidence intervals for these coefficients. Suppose we want to test the null hypothesis that the slope is a particular constant, vs. the alternative that it is not. This hypothesis can be written as

$$\begin{aligned} H_0 &: \ \beta_1 = c_1 \\ H_1 &: \ \beta_1 \neq c_1 \end{aligned}$$

where c_1 is any constant. In this case, the alternative hypothesis can be either greater than or less than c_1. Since there is an inherent assumption in linear modeling that the errors, ε_i, are normally distributed and independent with a mean of 0 and a variance of σ^2, $N(0,\ \sigma^2)$, the dependent variables y_i are normally distributed and independent with mean $\beta_0 + \beta_1 x_i$ and variance σ^2, denoted $N\left(\beta_0 + \beta_1 x,\, \sigma^2\right)$. Now $\hat{\beta}_1$ is a linear combination of the y_i values, so it also has a normal distribution with a mean of β_1 and a variance of $\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sigma^2}{S_{xx}}$, denoted $N\left(\beta_1,\ \dfrac{\sigma^2}{S_{xx}}\right)$. Therefore,

$$Z = \frac{\hat{\beta}_1 - c_1}{\sqrt{\sigma^2 / S_{xx}}}$$

has a normal distribution with a mean of 0 and a variance of 1. This is the Standard Normal distribution. Therefore, the *Z*-statistic can be used to test the hypothesis if the variance σ^2 is known. As is more often the case, however, the variance is unknown, and must be estimated. The estimate $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \dfrac{SS_{\text{res}}}{n-2}$ is called an *unbiased estimator* of σ^2, since $E\left(\hat{\sigma}^2\right) = \sigma^2$. Therefore, σ^2 in the *Z*-statistic can be replaced by its unbiased estimator to yield a new statistic

$$t = \frac{\hat{\beta}_1 - c_1}{\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}}},$$

and this statistic has a *t* distribution with $n-2$ degrees of freedom if the null hypothesis is true. Therefore, the null hypothesis is rejected if $|t| > t_{\alpha/2,n-2}$, where α is the significance level of the test, and $t_{\alpha/2,n-2}$ is called the *critical value* of the test. Otherwise, we fail to reject the null hypothesis.

A hypothesis test to determine whether the intercept, $\hat{\beta}_0$, is a particular constant can be tested using a similar method. In this case the hypothesis test can be written as

$$\begin{aligned} &H_0: \ \beta_0 = c_0 \\ &H_1: \ \beta_0 \neq c_0 \end{aligned}$$

For this test, the test statistic would be

$$t = \frac{\hat{\beta}_0 - c_0}{\sqrt{\frac{SS_{\text{res}}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

by an argument similar to that of the hypothesis test for the slope. Again, the null hypothesis is rejected if $|t| > t_{\alpha/2,n-2}$. Otherwise, we fail to reject the null hypothesis.

Example 3.

Test the hypothesis for the data from Example 1 to see whether there is a linear relationship. The hypothesis is formed as

$$\begin{aligned} &H_0: \beta_1 = 0 \\ &H_1: \beta_1 \neq 0 \end{aligned}$$

In other words, the conclusion will be that a linear relationship exists if the null hypothesis is rejected. In this case, $\hat{\beta}_1 = -0.0064719$, $n = 8$, $\sum_{i=1}^{8}(y_i - \hat{y})^2 = 0.00279362$, and $SS_{xx} = \sum_{i=1}^{8}(x_i - \bar{x})^2 = 4200$. So, the *t*-statistic is computed as

$$t = \frac{\hat{\beta}_1 - c_1}{\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}}} = \frac{-0.0064719}{\sqrt{\frac{0.00279362}{8-2} \Big/ 4200}} = -19.4379$$

Since $t_{0.025,6} = 2.447$, the null hypothesis is rejected, and it is concluded that the data has a linear relationship.√

In addition to testing hypotheses about the regression coefficients, confidence intervals may also be constructed for $\hat{\beta}_1$ and $\hat{\beta}_0$. The statistics are based on the fact that the errors are independent and normally distributed. Based on this, both $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}}}$ and $\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{SS_{\text{res}}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$ have *t*-distributions with $n-2$ degrees of freedom. Therefore, a $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_1$ can be computed as

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}}$$

and a $100(1-\alpha)\%$ confidence interval for $\hat{\beta}_0$ can be computed as

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\frac{SS_{\text{res}}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\frac{SS_{\text{res}}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Example 4.

Construct a 95% confidence interval for $\hat{\beta}_1$ using the data from Example 1. From Example 3, we know that $\hat{\beta}_1 = -0.0064719$, $t_{0.025,6} = 2.447$, and $\sqrt{\frac{SS_{\text{res}}}{n-2} \Big/ S_{xx}} = 0.000332953$. Therefore, the 95% confidence interval for $\hat{\beta}_1$ is

$-0.0064719 - 2.447(0.000332953) \le \beta_1 \le -0.0064719 + 2.447(0.000332953)$ or

$-0.00728664 \le \beta_1 \le -000565716$ √

Multiple Linear Regression

Often data is 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 paper 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

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

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

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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.

There is also a geometric interpretation of the solution to the linear least squares problem. The $\hat{\beta}$ vector can be thought of as the *unique* orthogonal projection of *Y* onto the space of *X* as shown in the figure below.

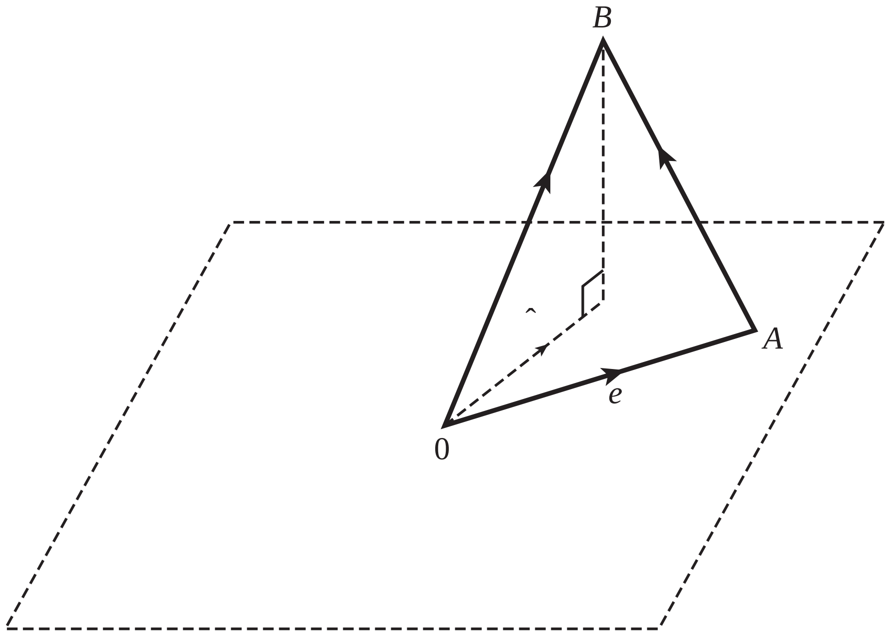

Figure 9. Orthogonal Projection of *Y* onto the space of *X*.
(Figure adapted from Seber, 1977)

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 least squares techniques. The methods for solving nonlinear problems are generally iterative.

Example 5.

Fit the data from Example 1 to a model that is quadratic in the *x*-data and examine the results.

To fit the data from Example 1 to a model quadratic in *x*, use the *X* matrix

$$X = \begin{bmatrix} 1 & x_1 & x_1{}^2 \\ 1 & x_2 & x_2{}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_8 & x_8{}^2 \end{bmatrix}, \text{ the beta vector } \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \text{ and the } Y \text{ vector } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_8 \end{bmatrix}$$

The matrix $X'X$ is a 3×3 square matrix. Multiplication yields

$$X'X = \begin{bmatrix} 8 & 480 & 33{,}000 \\ 480 & 33{,}000 & 2{,}484{,}000 \\ 33{,}000 & 2{,}484{,}000 & 198{,}285{,}000 \end{bmatrix}, \text{ and } X'X^{-1} = \begin{bmatrix} \frac{5923}{896} & -\frac{131}{560} & \frac{41}{22{,}400} \\ -\frac{131}{560} & \frac{37}{4200} & -\frac{1}{14{,}000} \\ \frac{41}{22{,}400} & -\frac{1}{14{,}000} & \frac{1}{1{,}680{,}000} \end{bmatrix},$$

and the optimal linear model is $y_i = 1.10348 - 0.00664714 x_i - 5.67857 \times 10^{-6} x_i{}^2$. Since the quadratic term is very close to 0, and given the high correlation found in previous analysis of this data, it can now be concluded that the quadratic term is 0 and the data fits the linear model found in Example 1.√

Residual analysis, hypothesis testing, confidence intervals, and other model adequacy tests can be performed on models with multiple regressor variables such as the ones described in this subsection, but the procedures are different from those of the simple linear model, and require techniques in linear algebra and knowledge of mathematical statistics that are too complex to discuss in this paper. Nevertheless, the procedures and techniques exist and are well known. They are covered in any advanced textbook that covers linear regression using linear algebra techniques (Seber, 1977).

