BRL R 1842

# BRL

REPORT NO. 1842

# LINEAR STATISTICAL REGRESSION AND FUNCTIONAL RELATIONS

Frank E. Grubbs

November 1975

Approved for public release; distribution unlimited.

USA BALLISTIC RESEARCH LABORATORIES
ABERDEEN PROVING GROUND, MARYLAND

Destroy this report when it is no longer needed.
Do not return it to the originator.

Secondary distribution of this report by originating or sponsoring activity is prohibited.

Additional copies of this report may be obtained from the National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22151.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

*The use of trade names or manufacturers' names in this report does not constitute indorsement of any commercial product.*

14 BRL-1842

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

**REPORT DOCUMENTATION PAGE** | READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
REPORT NO. 1842

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6 4. TITLE (and Subtitle)
Linear Statistical Regression and Functional Relations.

5. TYPE OF REPORT & PERIOD COVERED
9 Final Rept.,

6. PERF~~ORMING ORG. REPORT NUMB~~ER

10 7. AUTHOR(s)
Frank E. Grubbs

8. ~~CONTRACT OR GRANT NUMBER(s)~~
16 DA-1T161102A14B

9. PERFORMING ORGANIZATION NAME AND ADDRESS
US Army Ballistic Research Laboratories
Aberdeen Proving Ground, Maryland 21005

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
US Army Materiel Command
5001 Eisenhower Avenue
Alexandria, VA 22333

12. REPORT DATE
11 NOV [redacted] 75

13. NUMBER OF PAGES
57 12 52p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Linear Least Squares, Linear Regression, Functional Relations, Curve Fitting.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An approach to linear least squares or regression with errors of measurement in either one or both variables is covered, giving a computationally convenient procedure for all of the appropriate statistical significance tests or tests o: hypotheses concerning the true unknown parameters and the fitted line. The problem of meaningful physical functional relations is also discussed, showing the relation to and a comparison with usual least squares. Pertinent examples illustrating various applications of the theory are also given.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

050 750

# LINEAR STATISTICAL REGRESSION AND FUNCTIONAL RELATIONS

## SUMMARY

An approach to linear least squares or regression with errors of measurement in either one or both variables is covered, giving a computationally convenient procedure for all of the appropriate statistical significance tests or tests of hypotheses concerning the true unknown parameters and the fitted line. The problem of meaningful physical functional relations is also discussed, showing the relation to and a comparison with usual least squares. Pertinent examples illustrating various applications of the theory are also given.

# TABLE OF CONTENTS

# I. INTRODUCTION

A rather frequent and important problem in research and development is that of finding an appropriate relationship, or best fit, between variables of interest, i.e. fitting equations to data, and testing any hypotheses concerning the physical values or the relation of the parameters studied.In addition, and as usual, we would like to summarize experimental data in the form of an equation, and be able to predict future or expected occurrences from our fitted or "empirical" law. Indeed, in many problems it is now of some importance to be able to place confidence bounds on the various physical parameters which can be estimated or inferred from the data we have developed in an experiment.

Needless to say, this is more of an involved problem than appears on the surface, because errors of measurement are made in all determinations of values of the variables in an experiment, and in many cases we also run into the problem of random or unaccounted-for variations in addition to the physical laws (or functional relations) we seek to sort out of the "noise". Of course, we might say that we would really like to establish a law or enduring relationship between variables or parameters of interest, which is free of measurement error or other variations of extraneous interest. In addition, it also becomes of importance to know just how precise or accurate our final prediction is, since it might be desirable to conduct more experiments, depending on our subsequent uses of the fitted equation. A general, but simple and enduring law makes a very definite contribution to science and technology.

We should remark initially and keep in mind that the practice of transforming variables to linear functions or relations, as is often done in the physical sciences or in engineering, i. e. attempts toward "linearizing the data", is an excellent one indeed, as we will see, for it helps to establish relationships between complex quantities and simplifies much of the resulting analysis. Furthermore, it is usually not difficult to transfer statistical or physical statements about the transformed data back to equivalent ones about the original variables. For this reason, we will cover the case of linear least squares or linear regression in appropriate depth and take into account functional or "structural" relations of the variables involved. We will, therefore, start with the case of linear regression between an independent variable which is assumed to be free of (measurement) error and the dependent variable which is measured or found with error of determination, and later proceed to more complex cases. It is highly desirable in this connection to distinguish between "controlled" or "fixed" variables, random variables (variates) and errors of measurement which may be either random or systematic in character.

## II. LINEAR LEAST SQUARES OR REGRESSION FOR A DEPENDENT VARIABLE (MEASURED WITH ERROR) AND AN INDEPENDENT VARIABLE (WITHOUT ERROR)

In dealing with experimental data involving two variables x and y, for example, time and distance measurements, muzzle velocity and range measurements, etc., there may appear to be a trend or some relation (linear or otherwise) between the plotted values of x and y. We will be interested here in estimating the relation between x and y, and judging statistically whether or not the relation estimated is a significant one. The method we will use is generally referred to as the "Least Squares" technique, or i.e. the finding of the regression of y on x,although there are other methods of fitting a law between two or more variables, for example, the method of Maximum Likelihood. In the method of Least Squares we assume a model or relation, i.e. linear, quadratic, etc., which involves certain unknown parameters or coefficients, and fit the hypothesized curve to the two (or more) variables in a manner such that the sum of squares of residuals or deviations from the fitted curve is a minimum. The significance of the fitted curve will then be tested statistically, and confidence bounds placed on predictions.

Our approach will consist of combining the physical and statistical points of view, i. e., our models or assumptions will take into account both the functional or structural relation between true values of the variables and the statistical treatment of variates or errors of measurements and their probability distribution. In the model of this section, the independent variable is assumed to be free of error, with only the dependent variable subject to error.

### The Line - One Variable (y) Subject to Error

Suppose we are dealing with two observable variables, x and y, which are connected by an apparent linear relation. Suppose further that the variable y not only depends on x but is also subject to random errors of measurement (i.e.,y, as measured physically, includes an error of measurement), whereas x is a "fixed" or controlled variable which is free of errors of measurement, or relatively free of errors as compared to the variable y. Over the interval of physical interest, it will be assumed that variability in the errors of y is essentially constant. The mean value of y depends on the value of x considered and the variance of y, i.e.$\sigma^2_{y_x}$, about the hypothesized linear relation is independent of the value of x or is constant over the range of x considered.

