NOT FOR PUBLICATION

IS SIMPLE LINEAR REGRESSION REALLY SIMPLE?

Patrick O. Darrow
Research Statistician

IMPERIAL HOLLY CORPORATION

RESEARCH AND DEVELOPMENT CENTER
COLORADO SPRINGS, COLORADO

AMERICAN SOCIETY OF SUGAR BEET TECHNOLOGISTS
27TH BIENNIAL MEETING
ANAHEIM, CALIFORNIA
MARCH 3-6, 1993

ABSTRACT

This paper is concerned with the appropriateness of simple linear regression models in which a response is predicted based on a linear functional relationship to a regressor variable. Assuming that a cause-and-effect relationship exists, and that prediction is the objective, the usual method is to assume that least squares estimation of the parameters is appropriate. It is also assumed that the measured response is subjected to variation, while the regressor variable is a fixed quantity, measured without error. More often than not, such assumptions are not appropriate and the simple linear regression becomes more complicated. This paper focuses on the assumptions required for the usual least squares, fixed regressor variable case. Alternative means of constructing the simple linear regression model is outlined. An example involving beet quality determination is used for illustration.

INTRODUCTION

Early in the 1900's, correlation analysis was commonly used to characterize the relationship between two variables, x and y. The advantage of such a simple analysis was that it treated both variables on equal footing. Unfortunately, correlation analysis measures the strength of a linear relationship between the two variables. If x and y were related nonlinearly, correlation analysis would yield a low correlation and the researcher would look to other means for relating the two variables. While correlation analysis is still common today, researchers tend to focus on regression, or the prediction of y based on values of x.

The reason for expressing y as a mathematical function of x depends on the objective of the study. Generally, three objectives are recognized:

(1) a functional relationship exists based on theory - the purpose of regression is to obtain the estimates necessary to characterize this relationship;
(2) a functional relationship exists and parameter estimates are known - the purpose of regression would then be to validate the theory;
(3) no specific functional relationship can be deduced from theory - the purpose of regression would be to assist in determining possible relationships, or to provide an empirical characterization of the data.

Objective (3) is the most common. It is particularly useful when direct measurement of y is not possible or feasible. The researcher strives to find a functional relationship that approximates reality.

The simplest form of regression is the simple linear regression involving two variables, x and y. It may be that the researcher, through correlation analysis, determined an adequate linear relationship between the two variables. Furthermore, the researcher desires to predict y as a linear function of x. As simple as this sounds, several considerations must be made in performing such a regression.

The purpose of this paper is to detail the simple linear regression model and assumptions necessary for application. When these assumptions are violated, the simple linear regression becomes more complicated and requires less contemporary methods for application.

SIMPLE LINEAR REGRESSION MODEL

The simple linear regression model (SLRM) is characterized by an equation of the following form,

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where Y_i is the response of interest (commonly referred to as the dependent variable), α is the parameter representing the Y-intercept, β (referred to as the slope) is the parameter associated with X_i (referred to as the independent variable), and ε_i is the experimental error associated with Y_i.

The usual method is to collect data on Y for selected values of X, and obtain, via least squares, estimates for the parameters α and β. Fitting a line by least squares is preferred over other methods since least squares estimation strives to minimize the sum of the squared prediction errors. That is, least squares allows us to choose estimates for α and β that provides the least amount of uncertainty associated with the estimation of these parameters. Thus, the least squares estimates for α and β are given by,

$$\hat{\beta} = \frac{\sum_i x_i y_i - \frac{1}{N} (\sum_i x_i)(\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{N} (\sum_i x_i)^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

where $\bar{x}$ is the mean x value, $\bar{y}$ is the mean y value, x_i and y_i are the corresponding observations, and N is the number of (x_i, y_i) pairs. If the goal is to predict y based on selected values for x, the following equation is derived:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i \, .$$

The advantages in using least squares estimation are mathematical, including simplicity of the calculation, optimality of the estimates, and quantifying the uncertainty of predictions. If the data follow a gaussian distribution, the mathematical elegance of least squares is enhanced. However, disadvantages do exist and are principally concerned with the nature and "pattern" of the data. If the data exhibit unusual patterns (i.e., non-normality) or include extreme values (outliers), least squares can be very misleading. If the regressor variable, x, is also measured with uncertainty, usual approaches must be modified. Gunter (1992a) discusses an example involving outliers and their effect on simple linear regression. Other examples are available as well (e.g., Mandel, 1964; Draper and Smith, 1981).