Weighted Regression

The computation of the optimal least squares coefficients requires that the variance of the errors be constant. When this is known to be false, or when residual plots indicate that it may be false, weighted least squares is sometimes used to correct for this problem. Suppose that $\mathrm{Var}(\varepsilon) = \sigma^2 V$, where *V* is a known $n \times n$ matrix. If the *V* matrix is diagonal but has unequal diagonal elements, then the variances of the observations are uncorrelated but unequal. If any of the off-diagonal elements are nonzero, then the variances of the observations are correlated with each other. In the case of correlated variances, an alternative modeling technique called *generalized least squares* is used. This discussion will be limited to the case where the *V* matrix is diagonal. In this case, the modeling technique is called *weighted least squares.* Suppose that the *V* matrix is of the form

$$V = \begin{bmatrix} \frac{1}{w_1} & & & 0 \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_n} \end{bmatrix}$$

and define $W = V^{-1}$. Note that *W* is also a diagonal matrix and its “weights” are $w_1,\ w_2,\ \cdots,\ w_n$. The sum of squares function to be minimized is defined as

$$\begin{aligned} L(\beta) &= \varepsilon' V^{-1} \varepsilon \\ &= (Y - X\beta)' V^{-1} (Y - X\beta) \end{aligned}$$

Using the same arguments and calculations as in the previous subsection, and the fact that $W = V^{-1}$, the least squares normal equations are

$$(X'WX)\hat{\beta} = X'WY.$$

This can be solved for $\hat{\beta}$ by multiplying the left sides of both sides of the equation by $(X'WX)^{-1}$. This yields

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

also called the *weighted least squares estimator*.

Note that the weights must be known or chosen in order to use this technique. If weighting is being used to correct for observations with variances that are not constant, the weights are sometimes chosen so that they are inversely proportional to the variances of the measurement errors. In many cases like this, the first choice for weights is not ideal. The weights can be estimated, a model constructed and analyzed, and then the weights adjusted based on the results. This can be repeated until satisfactory results are obtained (Farebrother, 1988; Montgomery, 2001; Seber, 1977).

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. This kind of estimator is often called a maximum liklihood estimator since the function ρ is related to the likelihood function for an appropriate choice of the error distribution.

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 $(j = 0,\ 1,\ \cdots,\ k)$ 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 more 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.

Conclusion

Statistics is a huge area of mathematics, and linear regression analysis is in itself a huge area of study within statistics. A paper such as this can provide only a brief glimpse of a complex and ever evolving analysis that has been going on for over 200 years (Plackett, 1972). This paper has nevertheless laid the groundwork for the depth section where a new finding regarding least squares predictions will be revealed. The general area of prediction is one of the most important in least squares analysis since it is often the purpose for modeling. The foundational work from this section will be used in the depth section to describe and prove a new theorem showing that some data points drop out of prediction computations altogether in a way that is virtually invisible.

.

References

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

Byers, C. & Williams, D. (1987). Viscosities of Binary and Ternary Mixtures of Polyaromatic Hydrocarbons. *Journal of Chemical and Engineering Data, 32,* 349–354.

Farebrother, R. W. (1988). *Linear Least Squares Computations,* New York: Marcel Dekker.

Larsen, W. & McCleary, S. (1972). The use of partial residual plots in regression analysis. *Technometrics, 14,* 781–790.

Montgomery, D., & Peck, E. (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.

Seber, G. (1977). *Linear Regression Analysis.* New York: John Wiley & Sons.

Stefansky, W. (1971). Rejecting outliers by maximum normed residual, *Annals of Mathematical Statistics, 42,* 35–45.

Stefansky, W. (1972). Rejecting outliers in factorial designs, *Technometrics, 14,* 469–479.

Van Huffel, S. (1997). *Recent Advances in Total Least Squares Techniques & Errors in Variables Modeling,* Philadelphia: Society for Industrial and Applied Mathematics.

Younger, M. S. (1979). *A Handbook for Linear Regression.* Belmont, CA: Wadsworth Publishing Company, Duxbury Press.

The World Almanac (2005). New York: St. Martin's Press.

DEPTH DEMONSTRATION

DEPTH TABLE OF CONTENTS

Annotated Bibliography

Depth Essay

Citation 1: A Comparison of Weighting Methods in Latent Variable Modeling

Asparouhov, T. (2005). Sampling Weights in Latent Variable Modeling, *Structural Equation Modeling, 12(3),* 411–434.

Critical Summary

It is well known that unequal probability of selection in sampling causes significant biases if the unequal probability of selection is not taken into account when the data analysis is performed. Different software packages currently on the market use different techniques to take this problem into account. This article analyzes the problems associated with unequal probability of selection and then compares several techniques used by *M*-Plus, MLWIN, and HLM/SAS, which are three widely used statistical packages.

Psuedomaximum Likelihood (PML), Weighted Maximum Likelihood (WML), Weighted Least Squares (WLS), and Unweighted Least Squares (ULS), were compared in modeling under various sampling techniques. These techniques were prepared by simulating the sampling techniques. The various methods under consideration were then used for analysis of the data. It was found that PML works consistently under any sampling scheme, but consistently produces confidence intervals that are too short, and classic chi-square difference testing rejects more frequently than the designated rejection level. WML was found to be biased in the same direction, However, when robust chi-squared tests performed adequately. Even WLS produces inflated Chi-squared statistics and underestimates confidence intervals,

When the software packages were compared, *M*-Plus was found to have consistently adequate methods implemented in its software, while the HML estimates were found to have substantial biases. The methods used by MLWIN were mostly found to be equivalent to *M*-Plus. While *M*-Plus did show a major flaw with one method, it was generally shown to be equal or superior to the other packages overall given the particular techniques that were compared.

Critical Analysis

This article is an important caution to statisticians using software. Software packages do not always use the same methods or algorithms, especially when complex analysis is used, and the user must note the importance of unbiased results before choosing a software package when unequal probability of selection is a problem with the sampling of data. The author did a thorough analysis of the problems involved in unequal probability of selection and the biases that result. It is clear that certain robust techniques must be used in order to avoid bias.

The author fairly noted problems found with the software and limitations in his own analysis, including limitations with using simulated data and the dangers of making broad inferences from the results. However, the author did seem to be quite biased towards *M*-Plus, even before he discussed the completed analysis. That particular software package was discussed far more than the others, and the discussion comparing the three did not seem neutral, especially given that MLWIN was found to have virtually identical results to *M*-Plus for the most important tests performed. Nevertheless, this article is worthwhile reading for anyone trying to select a software package for problems with complex sampling and analysis.

Citation 2: Biased Regression: The case for cautious application

Burr, T., & Fry, H. (2005). Biased Regression: The case for cautious application. *Technometrics, 47(3),* 284–296.

Critical Summary

It is well known that biased regression methods can produce much better results in prediction that ordinary least squares. However, if applied recklessly these methods can also produce worse results than ordinary least squares, and their implementation is fraught with complications. It is for this reason that many statisticians have recommended against the use of all biased regression techniques in real problems. However, in the particular application of chemometrics, biased regression is still widely used because the assumptions needed to have a high probability of bettering least squares are usually met in this application. Even so, the specific type of biased regression that is usually used may not provide the best results compared to other biased regression results.

These authors advocate a cautious application of biased regression in real applications. Principal Components Regression, Ridge Regression, Minimax Regression, Partial Least Squares, and basic James-Stein-Like Estimators were covered in this article. The authors suggest testing all five of these biased regression techniques along with ordinary least squares, and applying a subset of specific criteria to determine which technique to use in their application. This method was tested via a simulation of chemometric data, and the simulation considered all 15 subsets of their four criteria over a wide variety of data.

Critical Analysis

The idea of applying biased regression methods cautiously, especially using objective criteria to ascertain which technique is best is a thoughtful and useful method of choosing among techniques. The criteria were carefully selected to remain unbiased over the range of methods, including ordinary least squares, and the authors did not throw out either ordinary least squares or any of the biased regression techniques. The suggestion is that each has its place, depending on the specific data to be modeled. The one caution is that all simulation data was from one application. It is possible that the suggested technique would not provide such clean results in other situations that do not ordinarily lend themselves to biased regression in the first place. Still, the study is useful, and it is particularly interesting that most of the chemometric analysis is done with partial least squares. According to the authors' results, this may not yield the best results as ridge regression in many instances. While implementation of this technique is time-consuming because of the need to model using six different methods and then do computations to compare them, the results may be fruitful if one has the time and resources to go through the extra work.