In order to illustrate our point clearly,we have selected a particular, yet simple example from the ASTM Manual on Fitting Straight Lines (1962). The observed data were obtained for calibration of a new method (gravimetric determination) for estimating the amount of calcium in the presence of large amounts of magnesium. The data are given in Table I for known amounts of CaO.

Table I. Gravimetric Determination of Calcium in the Presence of Magnesium

| x<br>CaO Actually Present<br>(mg) | y<br>CaO Found by New Method<br>(mg) |
|---|---|
| 20.0 | 19.8 |
| 22.5 | 22.8 |
| 25.0 | 24.5 |
| 28.5 | 27.3 |
| 31.0 | 31.0 |
| 33.5 | 35.0 |
| 35.5 | 35.1 |
| 37.0 | 37.1 |
| 38.0 | 38.5 |
| 40.0 | 39.0 |

The basic reasons for selecting this particular example should be clear. A plot of y against x indicates (near) linearity, as it should. The amount of CaO actually present should be "free of measurement error", the slope of the true line should be 45°, and the line should pass through the origin for the assumption of linearity and good calibration. In addition, the error of measurement for the new method should be acceptable. We will, therefore, look into each of these questions in sufficient detail.

We should remark here that x and y are not random variables. The CaO actually present, or x, has been purposely varied over the range,so that y will correspondingly vary but with random measurement error.

The observed values of x and y are:

$$\begin{array}{c} x_1, y_1 \\ x_2, y_2 \\ \vdots \\ x_i, y_i \\ \vdots \\ x_n, y_n \end{array}$$

The linear model or assumption considered for the observed pairs $(x_i, y_i)$ is

$$x_i = \mu_i$$

$$y_i = \alpha+\beta\mu_i+d_i = \eta_i+d_i \qquad (1)$$

We use the above notation to indicate that $d_i$ is an error in the measurement of y, and that $x_i$ is free of error as we can control or measure its true value $\mu_i$ in this case. (If $x_i$ contained error we would write it as $x_i = \mu_i+e_i$, where $\mu_i$ is the true value and $e_i$ is an error of measurement in x.) The relation

$$\eta = \alpha + \beta\mu \qquad (2)$$

is called the true (functional) relation between the parts of y and x we are interested in. It is also the true regression in our simple model.

The errors, $d_i$, have mean or expected value, $E(d_i) = 0$, and variance in the errors $E[d_i-E(d_i)]^2 = \sigma_d^2 = \sigma_{y_x}^2$ or simply $\sigma^2$, the constant variance about the fitted regression line.

Thus, the mean value of an observed y for a given value of x is

$$E(y) = E(\alpha+\beta x+d) = \alpha+\beta x = \alpha+\beta\mu$$

The variance of y about its population mean, $\alpha+\beta x = \alpha+\beta\mu$, is $E[y-\alpha-\beta x]^2 = E(d_i^2) = \sigma^2_{y_x} = \sigma^2_d$, i.e. the population "variance of residuals", or variance about the regression line.

Of course, for a small sample of n observed pairs $(x_i, y_i)$, it will not be possible to estimate $\alpha$ and $\beta$ very precisely. Our fitted line will therefore be of the form

$$y = a+bx$$

where a and b are estimates of $\alpha$ and $\beta$, respectively, and therefore subject to "error" or statistical variation.

We estimate $\alpha$ and $\beta$ from a and b respectively by determining a and b so that

$$\phi = \sum_{i=1}^{n} (y_i-a-bx_i)^2 \qquad (3)$$

is a minimum.

Now

$$\frac{\partial\phi}{\partial a} = -2\sum_{i=1}^{n}(y_i-a-bx_i) = -2[\Sigma y_i-na-b\Sigma x_i],$$

and we find also that

$$\frac{\partial\phi}{\partial b} = -2\sum_{i=1}^{n}(y_i-a-bx_i)x_i=-2[\Sigma x_i y_i-a\Sigma x_i-b\Sigma x_i^2].$$

Equating $\frac{\partial\phi}{\partial a}$ and $\frac{\partial\phi}{\partial b}$ to zero, we obtain the well-known Normal equations:

$$na + (\Sigma x_i)b = \Sigma y_i \qquad (4)$$

$$(\Sigma x_i)a + (\Sigma x_i^2)b = \Sigma x_i y_i.$$

Solving for a and b, we find

$$a = \text{Est. of } \alpha = \frac{(\Sigma y_i)(\Sigma x_i^2) - (\Sigma x_i y_i)(\Sigma x_i)}{A_{xx}}$$

$$= \bar{y} - b\bar{x} \quad \text{or} \quad \frac{1}{n}[\Sigma y_i - b\Sigma x_i] \tag{5}$$

$$b = \text{Est. of } \beta = \frac{A_{xy}}{A_{xx}} \quad , \tag{6}$$

where $A_{xx} = n\Sigma x_i^2 - (\Sigma x_i)^2$ and $A_{xy} = n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)$.

These quantities are established for computational purposes, as they may be used free of rounding error and have advantages which the reader will appreciate in what follows no doubt.

The variance of residuals $\sigma_y^2 = \sigma^2$,cr that is the variance of an individual deviation from the fitted line is estimated from

$$S_{y_x}^2 = S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \frac{\Sigma y_i^2 - a\Sigma y_i - b\Sigma x_i y_i}{n-2} \tag{7}$$

$$= \frac{1}{n(n-2)}\left[A_{yy} - \frac{A^2{}_{xy}}{A_{xx}}\right].$$

The quantity,

$$r = r_{xy} = \frac{A_{xy}}{\sqrt{A_{xx}A_{yy}}} \quad , \tag{8}$$

is called the product moment correlation coefficient. For very large samples, $\sigma_{y_x}^2 = \sigma_d^2 = \sigma_y^2\,(1-\rho^2)$, where $\rho^2$ is the population correlation coefficient between the variables x and y. (Note that $\beta = \rho\sigma_y/\sigma_x$ also.) Now it can be shown that the mean or expected values of a and b are $\alpha$ and $\beta$, respectively, and therefore are unbiased estimates. That is,

$$E(a) = \alpha \quad \text{and} \quad E(b) = \beta,$$

since $A_{xx}$ is a constant, $E(A_{xy}) = \beta A_{xx} + E(A_{xd})$, $E(A_{xd}) = 0$, and $E(a) = E(\bar{y} - b\bar{x}) = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha$.