Statisticians have a well-defined arsenal of diagnostic procedures for assessing the applicability of the simple linear regression model (e.g., normality plots, outlier tests). In many cases, however, researchers fail to indicate what diagnostic procedures, if any, were used to validate the use of the simple linear regression model. In fact, the availability of "canned" software tends to lull the researcher into a false sense of security when developing parameter estimates. Most popular software fails to outline the advantages and disadvantages of applying regression techniques. Unless a statistician or appropriately trained personnel are consulted, the adoption of these equations may prove erroneous.

SIMPLE LINEAR REGRESSION: IF NOT LEAST SQUARES, WHAT?

The most important criterion for wanting to fit a straight line to data is the underlying system. Is the line supported by theory? Linear regression is a tool, not a replacement for underlying processes. Assuming that simple linear regression is deemed theoretically sufficient, diagnostic tools are available for assessing the utility of the line.

One of the easiest diagnostic tools available for assessing the estimated regression line is the residual plot. Recall from above that the experimental error or residual, ε_i, is the actual data value, y_i, minus the predicted value, $\hat{y}_i$, based on the observed x_i's. A plot of these residuals against the predicted values should provide a graph portraying a structureless pattern or "noise". Additionally, residuals could be assessed for conformance to the gaussian (normal) distribution or formally tested (Draper and Smith, 1981). Extreme points would be recognizable and if warranted, omitted, with new estimates and refitting of the line performed. A cautionary note: extreme values are <u>NOT</u> necessarily wrong. If there is justification for omitting these points, then do it. If justification is not available or warranted, the extreme data may be important!

Popular software, while having the ability to perform calculations, often lacks diagnostic capabilities. As far as the software is concerned, the data are well-behaved. One disadvantage of least squares is that it assigns an equal weight to all data points, even those points deemed extreme. If extreme values are determined to be suspect, weighted least squares can be performed by assigning full weights to those data that are "OK" and lesser weights to suspect data (Gunter, 1992b). Unfortunately, this process is an iterative one (successive computing of estimates) and is not available in most non-technical software.

Another least squares disadvantage is that the procedure is based on averages. Extreme values have the effect of distorting the average. In normal data analyses, this would lead to the use of the median in place of the mean. Simple regression is no different in that instead of minimizing the mean of the squared residuals, we minimize the median of the squared residuals. This technique is computationally demanding and can often provide erroneous results.

A final problem that can occur concerns the assumption of a fixed regressor variable. If *x* is fixed (i.e., measured without error), and all the data are well-behaved, least squares estimation is inarguably the best method. In far too many cases, *x* is not measured without error. When this is the case, lines obtained by the usual methods can be very misleading. In fact, instead of a single regression line, *Y* on *X*, there exists another line, *X* on *Y*, which is the basis for calibration. The reason is that not only is there uncertainty of the estimates caused by the unknown components of experimental error, ε_i, but now uncertainty arises from the unknown error associated with the determination of *X*. I will not discuss the technical problems associated with this problem but interested readers will find a great deal of information on this topic (e.g., Bartlett, 1949; Sampson, 1974; Fuller, 1987; Draper, 1991).

To illustrate some of the ideas presented in this paper, a practical set of data will be investigated.

APPLICATION

The data in Table 1 were used by Carruthers and Oldfield (1962) to illustrate the relationship between select raw juice impurities (g/100 S) and second carbonation purity. This relationship forms the basis for the so-called 'Carruthers Equation' and has been used extensively throughout the beet sugar industry.

They found a model

$$Purity \ = \ 100.9 \ - \ 0.00143 \ impurity \ value$$

which was deemed appropriate over a wide range of conditions. The impurity value is calculated via the following equation:

$$impurity \ value \ = \ 2.5 \ K^+ \ + \ 3.5 \ Na^+ \ + \ 10 \ Amino\text{-}N \ + \ betaine$$

Fitting this model to the data of Table 1 yields the following information:

```
Coefficients          Estimate      Std. Error      P-value

Intercept             100.8971        1.0032         0.0001
impurity value        -0.0014         0.0002         0.0001

R² = 0.82        S² = 0.18
```

Table 1. Data from Carruthers and Oldfield (1961) comparison of second carbonation purities and impurity value.