Citation 3: The Effect of Maternity Leave on Maternal Depression

Chatterji, P., & Markowitz, S. (2005). Does the length of maternity leave affect maternal health? *Southern Economic Journal, 72(1),* 16–41.

Critical Summary

The authors of this paper attempted to determine whether the length of maternity leave had a statistically significant effect on maternal depression. Data from a 1988 U.S. federal study was used that purposely oversampled black and disadvantaged mothers, and low birth-weight babies. About 40 different variables were combined into a single *X* vector in a rather artibrary way. These included the mother's age, education, income, and ethnicity, as well as many others. The other independent variable was the number of weeks since birth before a mother returned to work. The dependent variable was a measure of maternal health which was a measure of depressive symptoms and a dummy variable representing whether or not the mother had sought health services three or more times in the first six months after childbirth. Some basic questions were asked to determine depression, which was then given a numerical value, which was then combined with the dummy variable. Finally, ordinary least squares was used on the model

$$H = b_0 + b_1 E + b_2 X + u + e$$

where *H* represented the health of the mother, *E* was the number of weeks after birth before a mother returned to work, and the *X* vector was a combination of some 40 variables described above. From this, the authors concluded that maternal health is positively correlated to lengthened time at home after birth before returning to work.

Critical Analysis

The authors did many complex statistical tests to attempt to validate their claims. They carefully described their work and data. However, ordinary least squares is not meant to be a method to determine causality between variables. It simply determines a relationship between the variables. The relationship may mean that *X* and *E* cause *H*, or it may mean that some other factor or factors cause all of the variables. Further weaknesses in the method are the use of data that purposely oversampled black, poor mothers, and low birth-weight babies. Therefore, any results from this study cannot be generalized to the general population of working mothers since it represents a special group. While this study could be of some benefit, the results should be used with great caution because of the arbitrary way in which variables were combined, the specific data that was used, and the reckless use of statistical tests to argue causality between variables.

Citation 4: Event Predictions

Eye, A., Brandtst, J., & Rovine, J. (1993). Models for prediction analysis. *Journal of Mathematical Sociology, 18(1),* 65–80.

Critical Summary

This paper discusses two alternative ways to predict events. The first is a method described by Hildebrand, Laing, and Rosenthal (1977), which describes a way to predict future states of an independent variable given bivariate or multivariate data. The second is a log-linear modeling method described by the authors. The first method consists of specifying a set of event predictions, grouping them into events that confirm the predictions or disconfirm the predictions, calculating expected frequencies assuming independence between predictors and criteria, and comparing observed errors with expected ones. The authors suggest a method of log-linear modeling using special design matrices. In general, they suggest the model

$$\log(F) = X\lambda + \varepsilon,$$

where F is a vector of frequencies, X is the indicator matrix, λ is the unknown parameter vector, and ε represents the random errors.

The authors criticize Hildebrand, et al. (1977) based mostly on their required assumptions of independence between the predictors and criteria as a baseline for evaluation of model success, and based on the way hypotheses are formulated using their approach. However, the authors admit several drawbacks to their own approach as well.

Critical Analysis

The authors clearly describe both methods. However, most of the support for the authors' method from is derived from other papers written by the same authors. The arguments in this paper give one the feeling of a group of authors arguing for their method against an established, classical one. Every statistical method has advantages and drawbacks, and these two would appear to fit that model. In certain instances, the log-linear approach may well be superior.

References:

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction Analysis of Cross-Classifications,* New York: Wiley.

Citation 5: Optimal Prediction for Least Squares with Infinitely Many Parameters

Goldenshluger, A., & Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis, 84(1),* 40–60.

Critical Summary

Stochastic multivariate linear regression models are considered in this paper, particularly when the number of parameters is large or approaching infinity. The authors develop a minimax modeling method that is superior to least squares in this case when certain other assumptions are met. The case where the random variables are independent and identically distributed are considered first, and then the case of correlated regressors are considered later in the paper. When the assumptions are met, the minimax prediction methods that produce optimal estimators are found via properly weighted least squares, with the weights defined by a filter developed by another author. The authors also point out that the optimal solution is based on certain a priori information. When the random errors are Gaussian, the method is asymptotically minimax over ellipsoids in L2, where the L2 is the least squares norm. The case of dynamic linear modeling is also briefly considered.

Critical Analysis

The results from this paper are important when doing transfer function modeling, a special case of multivariate modeling where the number of regressors is unusually large. The use for modeling general multivariate data may therefore be limited. The authors did a thorough job stating their assumptions, conclusions, and proving their results. As always, it is extremely important to check that all assumptions are met before using any statistical method to model real data.

Depending on the nature of future results concerning data dropping out of multivariate least squares, this paper could be of future interest. If r data points are affected for a multivariate model with r degrees of freedom as suspected, then as the number of parameters approaches infinity, so would the number of data points affected by data dropping out of least squares predictions.

Citation 6: Monitoring Changes in Linear Models

Horváth, L., Husková, M., Kokoszka, P., & Steinebach, J. (2004). Monitoring changes in linear models. *Journal of Statistical Planning & Inference, 126(1),* 225–251.

Critical Summary

When data is being collected continuously it is often necessary to monitor the model developed to fit the data to make sure that it still applies. To this end, monitoring schemes have been developed. Some monitoring schemes are slow but have a low probability of false alarm, while some are quick and have a higher probability of false alarm. The authors note the importance of assessing the particular application and the needs of monitoring before choosing a monitoring scheme. These authors developed four new monitoring methods using hypothesis testing and weighted least squares. The weighted residuals are calculated with a recursive method. The authors designed these methods to calculate changes fairly quickly after a training period, and to have a fairly low probability of false alarm. Several new lemmas and theorems are stated and proven, and the methods are then tested using a simulation. Results from the simulation showed that the authors' methods for monitoring changes in data worked as designed.

Critical Analysis

The four methods developed by the authors seem to be an improvement on older methods that had higher probabilities of false alarm. The fairly low probability of false alarm with these methods may imply that follow-up methods are not always needed before action is taken to change a model. The only caution is that users must be sure that all assumptions are met before employing any particular monitoring method.

Citation 7: A Weakly Consistent, Lower Variance Estimator than OLS

Jia, X., Rao, B.. & Zhang, H. (2003). On weak consistency in linear models with equi-correlated random errors. *Statistics, 37(6),* 463–473.

Critical Summary

A new estimator is proposed in this paper that is weakly consistent (has a limit that approaches the true parameters), and has lower variance than ordinary least squares estimators under certain conditions. Examples are shown that compare both the ordinary least squares estimator to the new estimator proposed by the authors. The estimator proposed is called the "centered" estimator for the unknown parameter vector β.

Critical Analysis

The authors propose that their estimator has lower variance than the ordinary least squares estimator under some "weak" conditions. However, those conditions will not be met in many problems. The new estimator is only weakly consistent if and only if $\lim_{n\to\infty} \lambda_{\max}\left[X_{(n)}^T X_{(n)} - n\overline{X}_{(n)}\overline{X}_{(n)}^T\right]^{-1} = 0$, where $\lambda_{\max}[A]$ is the maximum eigenvalue of the matrix *A.* A stronger requirement is imposed for the new estimator to be consistent. For consistency it is necessary that $\lim_{n\to\infty}\left(X_{(n)}^T X_{(n)}\right)^{-1} = 0$. The new estimator will beat least squares in terms of variance for problems where the conditions are met. However, prediction using those estimators was not dealt with in the paper.

Citation 8: Pathways to PTSD

Kaplow, J., Dodge, K., Amaya-Jackson, L., & Saxe, G. (2005). Pathways to PTSD, Part II: Sexually Abused Children. *American Journal of Psychiatry. 162(7),* 1305–1310.

Critical Summary

Nested ordinary least squares multiple regression was used to develop what the authors believe to be the first actual model defining pathways to post traumatic stress disorder (PTSD) for sexually abused children. One hundred fifty six children from North Carolina who had confirmed, probable, or suspicious sexual abuse were interviewed and assessed for risk factors known to be associated with PTSD. These include dissociation, age of onset of abuse, life stress prior to the abuse, avoidence behavior, and anxiety responses. Pathways and their respective probabilities of leading to PTSD later in life were developed. For example, children with high life stress before the abuse who exhibited anxiety responses during the interview were deemed to have probability (0.33)(0.22) ≈ 7.3% probability of developing PTSD later in life.