Under the assumptions, the following can also be proven:

$$\sigma_b^2 = E(b-\beta)^2 = \frac{n}{A_{xx}} \sigma_d^2 = \frac{n}{A_{xx}} \sigma^2 , \qquad (9)$$

$$E(A_{xd}^2) = n\sigma^2 A_{xx}, \qquad \text{and}$$

$$\sigma_a^2 = E(a-\alpha)^2 = E(\bar{y}-b\bar{x}-\alpha)^2 = \frac{\sigma^2}{n} + \frac{n\bar{x}^2\sigma^2}{A_{xx}} = \frac{\sigma^2\Sigma x^2}{A_{xx}} , \qquad (10)$$

the expectation of the cross-product term vanishing. Finally, the expectation of (7) turns out to be

$$E(S_{y_x}^2) = \sigma_d^2 = \sigma^2. \qquad (11)$$

Since the x's are free of error under the assumptions, then it can be seen from (5) and (6) that a and b are both linear functions of the errors, $d_i$.

Now (11), (10) and (9) give us the mean value of the computed variance of residuals, $S_{y_x}^2$, which is based on (n-2) degrees of freedom (d.f.), and the variances of the estimates, a and b. Thus, if we now assume that the errors, $d_i$, are normally distributed, and since $S_{y_x}^2 = S^2$ is an estimate of $\sigma^2$ based on n-2 d.f., then for independence of the $d_i$, and b and S,

$$t_b = \frac{(b-\beta)\sqrt{A_{xx}}}{S\sqrt{n}} \qquad (12)$$

follows Student's "t" distribution with n-2 d.f. Hence (12) can be used for testing the hypothesis that $\beta=0$, or that the true slope $\beta$ equals any other constant value, $\beta_0$, we may choose. Moreover, a confidence bound on the true unknown value of $\beta$ may be found from (12).

The customary test of significance for the intercept, widely used in textbooks on statistics, is in a manner similar to (12) given by

$$t_a = \frac{(a-\alpha)\sqrt{A_{xx}}}{S\sqrt{\Sigma x_i^2}} = \frac{a-\alpha}{S\sqrt{1/n+n\bar{x}^2 / A_{xx}}} \quad , \qquad (13)$$

which follows Student's "t" distribution with n-2 d.f. under the null-hypothesis. Furthermore, a confidence bound is easily found on the true unknown intercept, $\alpha$, from (13). The use of (13) in this connection, is quite proper if before looking at the data we decide in advance to use the "t" test for a hypothesized value of $\alpha$ in (13), or to place a confidence bound on the true, unknown intercept $\alpha$. It is proper if we intend to place confidence bounds on $\eta_0=\alpha+\beta x_0$ for a selected $x_0$, for which case we would in (13) replace a by $a+bx_0$, $\alpha$ by $\alpha+\beta x_0$ and $\bar{x}$ by $(x_0-\bar{x})$. However, if we make multiple statements about the line by picking several or many values of x, then $t_{\gamma/2}(n-2)$ must be replaced by $\sqrt{2F_\gamma(2, n-2)}$. Here, the probability is now $\geq 1-\gamma$ that all such statements are simultaneously correct. The reader is referred to Scheffe', Section 3.5 (1961). Thus, if a confidence bound on $\alpha$ is one of many such statements, one should use

$$a \pm \sqrt{2F(2,n-2)}\ (S)\ \sqrt{1/n+n\bar{x}^2/A_{xx}} \quad , \qquad (14)$$

where F(2,n-2) follows the Fisher-Snedecor "F" distribution with 2 and n-2 degrees of freedom.

If we pick some values of x, say x* (including x=0), and substitute these values of $x = x^*$ into the equation of the fitted line, i. e. $y = a + bx^*$, then all confidence bounds desired may be found from (14) by replacing a by $a+bx^*$, the $\bar{x}^2$ under the radical by $(x^*-\bar{x})^2$, and proper selection of the percentage point of F, using Scheffe's theorem.

To test the joint hypothesis that $\alpha = \alpha_0$ and $\beta = \beta_0$, we use the F distribution with 2 and n-2 degrees of freedom, F(2,n-2)

$$= [n(a-\alpha_o)^2 + 2n\bar{x}(a-\alpha_o)(b-\beta_o) + (\Sigma x^2)(b-\beta_o)^2]/2S^2. \quad (15)$$

A joint confidence region on α and β may be found from (15) by determining various pairs of $\alpha_o$ and $\beta_o$ for which (15) gives the values of F not exceeding the selected confidence level $F_\gamma(2, n-2)$.

A confidence region on any number of future values of y for given values $x = x_o$ may be found from

$$a+bx_o + \sqrt{2F(2,n-2)}(S) \sqrt{1+1/n+n(x_o-\bar{x})^2/A_{xx}}, \quad (16)$$

where we have simply added the variance of an individual.

Example

Using the data of Table I, we compute the following:

$n=10$, $\Sigma x=311$, $\Sigma x^2 =10100$, $\bar{x}=31.10$, $S_x=6.90$, $A_{xx}=4279$

$\Sigma y=310.10$, $\Sigma y^2=10055.09$, $\bar{y}=31.01$, $S_y=6.98$, $A_{yy}=4388.89$

$\Sigma xy=10074.80$, $A_{xy}=4306.90$, $S_{xy}=\sqrt{A_{xy}/n(n-1)} = 6.92$

$b= A_{xy}/A_{xx} = 1.0065$, $a = \bar{y} - b\bar{x} = -.2922$

$S^2_{y_x} =(A_{yy}-A^2_{xy}/A_{xx})/n(n-2) = .6729$ and $S_{y_x} = .8209$

As already indicated, we are particularly interested in whether the true slope of the line is 45° ($\tan\theta=1$) and whether the true intercept can be considered to be zero, indicating proper calibration for the gravimetric determination (new) method. To test whether $\beta=1$, we compute

$$t_b = (b-\beta)\sqrt{A_{xx}}/S\sqrt{n} = (1.0065-1.0000)\sqrt{4279}/(.8209)\sqrt{10} = .16,$$

which is not statistically significant at the 95% level. To test whether $\alpha = 0$, we compute

$$t_{\gamma/2}(n-2) = a/S(1/n + n\bar{x}^2/A_{xx})^{1/2} = -.23,$$

which is not significant either.

To make the joint test of hypothesis that $\alpha=0$, $\beta = 1$, we use (15) and find that the observed $F(2,n-2) = F(2,8)= 3.21$, which is not significant at the 95% level, concluding that the line is indeed a good fit to the data. [The observed F does exceed the 90% significance level, since $F_{.90}(2, 8) = 3.11$].