Factory and year	Second carb purity	Impurity value (mg/100 S)
1958/59 Campaign		
Cantley	93.3	5370
Colwick	92.7	5660
Brigg	92.7	5690
Poppleton	92.4	6300
Kelham	92.3	5970
Kidderminster	91.8	6730
Spalding	90.7	7570
Bardney	90.7	7340
1959/60 Campaign		
Poppleton	92.4	6320
Selby	92.2	6320
Colwick	91.8	6270
Bury	91.8	6420
Brigg	91.7	6470
Ipswich	91.1	7010
Spalding	91.0	6620
Kidderminster	91.0	7300
Felsted	90.7	6720
Wissington	90.5	6800
Kelham	90.4	7240
Cantley	90.3	6940
Bardney	89.4	7720
Mean	91.5	6609
Variance	1.0	401143

The results match those of Carruthers and Oldfield. Although not reported in the original paper, regressing purity on impurity value accounted for 82% of the variation in the data. Both parameter estimates were significantly different from 0 ($p < .01$ in both cases).

With slight modification, this equation is the currently adopted form used in estimating sugar loss to molasses and second carbonation purity. The modifications usually include removal of betaine from the equation and using a 9.5 multiplier for amino-N instead of 10. For details on the methodology, see Carruthers and Oldfield (1962) or Carruthers, Oldfield, and Teague (1962).

The first step in assessing the aptness of the adopted model was to construct a histogram of the actual data and residuals. Although not presented here, both histograms indicate well-behaved data. A test for normality did not indicate any problems. The next step was to obtain plots of the residuals versus predicted and impurity value data. Typically, such plots, assuming a correct model, will reveal a random scattering of points. Figures 1 and 2 provide these plots. Both figures indicate that the variance is not constant, as assumed by the model employed. Figure 1 shows that residuals increase with increasing predicted values, while in Figure 2, decreasing residuals are associated with increasing impurity values. Such "funneling" effects give evidence for unequal variation (i.e., heteroscedasticity). The usual method to correct this variance fluctuation is to use weighted least squares. An alternative is to transform the purity values to make the variance more stable.

If the information provided by the plots above had indicated no abnormalities in the data, the equation as developed, would be optimal. The fact that there are some discrepancies indicates that the least squares estimates above are misleading. The fluctuation in variance noted earlier can be traced to the raw data presented in Table 1. Note that the data were taken from two campaigns. Separating the means and variances for each year yields,

	Mean		Variance	
Variable	1958/59	1959/60	1958/59	1959/60
Purity	92.1	91.1	0.9	0.7
Impurity Value	6329	6781	661184	198558

The purity means and variances are similar for both years, but the impurity values differ, particularly the variances. This implies that the impurity observations arose from different populations and hence, have different variances. Even more critical, the assumption of no measurement error in *X* appears to be unjustified since a great deal of uncertainty is associated with the determination of the impurity value. This uncertainty is a total of all the variability surrounding the determination of each component used to calculate the impurity value. In light of this information, the simple linear regression model is not appropriate for this data.

In order to complete our discussion, an alternative is presented that would allow us to use impurity value to predict purity. For this example, it is assumed, perhaps incorrectly, that the measurement error associated with *X* is negligible. Given that the data arise from different sources, we can employ a regression model, using campaign year as a dummy variable (Draper, 1981). First, fitting a SLRM to both sets of data gives,