Critical Analysis

The model is interesting, and could potentially be useful in attempting to identify and treat children who are predisposed to PTSD before the symptoms actually develop. However, it should be noted that the authors did not clearly describe how exactly least squares was used, and it was main statistical tool used in the analysis. The R^2 value of 0.57, while not very low, was also not overwhelmingly high, and the model therefore needs to be used with some caution. The authors also noted several weaknesses in their analysis, including the fact that the findings may not generalize to all abused children because of the particular population of children used in the study. It was also noted that some of the findings may have been a response of subjects to the interview itself and not to the child's ability to cope with the underlying abuse.

Citation 9: Modeling approaches to risk adjustment in skewed outcomes

Manning, W., Basu, A., & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes. *Journal of Health Economics, 24(3),* 465–488.

Critical Summary

This paper dealt with the analysis of health care data, specifically health care costs and their responses to insurance, treatment or patient characteristics. The authors noted that ordinary least squares is sensitive to skewed data, and yet it is often used to analyze such data without dealing with the problem. This particular distribution was chosen partly because it includes some of the standard alternatives as special cases of the distribution. These include ordinary least squares with normal error, ordinary least squares for the log-normal, the standard Gamma and exponential with a log link, and the Weibull distribution. All of these distributions have been found to be helpful in dealing with health care data. This paper suggests using a generalized Gamma distribution with maximum likelihood estimation as an alternative to ordinary least squares when the data is skewed. Data that follows a Gamma distribution apparently occurs often in healthcare data analysis.

The authors tested their procedure using a Monte Carlo simulation and generating data in various ways. They found that their procedure was robust in dealing with skewed data.

Critical Analysis

The results are interesting, and are certainly better than ordinary least squares when data is skewed and the data to be analyzed fits the Gamma distribution. One must always be cautious about simulated data, because noise components are generally programmed in to fit nice conditions, and the kinds of problems that one sees with real measured data are often not seen in simulated data. Therefore, this modeling method should also be tested using real measured data from the health care field. The authors carefully documented their own cautions to the method as well, including making sure that assumptions are met and that every model needs to be examined for adequacy before being used in any particular case.

Citation 10: Selection of Variables for 1 Dimensional Regression Models

Olive, D., & Hawkins, D. (2005). Variable Selection for I D Regression Models. *Technometrics, 47(1),* 43–52.

Critical Summary

Fitting a model of *j* parameters from a larger group of *p* candidates for 1 dimensional models is a common problem in regression analysis. The authors note that when a large number of subgroups of parameters are being considered, the amount of computation involved is huge. This is often such a large problem that testing every candidate is not feasible. The paper suggests a graphical approach to the problem. Various plots are made, and these plots lead to a much smaller number of candidates to consider computationally. This approach is a much more efficient means of determining the optimal, or near-optimal candidates from among a large group of potential answers.

Critical Analysis

The approach is both innovative and simple. It should be considered, however, that the graphical method requires some interpretation from the statistician. It also is not guaranteed to yield the absolute optimal solution. However, it will yield either an optimal, or near-optimal solution. Therefore, when time and sheer amount of computation is a problem, this algorithm is a nifty solution.

Citation 11: Modeling Model Uncertainty

Onatski, A., & Williams, N. (2003). Modeling model uncertainty. *Journal of the European Economic Association, 1(5),* 1087–1122.

Critical Summary

The uncertainty in economic policy models can cause too much aggression or too much caution in making economic policy decisions. The authors of this paper modeled the uncertainty in economic policy models. They did this with Bayesian and Minimax techniques, and both parametric and non-parametric modeling. The minimax norm is appropriate in economic policy modeling because concern generally revolves around the "worst-case" scenario. The authors found that so-called robust models of errors performed dismally, and suggested that the reason for this is that these robust modeling techniques were not designed for this particular problem. In other words, statistical techniques may have been misused by some economists. The results they obtained suggested that the aggressiveness recently found in economic policy decisions is likely to be caused by an overemphasis of the uncertainty dynamics at low frequencies.

Critical Analysis

This fascinating paper was extremely thorough, and gives a clear introduction to the consequences of over-aggressive or over-cautious behavior in economic models. The authors point out that they made an oversimplifying assumption that once a model was in place that it would not be changed. They suggested that their results be viewed in light of this assumption, which is certainly not realistic in any real economy.

Citation 12: Local influence in multilevel regression

Shi, L., & Ojeda, M. (2004). Local influence in multilevel regression for growth curves. *Journal Of Multivariate Analysis, 91(2),* 282–304.

Critical Summary

Influence analysis is the study of how changes in a particular data point or simultaneous changes in a subset of the data changes linear regression models. This paper focused on approximating the influence of changing data points for mixed linear models. Several different assumptions and methods were considered. Local influence in multilevel regression was considered using a full iterative algorithm and using a one-step approximation. The authors found that the one-step approximation worked surprisingly well for detecting which data would be influential on a model. The authors thus recommended that the one-step approximation method be applied in real problems.

Critical Analysis

While the paper seemed overly long for what the author was trying to say, and the mathematics was tedious to read, the whole area of influence analysis may help to uncover the phenomenon of data dropping out of least squares predictions. I plan to research this area of study further for my dissertation.

Citation 13: Interpreting the Relationship between Staffing and Worker Injuries

Trinkoff, A., Johantgen, M., Muntaner, C., & Le, R. (2005). Staffing and Worker Injury in Nursing Homes. *American Journal of Public Health, 95(7),* 1220–1225.

Critical Summary

This paper was a study that attempted to make a link between staffing levels at nursing homes and professional caregiver injury levels. The authors hypothesized that they would find a statistically significant relationship between the variables. In order to test their hypothesis, the authors collected data on staffing levels and injury claims from Ohio, West Virginia, and Maryland, The 2000 Medicare Online Survey, Certification and Reporting (OSCAR) database was used as the data source. The staffing level variable was operationalized with some combination of the number of beds in each home, services provided by the homes, nursing and other personnel staffing, ownership of the homes, and resident acuity. The authors did not make it clear how these factors were combined or used to determine what "staffing level" meant, nor did they explain how any of these variables were connected with what was meant by staffing levels. First Report of Injury (FROI) databases were used to collect worker injury data in Ohio and West Virginia, while Workman's Compensation data was collected in Maryland since the authors felt that each of these sources had strengths and weaknesses compared to true injury figures.

Once the data was collected, an ordinary linear regression was run on the data, and the results were analyzed. It was found that the R^2 value was .25, which the authors reported as a high correlation between the data. The authors also reported a skew at one end of the data, so they also performed a regression on the independent variable versus the log of the worker injury data. They reported a consistently high correlation in this case. The authors concluded that there is a statistically significant linear relationship between staffing levels and worker injuries.

Critical Analysis

The purpose of reading this paper was to analyze a recent application of least squares in the research literature. While the authors performed a linear regression analysis and computed R^2 for the data, they did not look at residuals, they performed no formal hypothesis test, and did not do a goodness of fit test to establish statistical significance. The authors reported a high correlation between the data, but an R^2 value of .25 is hardly a strong association in the eyes of most statisticians. In fact, it is so low that most statisticians would conclude a weak association. If there was a persistent suspicion that a strong relationship did indeed exist then other tests such as a goodness of fit test and a formal hypothesis test would have been minimum requirements to reach the conclusions that the authors claimed.

This paper is an illustration of the misuse of statistics. In this case it was the misuse of linear regression analysis, but there were many problems with the analysis, including the lack of description of the operationalization of the variables, the lack of any data in the paper other than some of the figures resulting from the analysis, and the fact that the authors deleted some data, changed some data, and otherwise manipulated data in questionable ways before analyzing it. This paper reinforces the idea that claims of relationships between variables must be carefully examined before they are accepted. In addition, even high correlation does not imply causation. The authors of this paper claim causation without giving any basis for the claim other than what is arguably a weak relationship between the variables.

Citation 14: Simultaneous variable selection

Turlach, B., Venables, W., & Wright, S. (2005). Simultaneous Variable Selection. *Technometrics, 47(3),* 349–363.

Critical Summary

Linear regression problems that arise from real measured data often contain highly correlated predictors. Ordinary least squares is not an appropriate technique in this case. To this end, many algorithms have been produced to deal with these ill-conditioned problems. The authors of this paper propose a new algorithm for solving this sort of problem. While most of the previous algorithms produced by others were based on a Bayesian approach, the new algorithm proposed in this paper does not. This is a huge advantage, because selecting the parameters necessary in order to get convergence with Bayesian-based algorithms can be nontrivial. The method proposed is a way to select groups of regression estimates from among a large group of candidates based on the residual sum of squares while constraining the estimates to lie within a polyhedral region.

Critical Analysis

This algorithm appears to be an extremely useful one, especially in light of the non-Bayesian approach. It is most appropriate for extremely large regression problems where predictors are correlated. The method is an interesting mixture of a classic least squares fitting problem with constraints, and a quadratic programming problem. The authors did a beautiful job of illustrating their algorithm and its potential applications with several examples that would be useful in engineering and business problems.