For any given level of CaO actually present, such as $x = x^* = 20$, or 40, say, the standard error of prediction for that value from the fitted line, $y=a+bx=-.2922+1.0065x^*$, is given by

$$S\sqrt{1/n+n(x^*-\bar{x})^2/A_{xx}}. \tag{17}$$

Thus, if we take $x^* = 20$, and substitute this value in the above equation of the fitted line, we get its standard error

$$S_y(\text{predicted}) = (.8209)\sqrt{1/10+10(20-31.1)^2/4279} = .51 \text{ mg}.$$

As already indicated, the confidence interval for a future (individual) observation on y, call it $y_o$, corresponding to a given true value of $x = x_o$, say, may be found from (16).* Thus, a 95% confidence bound on a new observed y for $x = x_o = 20$, is given by

$$-.2922+1.0065(20)\pm t_{.975}(8)(.8209)\sqrt{11/10+10(20-31.1)^2/4279}$$

$$= 19.84 \pm 2.23 = 17.61 \text{ to } 22.08 \text{ mg}.$$

(Note that the standard error for the single future observation is .97 as compared to value of .51 mg based on the same point of the line.)

Since x is regarded as the "true" value, measured or determined without error, then of more particular interest might be confidence bounds on the true amount of CaO for a given measurement by the (new) gravimetric method. Thus, suppose we have measured y to be $y = y' = 20.1$ mg. Then, the approximate confidence bound on x, obtained by substituting y' in the equation of the line $y'= a+bx$ and solving for x, may be found for the a priori y' from

$$(y'-a)/b + t_{\gamma/2}(n-2)(S/b)\ \sqrt{1/n+n[(y'-a)/b-\bar{x}]^2/A_{xx}}. \tag{18}$$

* With $\sqrt{2F}$ replaced by "t" for a particular a priori value of $x = x_o$

For y' = 20.1, substitution in (18) gives a confidence bound on x of

$$20.26 \pm 1.15 = 19.11 \text{ to } 21.41,$$

so that the probability statement for y' = 20.1 mg is

$$\Pr[19.11 \text{ mg} \leq \text{True CaO} \leq 21.41 \text{ mg}] \equiv .95.$$

In the above, note that we have used the fitted line to improve on accuracy of prediction, as compared to that of a single determination by the new method. If the error of prediction is too large for the practical problem involved, then we might improve on precision by taking more points (especially at the ends for a fitted line), or conclude that a better measurement method is needed.

Finally, concerning the example, we did not have a physical law or hypothesis for the fitted equation. We thus had to use the line. In examples below, we will nevertheless consider functional relationships, or appropriate physical laws in our analyses.

Suppose that instead of fitting the line y = a + bx, we had fitted $y = a_o + (x_i - \bar{x})$, i. e. we measure each $x_i$ from its mean. In this case, our Normal equations become

$$na_o + [\Sigma(x_i - \bar{x})]\, b = \Sigma y_i$$

$$[\Sigma(x_i - \bar{x})]\, a_o + [\Sigma(x_i - \bar{x})^2]\, b = \Sigma(x_i - \bar{x})y_i = \Sigma x_i y_i - \bar{x}\Sigma y_i$$

$$= \frac{1}{n} A_{xy}.$$

But $\Sigma(x_i - \bar{x}) = \Sigma x_i - n\bar{x} = 0$

Hence,

$$na_o = \Sigma y_i \quad \text{or} \quad a_o = \frac{1}{n}\Sigma y_i = \bar{y}, \tag{19}$$

$$b = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2} = \frac{A_{xy}}{n\Sigma(x_i - \bar{x})^2} = \frac{A_{xy}}{A_{xx}} \tag{20}$$

as before (no change).

Note, however, that $a = a_o - b\bar{x} = \bar{y} - b\bar{x}$, which agrees with the intercept a fitted from the equation $y = a + bx$ as before. The importance of this result is that by a simple transformation of the independent variable (i. e. by choosing the origin of x at its mean value), we can always eliminate the constant term if desired.

The variance of the intercept a was found to be $\frac{\Sigma x^2}{A_{xx}} = \sigma_d^2$.

The variance of $a_o$, however, turns out to be $\frac{1}{n}\sigma_d^2$, as one might surmise as it is simply the variance of an average value.

## Transformation of Original Data.

In many problems the original variables x and y may be so large (or small) that it would be inconvenient to work with them. Hence, we may want to subtract some constant from one or both variables, or multiply or divide the original numbers by some constant factor. Thus, suppose we transform the $x_i$ and $y_i$ as follows:

$$u_i = c(x_i - h) \quad ; \quad v_i = d(y_i - k)$$

where c, d, h and k are selected constants.
Making these transformations, we find:

(i) $A_{uv} = n\Sigma uv - (\Sigma u)(\Sigma v) = c\ d\ A_{xy}$ or $A_{xy} = A_{uv}/c\ d$

(ii) $A_{uu} = c^2 A_{xx}$, or $A_{xx} = = A_{uu}/c^2$

(iii) $A_{vv} = d^2\ A_{yy}$, or $A_{yy} = A_{vv}/d^2$

(iv) $\Sigma u_i = c\Sigma x_i - nch$ ; $\Sigma v_i = d\Sigma y_i - ndk$

Hence,

$$b = \frac{A_{xy}}{A_{xx}} = \frac{A_{uv}}{cd} \cdot \frac{c^2}{A_{uu}} = \frac{c}{d} \cdot \frac{A_{uv}}{A_{uu}},$$

and

$$a = \frac{1}{n}[\Sigma y_i - b\Sigma x_i] = \frac{1}{nd}[\Sigma v_i + ndk - \frac{A_{uv}}{A_{uu}}(\Sigma u_i + nch)]$$

$$= \frac{1}{d}\bar{v} - \frac{A_{uv}}{dA_{uu}}\bar{u} + k - \frac{c}{d}h\frac{A_{uv}}{A_{uu}},$$

$$S^2 = \frac{1}{n(n-2)d^2}[A_{vv} - \frac{A_{uv}^2}{A_{uu}}],$$

and $\Sigma x_i^2 = \frac{1}{c^2}\Sigma u_i^2 + 2\frac{h}{c}\Sigma u_i + n h^2.$

Therefore, using the above formulae, we may work with the transformed variables u and v, and find the required parameter estimates for the original variables x and y. Indeed, such transformations are often very convenient or necessary in regression analyses.

## Equal Spacing of the Independent Variable

In some problems it may be that the x's are equally spaced, i. e.