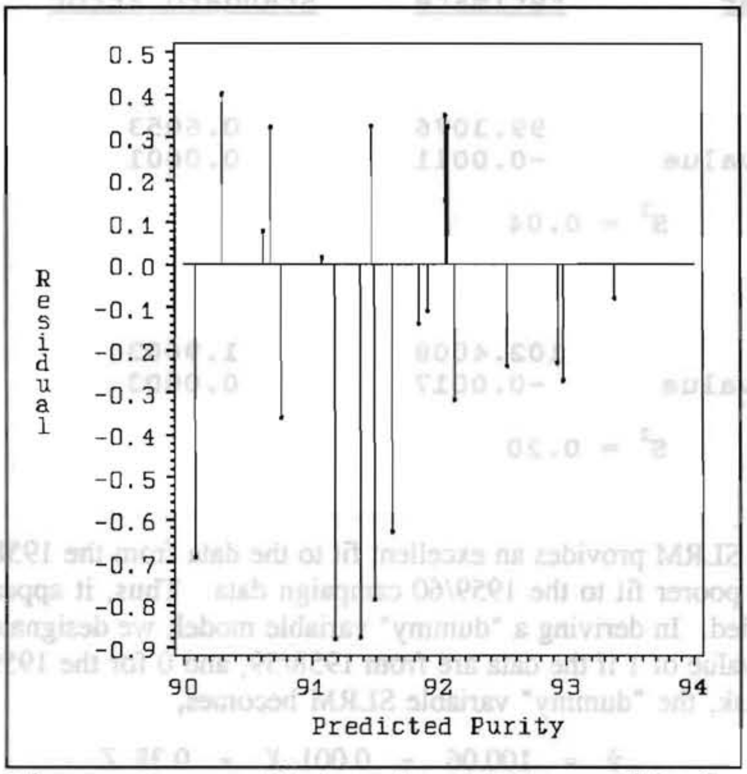

Figure 1. Plot of SLRM residuals versus predicted purity.

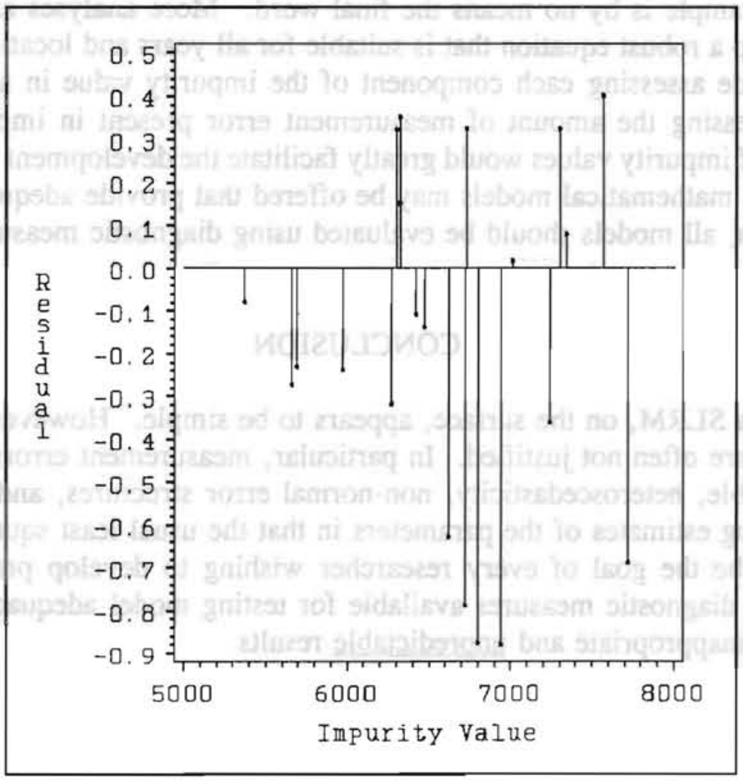

Figure 2. Plot of SLRM residuals versus impurity value (g/100S).

Year/Parameter	Estimate	Standard Error	Prob.
1958/59			
intercept	99.3076	0.6053	.0001
impurity value	-0.0011	0.0001	.0001
$R^2 = 0.96$	$S^2 = 0.04$		
1959/60			
intercept	102.4008	1.9603	.0001
impurity value	-0.0017	0.0003	.0001
$R^2 = 0.75$	$S^2 = 0.20$		

It is evident that a SLRM provides an excellent fit to the data from the 1958/59 campaign year, while providing a poorer fit to the 1959/60 campaign data. Thus, it appears that splitting out the data was justified. In deriving a "dummy" variable model, we designate a new variable, Z, which takes on a value of 1 if the data are from 1958/59, and 0 for the 1959/60 campaign data. Performing this task, the "dummy" variable SLRM becomes,

$$\hat{y} = 100.06 - 0.001 X + 0.38 Z$$

This equation yielded a $R^2 = 0.86$ and $S^2 = 0.16$. Thus, we have increased the amount of variation accounted for by the model, while improving upon the original precision.