Citation 15: Estimating the Conditional Variance of Y, given X

Wilcox, R. (2005). Estimating the conditional variance of Y, given X, in a simple regression model. *Journal of Applied Statistics, 32(5),* 495–502.

Critical Summary

The condition variance of Y, given X, is of importance in many practical situations. Some of these include weighted least squares, quality control problems, calibration problems, and when trying to understand the relationship between X and Y. Estimators of the conditional variance proposed by others have been found to be inadequate for various reasons. Key reasons are the requirement of a large sample size, or simple poor performance of the estimators. Wilcox attempted to determine whether a better estimate of the conditional variance could be obtained.

Five different potential estimators for the conditional variance were developed, and then all five were compared using a simulation. Data with different distributions were used in the simulation, along with different sample sizes. The results suggested that the best estimator is a complex function of the underlying distribution function of the data. However, it was also found that more than one estimator should be retained, because one particular estimator was better for heavy-tailed distribution functions, and another estimator was optimal for more "centered" distributions. The optimal way of calculating the estimate of conditional variance would appear to be highly related to the shape of the distribution function.

Critical Analysis

The author's method of using simulations to compare several different estimators of the conditional variance was an innovative way to approach the problem. The problem is complex enough that theoretical analysis would have been prohibitive, and this is probably the reason that such an important problem has not yet been successfully tackled analytically. This paper is a “must-read” for anyone that needs to have an adequate estimate of conditional variance.

DEPTH DEMONSTRATION

Data Drops Out when Making Predictions using Least Squares Models

Introduction

In the previous section, the general theory of least squares modeling and prediction was described. The absolute value and minimax norms were also briefly described and analyzed. In this section, a new finding regarding least squares is developed. This involves the fact that certain conditions cause data points to drop out of predictions for particular $\hat{y}_i$ values. This finding is important because it is usually assumed that all data is being used in such predictions, and results can be skewed if data points are in fact dropping out of the prediction equations, especially for small values of *n*. It seems clear that the loss of a data point when predicting *y*-values in a linear model of the form $y = \beta_0 + \beta_1 x + \varepsilon$ is a loss of information, and such a prediction may be suboptimal in comparison to some other prediction technique that uses all the *y*-data points in its calculation. Further, this finding extends beyond straight line models to other models that are linear in the unknown coefficients.

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

The Phenomenon of Data Dropping Out in Least Squares Predictions

Linear modeling using the L2 (least squares) norm is a mature field of statistics. The mathematics associated with it is elegant, and the technique lends itself to many applications using closed form solutions that are efficient and convenient. However, a phenomenon that occurs only under certain circumstances may have escaped notice until now: some data is not being used when making certain predictions. This happens in a predictable way, and occurs for a range of models that are linear in the unknown coefficients, and in multivariate linear modeling as well.

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

Example 1.

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

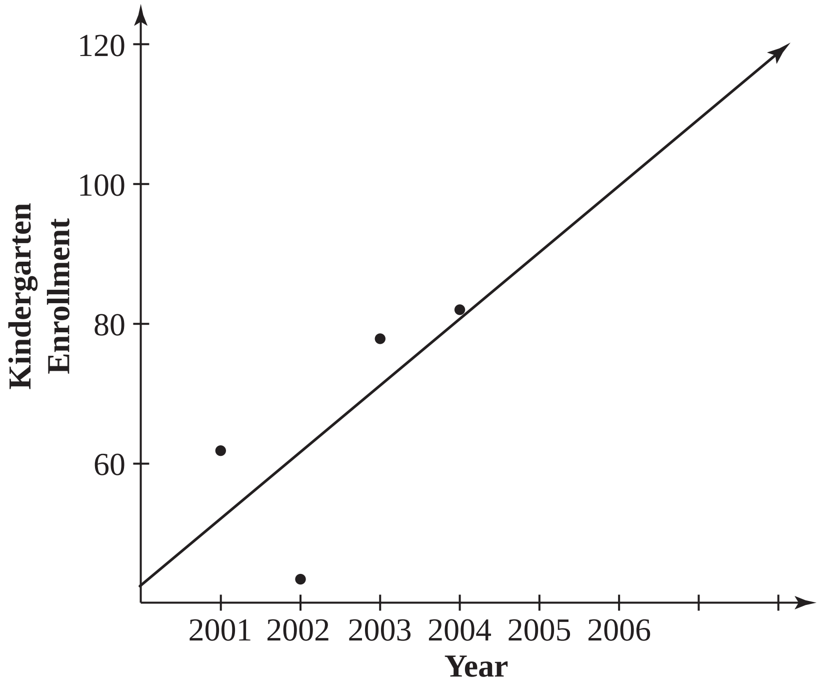

Figure 1. Allen Elementary School Kindergarten Enrollment from 2001 to 2004

Suppose the school wishes to estimate the enrollment for 2005. Assume that the enrollment behaves according to a linear function $Y = X\beta + \varepsilon$, with $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$, where the first row corresponds to 2001, the second row to 2002, etc., $Y = \begin{bmatrix} 62 \\ 43 \\ 78 \\ 82 \end{bmatrix}$, and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$.

When the least squares solution is computed, the solution is $\hat{y}_i = X\hat{\beta} = X(X'X)^{-1}X'Y = 9.5x_i + 42.5$. To estimate enrollment for 2005, $x_5 = 5$ is substituted into the model to obtain $\hat{y}_5 = 42.5 + 9.5(5) = 90$. Now, suppose the second y-value is changed so that $y_2 = 20$ instead of the original value of 43. This changes the regression line so that $\hat{y}_i = 31 + 11.8x_i$, but $\hat{y}_5 = 31 + 11.8(5) = 90$ as before. Similarly, the second y-value can be changed again so that $y_2 = 120$. Now the regression yields $\hat{y}_i = 81 + 1.8x_i$. Since the value of y_2 is so large, it seems reasonable to expect that the new estimate for 2005 enrollment would be much higher than before. Yet the computation for $\hat{y}_5$ is $\hat{y}_5 = 81 + 1.8(5) = 90$ just as before. This is the case even though the regression line itself has certainly shifted. In short, it appears that the value of y_2 has no effect at all on the estimate for the 2005 Kindergarten enrollment.

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

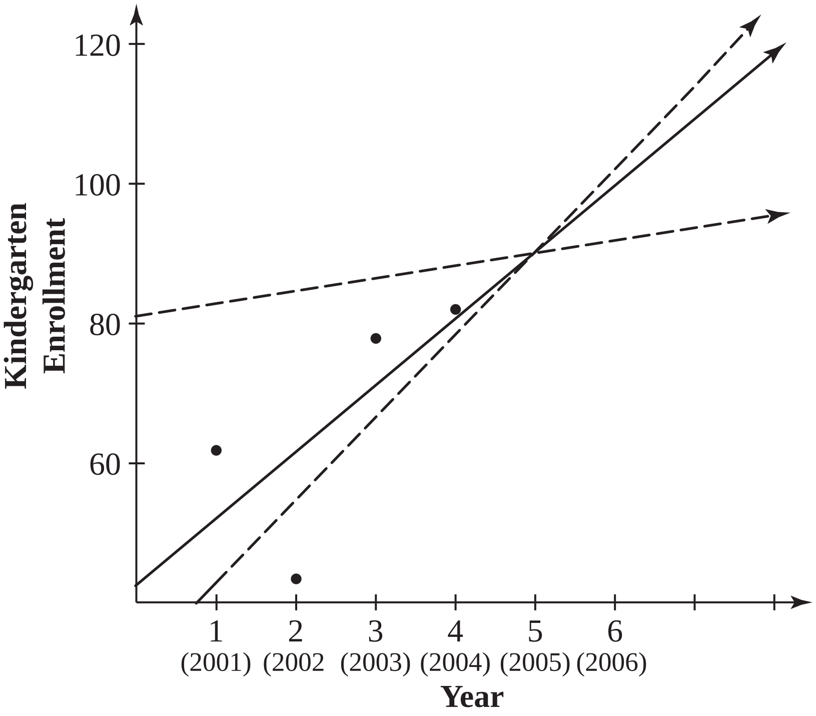

Figure 2. Allen Elementary School Kindergarten Enrollment Prediction Lines

It is easy to see in Figure 2 that the three lines intersect at (5, 90), which corresponds to the prediction that there will be 90 students in the 2005 Kindergarten class. In fact, it will later be shown that y_2 can be changed arbitrarily, and all of the regression lines will intersect at

(5, 90). Notice, however, that the lines only intersect at this one point, and that for all other values of *x* the predictions for enrollment will be different when y_2 is changed. This phenomenon occurs for other values of *n* and $\hat{y}$ as well, and it will be useful to derive a mathematical relationship between the data point that drops out and the predicted value for which it drops out.