$x_1 = e$, $x_2 = e + f$, $x_3 = e + 2f$, ..., $x_i = e + (i-1)f$, ...,

and $x_n = e + (n-1)f$,

where f is the width of the uniform interval. In this case, it can be shown that

$$\sum_{i=1}^{n} x_i = ne + \frac{n(n-1)}{2}f \qquad ;$$

$$\sum_{i=1}^{n} x_i^2 = ne^2 + 2ef\frac{n(n-1)}{2} + f^2\left(\frac{n-1}{6}\right)(n)(2n-1)$$

$$A_{xx} = \frac{n^2f^2}{12}(n^2-1) \quad ; \quad \text{and} \quad A_{xy} = \frac{nf}{2}\sum_{i=1}^{n}(2i-n-1)\,y_i.$$

$$\text{Hence, } b = \frac{A_{xy}}{A_{xx}} = \frac{6\sum_{i=1}^{n}(2i-n-1)y_i}{nf(n^2-1)},$$

$$a = \frac{1}{n}[\Sigma y_i - b\Sigma x_i] = \frac{1}{n}[\Sigma y_i - \frac{6e-3f(n-1)}{f(n^2-1)}\sum_{i=1}^{n}(2i-n-1)y_i],$$

and

$$S^2 = \frac{1}{n(n-2)}\left[A_{yy} - \frac{A_{xy}^2}{A_{xx}}\right]$$

$$= \frac{1}{n(n-2)}\left[A_{yy} - \frac{3\{\sum_{i=1}^{n}(2i-n-1)y_i\}^2}{n^2-1}\right]$$

The above formulae give all the information required to find also the values of $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$, $t_a$, $t_b$, etc., as needed.

## III. LINEAR REGRESSION AND FUNCTIONAL RELATIONS - BOTH VARIABLES SUBJECT TO ERROR, BUT INDEPENDENT VARIABLE CONTROLLED

The problem of fitting lines or linear functional relations of some physical significance becomes much more complex for the important case where both the dependent and independent variables are subject to (random)measurement error. Here, one has the problem of finding the physical or functional relation for the true unknown parts of x and y in the presence of "noise", and it clearly becomes of importance to have some knowledge of, or be able to estimate, the relative sizes of the errors in y as compared to those of x, whether these errors are correlated with each other, or whether errors of measurements in the variables depend on the magnitude of physical values studied, etc. Indeed, there are more parametric quantities of interest than can possibly be estimated without rather severe assumptions on what is actually happening. The reader will appreciate this in what follows; however, it will be instructive to first return to the data of Table I and formula (1) to check upon our assumptions in the analysis of that data. In particular, we assumed that x, the amount of CaO actually present, was "free of error". In this connection, suppose we now replace equations (1) by

$$x_i = \mu_i + e_i, \tag{21}$$

and

$$y_i = \alpha + \beta\mu_i + d_i = \eta_i + d_i. \tag{22}$$

In other words, x is now measured with (random) error, e, in addition to y having error d as before, so that our problem is to estimate the true relation $\eta = \alpha + \beta\mu$,which is covered with noise. $\mu$ is not a random variable here.

In the above analysis, we considered that the errors $e_i$ were zero, or quite inconsequential, and that the variance of errors was zero,i.e. $\sigma_e^2 = 0$. For the observed $x_i$ in (21), we have from the definitions of variances and covariances, that

$$S_x^2 = \Sigma(x_i-\bar{x})^2/(n-1) = S_\mu^2 + 2S_{\mu e} + S_e^2. \tag{23}$$

Likewise, for the observed $y_i$ in (22), one obtains

$$S_y^2 = \beta^2 S_\mu^2 + 2\beta S_{\mu d} + S_d^2, \tag{24}$$

and for the covariance between the observed x's and y's, we get

$$S_{xy} = \beta S_\mu^2 + S_{\mu d} + \beta S_{\mu e} + S_{de} \ . \tag{25}$$

Now for the hypothesized or true linear relationship,$\eta=\alpha+\beta\mu$, we must be able to estimate $\alpha$ and $\beta$ accurately from the data. The expected values of $S_d^2$ and $S_e^2$ are $\sigma_d^2$ and $\sigma_e^2$, respectively, i.e. the variances in errors (of measurement), of y and x, and the quantity $S_\mu^2(=\sigma_\mu^2$ also really), or $S_\mu$, is a measure of the variation over the range of interest of the experiment. It is certainly important to know something about the relative magnitudes of $\sigma_d$, $\sigma_e$, and $\sigma_\mu$ for such information is in fact needed for best estimates of $\alpha$ and $\beta$. Then finally, the problem is made more difficult because of the covariances, $S_{\mu d}$, $S_{\mu e}$ and $S_{de}$, which could have non-zero expectations equal to $\sigma_{\mu d}$, $\sigma_{\mu e}$ and $\sigma_{de}$, respectively,in some applications. Thus, we have the formidable problem of being perhaps interested in some eight parameters, $\alpha$, $\beta$, $\sigma_d^2$, $\sigma_e^2$,$\sigma_\mu^2$ $\sigma_{\mu d}$, $\sigma_{\mu e}$ and $\sigma_{de}$, and far too few conditions to estimate them from! By assuming that the errors are not correlated with each other or with the levels of the values taken by $\mu$, and have constant variance over the range, then the expectations of all the covariance terms vanish, and we are left with the expectations of (23), (24) and (25), which are

$$\sigma_x^2 = \sigma_\mu^2 + \sigma_e^2,$$

$$\sigma_y^2 = \beta^2\sigma_\mu^2 + \sigma_d^2, \qquad (26)$$

and

$$\sigma_{xy} = \beta\sigma_\mu^2$$

Nevertheless, even though $\alpha$ is absent from these three equations, we still have four unknowns,$\beta$, $\sigma_d^2$, $\sigma_e^2$, $\sigma_\mu^2$. Thus, it is quite evident that some knowledge,even from past experience of the relative sizes of the variances in errors,$\sigma_d^2$ and $\sigma_e^2$, becomes rather critical indeed. If we know for the problem at hand that $\sigma_d = \sigma_e$, then of course solutions are forthcoming (although from small samples we could still run into negative estimates of the variances). With this background,however, we may proceed with the analysis of the data of Table I, and later discuss needed aspects of the overall problems of estimation.