This simple example is by no means the final word. More analyses and data are required in order to develop a robust equation that is suitable for all years and locations. Some ideas for future work include assessing each component of the impurity value in a multiple regression situation, and assessing the amount of measurement error present in impurity determination. True replication of impurity values would greatly facilitate the development of a robust equation. More complicated mathematical models may be offered that provide adequate representation of the theory. Again, all models should be evaluated using diagnostic measures.

CONCLUSION

The fitting of a SLRM, on the surface, appears to be simple. However, many assumptions are required that are often not justified. In particular, measurement errors associated with the independent variable, heteroscedasticity, non-normal error structures, and extreme values can all yield misleading estimates of the parameters in that the usual least squares methods are not valid. It should be the goal of every researcher wishing to develop predictive equations to employ the many diagnostic measures available for testing model adequacy. Failure to do so can often lead to inappropriate and unpredictable results.

SUMMARY

When linear regression methods are used to predict a response, *Y*, from some independent variable, *X*, consideration as to the nature of the variables must be given. Usual least squares estimation is simple and appropriate when data exhibit no abnormalities. Extreme data can influence these estimates and be misleading. Heteroscedasticity can also lead to inconsistent estimates of the parameters. Failure of the data or residuals to conform to assumptions can also cause difficulties. Weighted least squares and "dummy" variable models can often alleviate such difficulties.

In using simple linear regression methods, usual practice is to regard the *X*'s as fixed quantities not subject to error. Under certain conditions (i.e., the range in *X* is large compared to the variance), this assumption is not detrimental to the analysis. If the *X*'s are measured with error, and this error is large relative to the range observed, then the usual simple linear regression model yields parameters that are misleading, and hence, inappropriate. Many methods exist for an2lyzing such data. A particularly useful class of techniques, measurement error models, allows estimation of parameters when *X* is subjected to error. These models are also useful in calibration models, where prediction of *X* is desired.

Methods outlined in this paper were applied to a set of data used in developing the so-called "Carruthers Equation". Results using the simple linear regression model indicated that the assumption of constant variance was not justified. Furthermore, *X* was measured with error. The lack of constant variance was attributed to the data arising from two different campaigns. A "dummy" variable model was suggested. The resulting estimates were more precise than with the SLRM. More of the variation was accounted for by the 'dummy' variable model.

LITERATURE CITED

Bartlett, M.S. 1949. Fitting a straight line when both variables are subject to error. Biometrics 5:207-212.

Carruthers, A., and J.F.T. Oldfield. 1962. Methods for the assessment of beet quality. pp 224-248. *In*: The Technological value of the sugar beet. XI Session of the Commission of Internationale Technique DeSucreria Proc. 1960. Elsevier Publishing Co., New York.

Carruthers, A., J.F.T. Oldfield, and H.J. Teague. 1962. Assessment of beet quality. 15th Annual Technical Conference of the British Sugar Corporation Ltd., Nottingham, England. 28 pp.

Draper, N.R., and H. Smith. 1981. Applied Regression Analysis. John Wiley & Sons, Inc., New York. 709 pp

Draper, N.R. 1991. Straight line regression when both variables are subject to error. Proc. 1991 KSU Conf. App. Stat. in Agric. Manhatten, KS. pp 1-18.

Fuller, Wayne A. 1987. Measurement Error Models. John Wiley & Sons, Inc., New York. 440 pp

Gunter, Bert. 1992a. Fitting a line to data. Part 1: Why a best-fitting line might not be. Quality Progress, October. pp 113-123.

Gunter, Bert. 1992b. Fitting a line to data. Part 2: Alternatives to least squares. Quality Progress, December. pp 89-92.

Mandel, John. 1964. The Statistical Analysis of Experimental Data. Dover Publications, Inc., New York. 410 pp

Sampson, A.R. 1974. A tale of two regressions. J. Amer. Stat. Assoc. 69:682-689.