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

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

Assume there is a linear model in *x* such that

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

or, in matrix form

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

It was shown in the Breadth section of this paper that the solution for this model is

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

To clarify, the matrix form of the solution for $\hat{\beta}$ is included below:

$$
\begin{aligned}
\hat{\beta} &= \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
&= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i{}^2 \end{pmatrix}^{-1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
&= \frac{1}{n\sum x_i{}^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i{}^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{n\sum x_i{}^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i{}^2 \sum y_i - \sum x_i y_i \sum x_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix} \qquad (1)
\end{aligned}
$$

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

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

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

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 $i = 1,\ 2, \cdots$ 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

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

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

The Case when n = 4

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

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

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 gives

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

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

Now the result for the simple linear model shown above is generalized for *n*, and the simplest case is a prediction of y_{n+1}. On the way to the general case, a brief look is taken at the case when $n = 7$ as a help in finding the pattern for general *n*.

If the *x*-values are 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

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

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

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

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

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

Therefore,

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

and

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

(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

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

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 3 below.

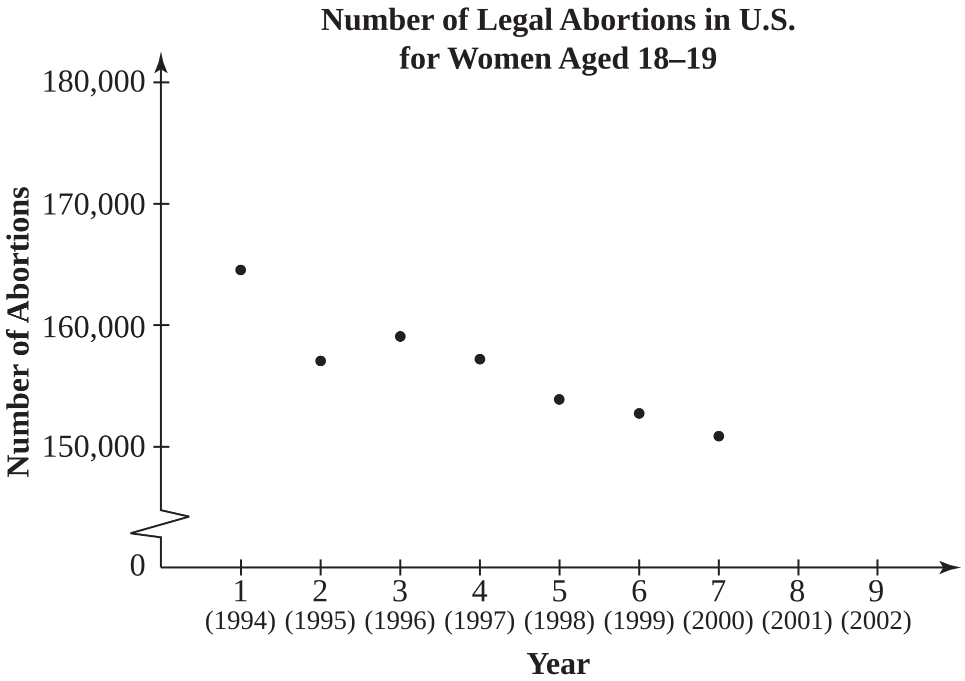

Figure 3. Legal Abortions in the U.S. from 1994 to 2000 for Women Aged 18 and 19.

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 4 below, using several different values for y_3 and showing that they all intersect at $(8, 148,457)$.

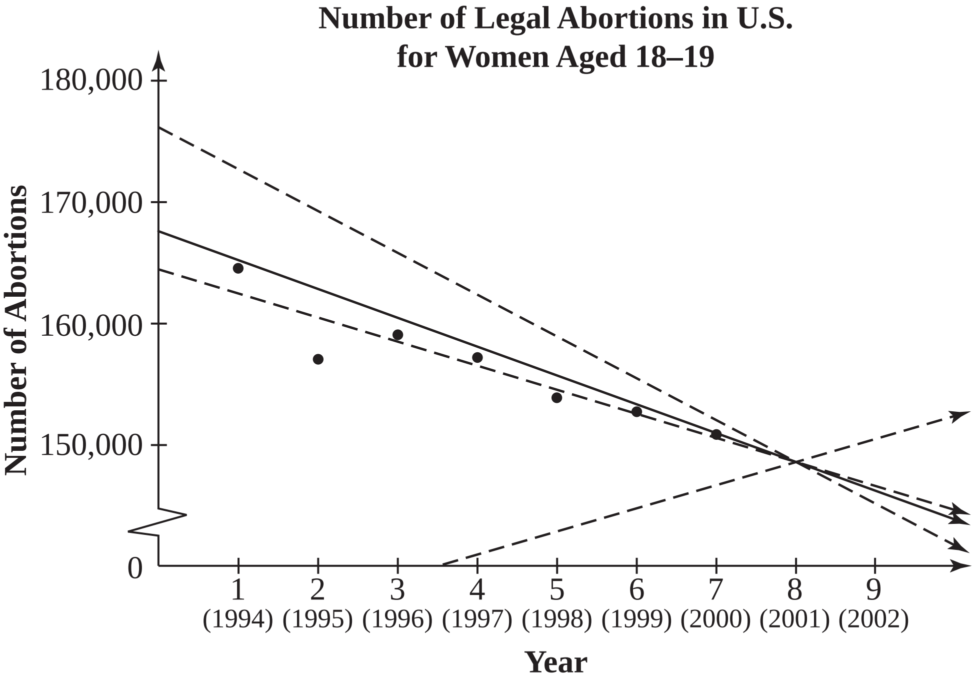

Figure 4. Regression Lines for U.S. Abortions with Various Values of y_3

It is clear that the predictor $\hat{y}_8$, represented 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 drops out can be found when the coefficient of y_i is 0. So,

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

Therefore,

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

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

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

When $n = 4 + 3k,\ k = 0,\ 1,\ 2,\ 3,\cdots$ then y_{k+2} drops out of the prediction for $\hat{y}_{n+1}$. In particular, substituting $n = 4 + 3k$ in the above equation for $\hat{y}_{n+1}$ and simplifying shows the result that the $(k + 2)$th y-data point drops out when estimating $\hat{y}_{n+1}$, as stated.

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

The simplified case where the x_i's are equally spaced helped to determine which integer values of *k* cause y_k to drop out 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. Now, assuming that for every value of *d,* there exists a point $\hat{y}_p$ 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 fact is illustrated in Examples 1 and 2, where the independence of $\hat{y}_5$ from y_2, and of $\hat{y}_8$ from y_3, does not depend on the other y_i values. Therefore, in order to derive the relationship between x_p and x_d, the values of the *y*-vector can be varied at will. Visually, the desired result is the point x_p at which all the various regression lines corresponding to different values of y_d intersect when the other *y*-values are held constant. For this purpose any two lines will suffice, and thus *y*-values can be chosen for maximal convenience. Thus, let all the *y*-values other than y_d equal 0, and let y_d be either 0 or 1. (Thus *Y* is either the zero vector or the indicator function on *Y* at *d*.) Therefore, in order to derive the relationship, the values of the *y*-vector can be varied at will. Therefore let the *Y* vector consist of zeros except let $y_d = 1$ (the indicator function on *Y* at *d*). Then

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_d & \cdots & x_n \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_d \end{bmatrix},$$

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

$$\hat{\beta}_1 = \frac{x_d \cdot n - \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \text{ and } \hat{\beta}_0 = \frac{\sum x_i^2 - x_d \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

Now since all the components of *Y* are 0 except for the value corresponding to x_d, the *y*-value corresponding to x_p is 0. This yields the equation

$$0 = \hat{\beta}_1\left(x_p\right) + \hat{\beta}_0, \text{ or } x_p = -\frac{\hat{\beta}_0}{\hat{\beta}_1}.$$

Therefore,

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

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

Theorem 1. Given $y = \beta_0 + x\beta_1 + \varepsilon,\ Y\ (n \times 1),\ X\ (n \times 2)$, specified. Let $\hat{y}_p$ (p real), be a prediction based upon $\hat{\beta} = (X'X)^{-1} X'Y$. Then there exists an integer value d, $1 \le d \le n$, and $d \ne p$, such that $\hat{y}_p$ does not depend on y_d, and the relationship between p and 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 $$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}, \qquad (8)$$

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

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

Note that with some simplification,

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

$$x_p = \frac{\sum (\bar{x} - x_i)^2}{n(\bar{x} - x_d)} + \bar{x} \qquad (10)$$

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

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

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

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

Expanding gives

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

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

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

If y_d drops out of 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

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

Example 3.