For the example (Table I), we found that b = 1.0065 for the estimate of $\beta$ and that this value did not depart significantly from unity. Thus, since $S_{xy} = A_{xy}/n(n-1)$, we might estimate $\sigma_\mu^2$ from the last equation of (26), i. e. from $S_{xy}/b$ $=47.85/1.0065 = 47.54$, (or even from $S_{xy}/1 = 47.85$),and then $\sigma_e^2$ from the first of equations (26). We get $\hat{\sigma}_e^2 = S_x^2 - \sigma_\mu^2 =$

$47.54 - 47.54 = 0$, so that the assumption that $\sigma_e = 0$, or that x is "free of error" (except for possible calibration bias) certainly seems valid for the analysis of Table I data. We are therefore confident in treating x as "free of error", as we did.

Next, in approaching the case of error in both variables, we proceed with a very important result due to Berkson (1950), which has a profound effect on regression problems in the physical sciences. Berkson's result states that if the independent variable x is "controlled" even though it is otherwise "measured with error", the ordinary least-squares estimate of the slope in (6), i. e. $b = A_{xy}/A_{xx}$, gives an unbiased estimate of $\beta$ for the linear fit, and $a = \bar{y} - b\bar{x}$ is also an unbiased estimate of $\alpha$. To appreciate this result, we first note that so far we have considered only the errors, $d_i$ and $e_i$, to be random variables which have zero means and variances $\sigma_d^2$ and $\sigma_e^2$. As yet, we have not considered the possibility that the $\mu_i$ could be random variables, for in the physical sciences there are so many cases of interest where random sampling with respect to the $\mu_i$ is not carried out. That is to say, the $x_i$ are varied systematically over some particular range of interest in the experiment. This being the case, then the $x_i$ are brought to near fixed or "controlled" levels by setting the dial of an instrument, presetting the time or distance measurement, etc., or aiming for a fixed or preset level which is measured as $x_i$. Thus, from (21) we have as before that $e_i$ is a random variable but also that $\mu_i$ has been in effect made to be random about $x_i$ by controlling the $x_i$. Hence, $\mu_i = x_i - e_i$, and upon substituting this in (22) we have

$$y_i = \alpha + \beta x_i + (d_i - \beta e_i)$$

But since the expectations of $d_i$ and $e_i$ are zero and $x_i$ is fixed or controlled, we have the problem of fitting $y_i = \alpha + \beta x_i$ + a random error, which reduces to that of Section II, so that the ordinary least squares b becomes an unbiased, estimate of the true and unknown slope $\beta$! This means that due to the imposed method of sampling or taking the data, we have controlled the $\mu_i$ to narrow random ranges about the selected or set $x_i$, which are brought to given levels, so that linear regression with error only in the dependent variable is still appropriate. Moreover, since the expectations of the errors are zero and that of b is equal to $\beta$, then $a=y-bx$ is an unbiased estimate of the intercept $\alpha$. Berkson's (1950) result is therefore of great importance in wide fields of scientific investigation and experimentation, since relatively the variance in errors of x, or $\sigma_e^2$, is small compared to the overall variance of the $\mu_i$ (made possible by controlling the $x_i$), and the measured $x_i$ consequently average out over the imposed range to give an unbiased estimate of $\beta$ anyway. In summary, therefore, we are fortunate indeed for a wide class of problems where we can simply ignore the errors in the independent variable. (The author's experience in Army research and development is that controlling the independent variable is very widely practiced in curve fitting problems, and one rarely runs into the case where the $\mu_i$ are random or statistical variates except in the narrow range about the controlled $x_i$ discussed above. Hence, the Berkson model has very wide application). Finally, as will be seen, we may still estimate the values of the variances in errors of x and y, i. e. $\sigma_e^2$ and $\sigma_d^2$, the most critical problem being that of estimating $\beta$ accurately.

In view of the Berkson development, we will now give an example in penetration mechanics, the data for which we are indebted to Mr. Chester Grabarek of the Terminal Ballistics Laboratory, BRL. Furthermore, the data are not linear, but rather lie on the branch of a hyperbola, so that we will transform the variables to near linearity for analysis, and also attempt to illuminate our analysis with some physical meaning or functional relationship.

The data are given in Table II, covering an experiment on striking velocities and residual velocities for a 27 gram penetrator fired at 1/2" armor plate.

Striking velocities and residual velocities are plotted on Figure 1. For the higher striking and residual velocities at the upper part of the curve the slope should approach unity (angle of 45°), whereas it becomes infinite at the value of $V_S$ for which $V_R = 0$. For the higher striking velocities, all rounds penetrate the plate until the knee of the curve is reached, at which the chance of complete penetration varies from near 100% down to zero or near zero per-