The finance manager of a major fast food chain suspects that the gradually increasing number of tacos sold can be usefully modeled by a linear function. She has decided to compile data on the number of tacos sold for several years in order to estimate the number of tacos likely to be sold over the next several years if the pattern continues. She knows she can compile data for the last 11 years, except that the year 3 data was irretrievably lost due to a computer crash several years ago. Therefore, she can compile data for years 1, 2, 4, 5, 6, 7, 8, 9, 10, and 11, where year 11 corresponds to last year. The finance manager has noted that it will take a considerable amount of effort to retrieve the data for the number of tacos sold each year, due to the way the data was originally recorded. Therefore, she wants to make sure that all the data she collects will be used in her estimates for years 12, 13, 14, and 15. Will any of the predicted values the manager wants be independent of any of the data values?

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

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

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

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

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

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

Therefore, the fourth data value (when $x = 5$) will not affect the predicted value of $\hat{y}_{14}$. This might be a reason for the finance manager not to bother collating the data for that year. Admittedly, given that the fourth data value is still apparently relevant to the predictions for years 12, 13, and 15, there might still seem to be a reason to go ahead and compile the fifth data value. However, we see a hint emerging that the fifth data value may not have much effect on the predictions for years 12, 13, and 15 after all. In fact, initial findings indicate that there are "degrees of relevance" for various data points that would strengthen the case for omitting the fifth data point. The analysis of this issue is beyond the scope of this paper and will be the topic of a future paper.

A Physical Application

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

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

Suppose there are *n* *x*-values where the *x*-values are allowed to be arbitrary, 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 5 below.

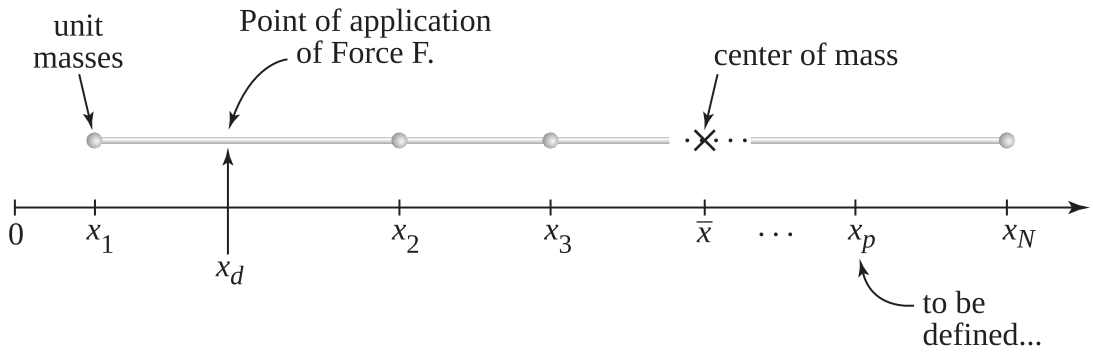

Figure 5. Massless Rods Joined to Form a Single Rigid Body

Press sideways against this linear body at some arbitrary point other than the center of mass. The body will now begin to translate and to rotate. There will, however be a point 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 6 below).

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 that was previously called x_d. The point that remains stationary is the analogue of the *x*-coordinate corresponding to the point $\hat{y}_p$, the point that was previously called x_p. The relationship between x_d and x_p can be derived by the following equations, derived from simple physical relationships.

$F = ma = na_{\overline{x}}$ $\quad$ ($a_{\overline{x}}$ = acceleration of the center of mass, *n* is the number of points)

$\tau = I\alpha$, where

τ = torque and $I\alpha$ = moment of inertia multiplied by the angular acceleration.

Then

$$F(\overline{x} - x_d) = \sum(\overline{x} - x_i)^2 \times \alpha.$$

The displacement of a point along the rod, a distance *x*, is the displacement of the center of mass plus the displacement due to the rotation about the center of mass (see Figure 6), so

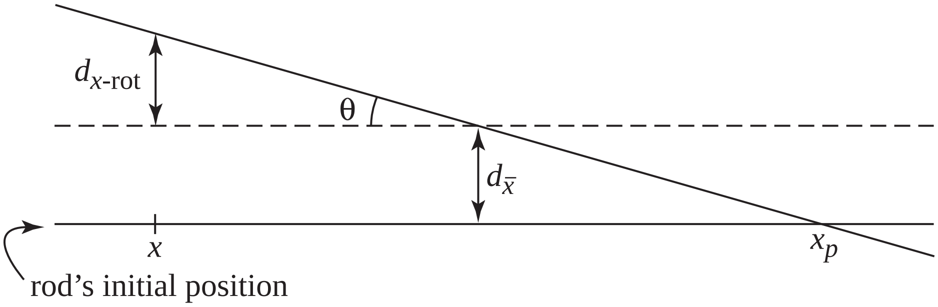

Figure 6. Pivoting Massless Rod

$$d_x = d_{x-\mathrm{rot}} + d_{\bar{x}}.$$

Differentiating twice, and using

$$d_{x-\mathrm{rot}} = (\bar{x} - x)\theta, \text{ yields}$$

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

Now, the point where $a_x = 0$ is the point x_p, which is the point that doesn’t move when force is applied at x_d. At that point,

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

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

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

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

which, after some simplification becomes

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

where $\bar{x}$ is the statistical notation for the physical quantity of the center of mass, x_d is the *x*-value that drops out when predicting $\hat{y}_p$, and x_p is the *x*-value corresponding to $\hat{y}_p$.

Future Work

Preliminary results show that the result regarding data points dropping out of predictions holds for models of the form $Y = X\beta + \varepsilon$, which are linear in the unknown coefficients but polynomial equations in *X*. Exactly *k* values of $\hat{y}_i$ will be independent of some data point for any linear model that is polynomial in *X* with power *k*. This will be the subject of a future paper. Preliminary work has also shown that the result is extendable to multivariate linear models as well. Future work will extend the results for these cases and compare the estimates of $\hat{y}_p$ for the least squares norm in this special case when data drops out to other methods for estimating $\hat{y}_p$.

Additionally, future work will explore the "degrees of relevance' as introduced at the end of Example 3. Finally, if we measure distance from $\bar{x}$ in one direction as positive and in the

other direction as negative, then the relationship between x_p (an *x*-data point at which we want to predict y_p) and x_d (the *x*-data point at which y_d is irrelevant to that prediction) has the general shape of the function $x_p = -\frac{1}{x_d}$. That is, a data point slightly to the left of $\bar{x}$ will be irrelevant to a prediction far to the right of $\bar{x}$, and a data point far to the left of $\bar{x}$ will be irrelevant for a prediction slightly to the right of $\bar{x}$. These topics will also be covered in a future paper.

References

California Department of Education, Educational Demographics Unit. District and School Enrollment by Grade, San Jose Unified School District, 2001-2004 [Electronic data retrieval]. Retrieved on September 9, 2005, from http://www.cde.ca.gov/ds/sd/cb.

The Alan Guttmacher Institute. (2004, February 19). U.S. Teenage Pregnancy Statistics With Comparative Statistics For Women Aged 20-24: Notes on Teenage Pregnancy Statistics [Electronic version], Retrieved on September 1, 2005, from http://www.agi-usa.org/pubs/teen_stats.html.

APPLICATION DEMONSTRATION

APPLICATION TABLE OF CONTENTS

APPLICATION DEMONSTRATION

Introduction

In the previous two sections of this KAM, the basic theory of least squares modeling and estimation was developed, and a phenomenon describing how some data drops out of least squares predictions was described and developed. In this section a look is taken at some real data using least squares modeling and prediction techniques. This data is derived from the same data used in one of the examples in the depth section of this KAM.

The data used in this section came from the same source as one of the examples used in the depth section of this paper. Namely, a company name Decision Insite Corporation collects school enrollment data for several school districts in the Los Angeles area, compiles the data, and among other things, prepares enrollment projections for future years for each district and for each school within the district. Of special interest are Kindergarten projections, since these must be made with some combination of enrollment trends, births in the district, and other demographic data. It is on the trend part of the Kindergarten enrollment projections that this paper will focus.

Problem Description

Data was provided by client school districts to Decision Insite Corporation each October. To date, Decision Insite Corporation has only been provided with enrollment data for the Oak Park School District in California from 2001 through 2004. Fairly accurate estimates were also available for 2005 Kindergarten enrollment. There are five schools in the Oak Park School District, and enrollment figures were available at the school level. The district's Kindergarten enrollment has been showing a sharply decreasing trend in most of the district's five schools. The problem was to model the data and provide reasonable projections for 2006 enrollment so that the school district could make decisions about hiring and purchases. The district notified Decision Insite Corporation that they preferred to have underestimates of Kindergarten enrollment rather than overestimates, presumably since they can always buy a few more books and supplies, but overbuying is a problem. This is also the case since the classrooms are not currently running at full capacity, and so a few extra children do not present a space problem.