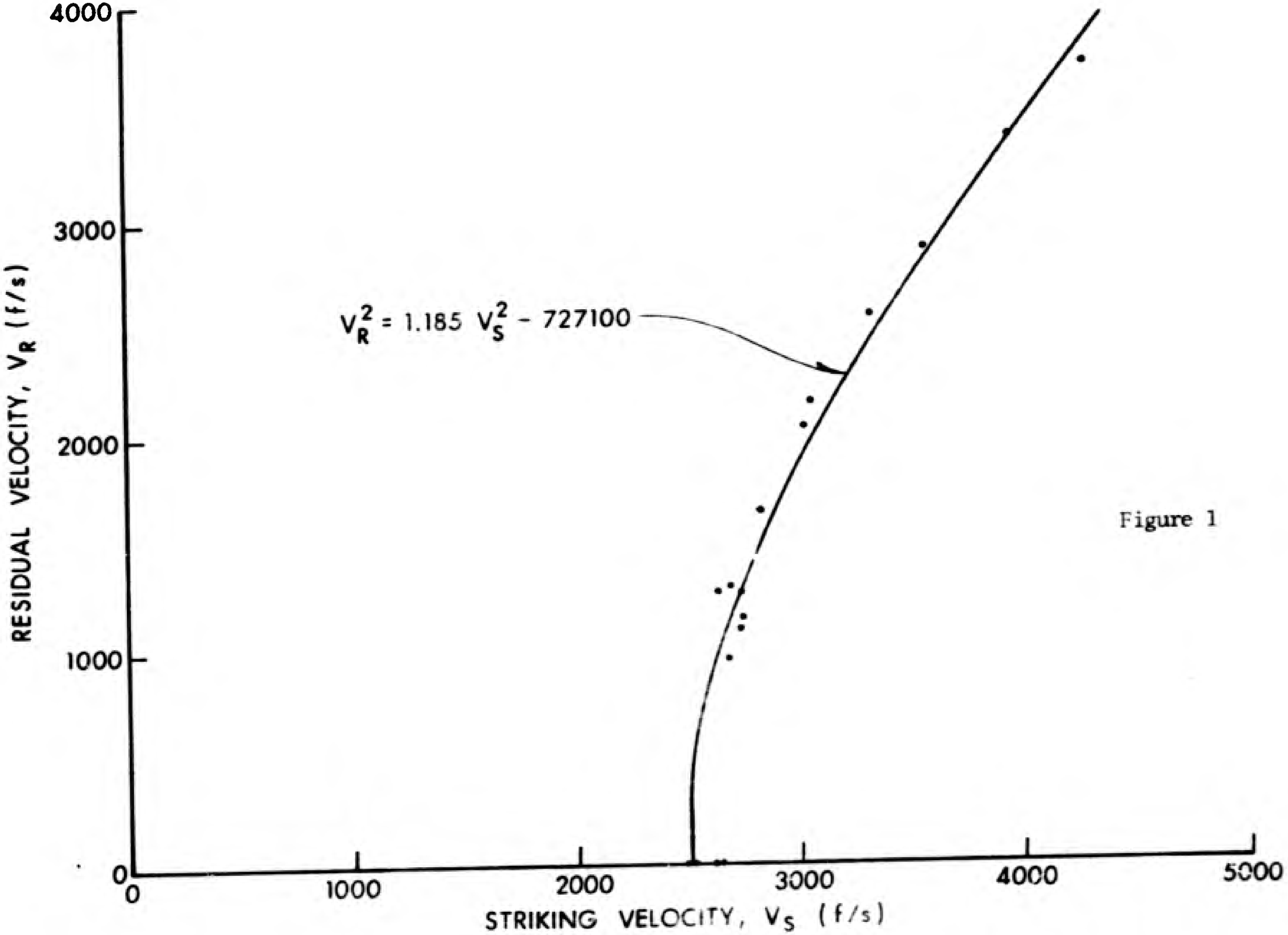
4000
3000
2000
1000
0
RESIDUAL VELOCITY, $V_R$ (f/s)
$V_R^2 = 1.185\ V_S^2 - 727100$
0
1000
2000
3000
4000
5000
STRIKING VELOCITY, $V_S$ (f/s)

Figure 1

TABLE II. Striking Velocities, Residual Velocities and Residual Masses for 27 gram Projectiles Fired Against 1/2" Armor Plate

| Striking Velocity (f/s) $V_S$ | Residual Velocity (f/s) $V_R$ | Residual Mass (grams) $M_R$ | y = $V_R^2/10^6$ | x = $V_S^2/10^6$ |
|---|---|---|---|---|
| 2487 | 0 | - | 0 | 6.185 |
| 2508 | 0 | - | 0 | 6.290 |
| 2611 | 0 | - | 0 | 6.817 |
| 2631 | 0 | - | 0 | 6.922 |
| 2680 | 950 | 14.267 | .903 | 7.182 |
| 2732 | 1102 | 16.572 | 1.214 | 7.464 |
| 2735 | 1154 | 14.204 | 1.332 | 7.480 |
| 2718 | 1265 | 12.527 | 1.600 | 7.388 |
| 2646 | 1273 | 11.816 | 1.621 | 7.001 |
| 2707 | 1292 | 12.276 | 1.669 | 7.328 |
| 2846 | 1648 | 18.419 | 2.716 | 8.100 |
| 3023 | 2036 | 18.894 | 4.145 | 9.139 |
| 3051 | 2157 | 16.064 | 4.653 | 9.309 |
| 3331 | 2522 | 17.970 | 6.360 | 11.096 |
| 3579 | 2859 | 19.604 | 8.174 | 12.809 |
| 3971 | 3382 | 19.627 | 11.438 | 15.769 |
| 4274 | 3702 | 19.837 | 13.705 | 18.267 |

cent at the "limit" or "critical" striking velocity for which the residual or exit velocity is zero (i. e. partial penetration). In this particular problem, one is very much interested in fitting an appropriate curve or law so that confidence bounds can be placed on the limit or critical striking velocity (x intercept). Although one might be tempted to exclude the $V_S$ for the four cases where $V_R = 0$, i. e. the partial penetrations, these are nevertheless valid points and will be included in our least-squares procedure below.

A plot of the square of the residual velocities versus the square of the striking velocity (last two columns of Table II) indicates a nearly linear relationship. Therefore, we will analyze the transformed variables $y = V_R^2/10^6$

and $x = V_S^2/10^6$. Also,since the independent variable may for practical purposes be regarded as a controlled variable, we may treat it as being essentially "free of error" and it seems natural to regard any function of the residual velocity, $V_R$, as the dependent variable.

For the transformed variables, x and y, we obtain

$$n = 17, \quad \Sigma x = 154.546, \quad \Sigma x^2 = 1598.068, \quad A_{xx} = 3282.690$$

$$n = 17, \quad \Sigma y = 59.530, \quad \Sigma y^2 = 484.163, \quad A_{yy} = 4686.950$$

$$\Sigma xy = 770.092, \quad A_{xy} = 3891.441$$

$$b = A_{xy}/A_{xx} = 1.185, \quad a = \bar{y} - b\bar{x} = 3.502 - (1.185)9.091$$

$$= -7.271$$

$$y = a + bx \quad \text{or} \quad V_R^2 = 1.185V_s^2 - 7271000$$

When $V_R = 0$, $V_S = 2477$ f/s, the "limit" velocity.

The variance of residuals is

$$S_{y_x}^2 = [A_{xx}A_{yy} - A_{xy}^2]/n(n-2)A_{xx} = .290, \quad \text{or } S_{y_x} = .538.$$

95% confidence bounds on the true unknown "limiting" x, i.e. for y = 0, are obtained from (18), where y' = 0, or

$$-a/b \pm t_{\gamma/2}(n-2)\ S_{y_x}/b)[1/n + n(-a/b - \bar{x})^2/A_{xx}]^{1/2}.$$

This gives

$$Pr(5.827 \leq x_{limit} \leq 6.451) = .95,$$

and since $V_S^2/10^6 = x$, we have for the original data that

$$Pr(2414 \text{ f/s} \leq V_{limit} \leq 2540) = .95,$$

so that the 95% confidence bound on the true unknown limit or critical velocity is 2540 - 2414 = 126 f/s wide for the single intercept.