The Modeling Process

The small number of data points presented an immediate problem for modeling and prediction. In fact, past predictions were made using simple percentage changes from only one previous year, and no modeling at all has ever been done with the enrollment data. This is not surprising since the model would have been barely oversampled last year. Since exactly four data points of real data were available, there was a second problem of the second data point dropping out when estimating enrollment figures for 2006. It is for this reason that the estimates for 2005 were also included in the model. This eliminated the problem of the second data point dropping out of the model, and insured that all data would be used. A further assumption was also necessary. The enrollment figures needed to be assumed to follow some kind of trend that could be modeled with some kind of linear or linearizable equation.

The data was modeled using a simple linear model of the form $y_i = \beta_0 + \beta_1 x_i$, and also as a population model of the form $y_i = \beta_0 e^{\beta_1 x_i}$, which is a well known population model. It was hoped that the data would fit either a linear trend or a decreasing exponential trend. The decreasing exponential model has the additional desired effect that it is asymptotic to zero, since enrollment projections can never be less than zero. The exponential model is clearly not linear in the unknown coefficients. However, it is linearizable. In particular, take the natural logarithm of both sides of the equation. This yields

$$\log y_i = \log\left(\beta_0 e^{\beta_1 x_i}\right)$$

This can then be further simplified, giving

$$\log y_i = \log \beta_0 + \log\left(e^{\beta_1 x_i}\right)$$
$$\log y_i = \log \beta_0 + \beta_1 x_i$$

Finally, two substitutions of the form $\beta_0^* = \log \beta_0$, and $y_i^* = \log y_i$, giving the linear model

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

It is this final linear model that can be fitted using ordinary least squares. Note that there are no error components in either of the models since the enrollment data is assumed to be accurate and free of measurement error.

The Kindergarten enrollments for the Oak Park School District are from 2001 through 2005 are given in Table 1 below. Note that the 2005 enrollment figures are estimates.

Table 1
Red Oak School District Enrollment Data for 2001 through 2005

School	2001	2002	2003	2004	2005 (est.)
District Aggregate Data	224	232	199	169	169
RdOk	84	86	74	62	60
Brksd 1	50	47	34	31	28
Brksd 2	20	33	25	12	12
OkHs	52	54	48	55	60
ODist	18	12	18	9	8

Results

Both the simple linear model and the linearized exponential model were fitted using ordinary least squares using Mathematica as described in the breadth section of this paper. The resulting models were then used to obtain predictions for 2006 Kindergarten enrollment for the entire Oak Park School District, and for the individual schools within the district. Future years were also calculated using the decreasing exponential models for four of the five schools. Future years were not predicted for the Oak Hills School (OkHs) since the data did not appear to fit either a linear trend or a decreasing exponential trend. The data and residuals were examined for each school, the correlation coefficient calculated, and the results were examined.

The aggregate Kindergarten data for the district showed a correlation of over .92 for the straight line model, with a residual pattern that looked adequate. The estimated total Kindergarten enrollment for 2006 for the Oak Park School District using this model was 147 students, compared with Decision Insite Corporation's straight percentage prediction of 148 students. However, beyond 2006, the linear trend looks inadequate, and it appears that the enrollment would be seriously underestimated, This is not at all surprising given that least squares is not meant to predict what will happen far outside the realm of the collected *x*-data. When the exponential model was used, the correlation was just under .92, and the projected Kindergarten enrollment for 2006 was 151. While this prediction is higher than the linear prediction for 2006, the model seems much better for years past 2006 than the linear model, and aligns fairly closely with the predictions of Decision Insite Corporation out to 2007 or 2008. Past that point the model is probably underestimating the Kindergarten enrollment. It is interesting to note that Decision Insite's predictions have already accounted for demographics, while the model is only taking into account past enrollment figures. If the current trends continue, the exponential model seems like the best overall fit for the aggregate data. A compilation of results is shown in the Table 2 below.

Table 2
Comparison of Predictions for Different Models, Aggregate Kindergarten Enrollments, Red Oak District

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Actual	224	232	199	169						
Decision Insite's Estimates					169	148	140	135	135	131
Linear model	233	216	199	181	164	147	129	112	95	77
Exponential model	235	215	197	180	165	151	138	127	116	106

Each school within the Red Oak School District was also modeled individually. Decision Insite's method of prediction for each individual school was to take the previous year percentage of the overall district Kindergarten enrollment for each school and apply the same percentages to the new predicted value. All of the linear models provided what appeared to be reasonable estimates for 2006, but inadequate predictions for later years. Correlations were acceptable to very high. The exponential models seemed better for years after 2006 than the linear model, except in the case of Oak Hills School (Ok Hs), which did not appear to even have a decreasing trend, let alone an exponentially decreasing trend. One possible solution would be to look at estimates for the other schools in the district and then assign the difference between the aggregate enrollment estimates and the sum of the other schools to the Oak Hills School. The results are shown in Tables 3, 4, 5, 6, and 7 below. Any linear estimate that resulted in a prediction of fewer than 0 students was truncated to 0 students.

Table 3
Comparison of Kindergarten Enrollment Predictions for Different Models, Red Oak School

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	ρ
Actual	84	86	74	62							-
Decision Insite's Estimates					60	51	47	45	45	43	-
Linear model	88	80	73	66	60	52	44	37	30	23	.94
Exponential model	88	80	72	66	59	54	48	44	40	36	.94

Table 4

Comparison of Kindergarten Enrollment Predictions for Different Models, Bakerside 1

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	ρ
Actual	50	47	34	31							-
Decision Insite's Estimates					28	23	21	20	20	19	-
Linear model	50	44	38	32	26	20	14	8	2	0	.96
Exponential model	51	43	37	32	27	23	20	17	14	12	.97

Table 5

Comparison of Kindergarten Enrollment Predictions for Different Models, Bakerside 2

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	ρ
Actual	20	33	25	12							-
Decision Insite's Estimates					12	11	10	10	10	10	-
Linear model	28	24	20	17	13	9	6	2	0	0	.65
Exponential model	28	23	19	15	13	10	8	7	6	5	.57

Table 6

Comparison of Kindergarten Enrollment Predictions for Different Models, Oak Hills

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	ρ
Actual	52	54	48	55							-
Decision Insite's Estimates					60	56	55	54	54	52	-
Linear model	51	52	54	56	57	59	61	62	64	66	.61
Exponential model	47	46	45	43	42	41	40	38	37	36	.1

Table 7

Comparison of Kindergarten Enrollment Predictions for Different Models, ODistrict

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	ρ
Actual	18	12	18	9							-
Decision Insite's Estimates					8	7	7	7	7	7	-
Linear model	18	15	13	11	8	6	4	2	0	0	.76
Exponential model	18	15	12	10	8	7	6	5	4	3	.73

Potential Problems and Cautions

Unfortunately, the predictions calculated above need to be used only with extreme caution. As stated in the breadth section of this paper, ordinary least squares prediction is best used to predict data between or near the *x*-data used to create the model. For example, it is not meant to extrapolate far into the future. Therefore, the estimates for 2006 can be used with reasonable assurance, but the estimates beyond that should be used only very cautiously. Because of the nature of the models and the small number of data points, it is very likely that the enrollment figures are underestimated for years past 2007, especially so for predictions derived using the straight line models. These may be able to be adjusted upward using demographic data as Decision Insite has done in the past. Still, the use of all of the previous data to model future years is probably better than a simple percentage change based only on the previous year.

It must be stated again that the number of data points used for these models was extremely small. If only the actual figures from 2001 through 2004 had been used, the second data point would have dropped out of the predictions for 2005. Therefore, it was advantageous to use the 2005 estimates in the models, especially since the 2005 school year has already started. This added two data points of additional information when calculating the predicted values over what would have happened using four data points. Even so, such a small number of data points can cause unusually high correlation coefficients, even if the data is not highly correlated. The correlations also therefore need to be used cautiously.

Finally, it is highly unlikely that a continuous downward trend will be observed in any school district over a long period of time. The models therefore need to be rerun each year to check for model adequacy. If the data follows any kind of trend, then as the number of actual data points increases from year to year, the models should get better, and the predictions from those models more accurate.

Conclusion

The above analysis shows the best and worst of least squares predictions. Using several data points should be superior to using only the previous year enrollment figures for the purpose of making predictions for the following year if the data follow a trend. However, least squares is clearly not a good method for estimating Kindergarten enrollment far out into the future. However, enrollment figures for 2006 are probably fairly accurate and usable. While least squares prediction is a powerful technique, one must always be careful to note its limitations as well as its strengths.

References

California Department of Education, Educational Demographics Unit. District and School Enrollment by Grade, San Jose Unified School District, 2001-2004 [Electronic data retrieval]. Retrieved on September 9, 2005, from http://www.cde.ca.gov/ds/sd/cb.