Had the above statement been one of many similar ones about confidence bounds for various points on the line, then $t_{\gamma/2}(n-2)$ should be replaced by $\sqrt{2F_\gamma(2,n-2)}$ and the resulting confidence bounds for $V_{limit}$ would be 2396-2555 f/s.

The variance of residuals on the transformed scale is $S^2_{y_x} = .290$, but since $V_R = 1000\sqrt{y}$, we have $dV_R = 500\ y^{-1/2}dy$, and upon squaring and taking mean values we have the variance of residuals on the original scale of $V_R$, which is

$$\sigma^2_{V_R} \approx (250000/\bar{y})\sigma^2_{y_x} = (250000/3.502)(.290) = 2070\ 2$$

$$\text{or } \sigma_{V_R} = 144 \text{ f/s (for an individual value).}$$

At this stage, we might ask whether our assumption of x being "free of error" is met, or nearly so. In this connection, we note from the last equation of (26) that $\sigma^2_\mu = \sigma_{xy}/\beta$, and hence that

$$\hat{\sigma}^2_\mu = \text{Est. of } \sigma^2_\mu = A_{xy}/n(n-1)b = 12.07.$$

Now from the first of equations (26), we take

$$\hat{\sigma}^2 = \hat{\sigma}^2_x - \hat{\sigma}^2_\mu = A_{xx}/n(n-1) - \hat{\sigma}^2_\mu = 3282.69/(17)(16) - 12.07$$

$= 12.07 - 12.07 = 0$, which gives us some confidence in our procedure. We also observed from the second equation of (26) that our observed estimate of $\sigma^2_d$ turns out to be $\hat{\sigma}^2_d = .28$ or $\hat{\sigma}_d = .53$, which converted to the original scale of $V_R$ is 141 f/s versus the 144 f/s above.

In fitting the equation $V^2_R = 1.185\ V^2_S - 7271000$, we merely observed that the original data fall on the branch of a hyperbola type of curve, and hence we could linearize the data (or approximately so) by working with the squares of the striking and residual velocities. But what about the possibility of a "physical" fit or law? Here, we might consider fitting the residual energy versus the striking energy. In Table II, one notes that a third or more of the weight of the projectiles wear away in the penetration process. Nevertheless, it might make considerable sense to treat the "measured" residual energy as the dependent variable and the striking energy as the independent variable. We will actually take $x = m_S V^2_S/10^8 = 27V^2_S/10^8$ and $y = m_R V^2_R/10^8$,

$m_R$ varying and given in Table II. A plot of these new x's and y's indicates a nearly linear relationship. Our key computations now become

$n = 17, \quad A_{xx} = 239.301, \quad A_{yy} = 187.103, \quad A_{xy} = 210.721$

$b = .8806, \quad a = -1.523 \quad$ or $\quad y = -1.523 + .8806x.$

Using the average of the residual masses $(\bar{m}_R = 16.314)$,we now obtain the equation

$$V_R^2 = -9335540 + 1.457\ V_S^2$$

Also, $S_{y_x} = .078$, $V_s$(critical) = 2531, f/s, and

$$\Pr[2497 \text{ f/s} \leq V_S(\text{critical}) \leq 2565 \text{ f/s}] \approx .95$$

Thus, by using the "physical" law, the confidence interval has a width of 2565 - 2497 = 68 f/s, or 58 f/s shorter than the one based on $V_R^2$ and $V_S^2$. (We note that this "law" does not fit as well as the other one at the upper end of the curve, although the lower end is still of more interest nevertheless. We also note that raising the "measured" residual energy and the striking energy to about the .90 or .95 power might produce a slightly better linear relationship, but this would begin to depart from physical considerations.)

For the transformed data based on striking energy and "measured" residual energy, we have from (26) that

$$\hat{\sigma}_\mu^2 = .88, \qquad \hat{\sigma}_e^2 \approx 0.00, \qquad \text{and} \qquad \hat{\sigma}_d^2 \approx .10,$$

so that the assumptions still seem sufficiently valid, and the relation between striking and residual velocities is taken as $V_R^2 = 1.457\ V_S^2 - 9335540$. Moreover, the standard deviation of the random measurement error, d, is easily converted to the original scale of the residual velocity, $V_R$, and is approximately, $\sigma_{V_R} \approx 10^4 \sigma_{y_x} / 2\sqrt{m_R\ y} = 60$ f/s, much less than the value of 141 f/s above for $V_R^2$ versus $V_S^2$.

In summary, we have demonstrated the importance of trying to seek a physical relationship, transforming the original variables to near linearity for the regression analysis, and then being able to make statistical or probability statements about the original variables of interest.

If we knew that the slope of the line is unity from physical considerations,then there is no point in estimating it statistically, of course, the analysis thereby being simplified. Also, for more complex problems, one might well consider using various functions of the physical variables which result in linearity, with only the error following a statistical distribution. Indeed, regression problems are not all statistical, nor are they all physical,but rather it is the combination of both fields of interest which may result in wider practical value and utility.

We mentioned that proper estimation of the slope $\beta$ was important, and that unbiased estimates are needed. With regard to this matter, equations (26) are of considerably more help than might at first be realized. To begin with, if $\sigma_e = 0$, we note using the first and third equations that a proper estimate of $\beta$ is

$$\hat{\beta} = A_{xy}/A_{xx}$$

as we established in (6). If $\sigma_e$ is not zero but known, for example from past data or experience, then the first of (26) indicates that

$$\sigma_\mu^2 = \sigma_x^2 - \sigma_e^2,$$

so that an unbiased estimate of $\beta$ may be found, looking at the third equation of (26), from

$$\hat{\beta} = A_{xy}/[A_{xx} - n^2\sigma_e^2]. \qquad (27)$$

If $\sigma_d$ is known, then looking at the second and third equations of (26),we see that an estimate of $\beta$ is found from

$$\hat{\beta} = [A_{yy} - n^2\sigma_d^2]/A_{xy} \qquad (28)$$

If both $\sigma_d$ and $\sigma_e$ are known, then from the first and second equations of (26), we obtain the estimate

$$\hat{\beta} = [A_{yy} - n^2\sigma_d^2]^{1/2}/[A_{xx} - n^2\sigma_e^2]^{1/2}. \quad (29)$$

The estimates (27), (28) and (29) are not maximum likelihood estimates, but they do enjoy the property of being "consistent", i. e. for large samples, they tend in probability toward the true unknown parameter β.

Since we have seen the importance of estimating the slope accurately and that the method of estimating it depends on the values of the (often unknown) variances in errors of measurement, then continuing knowledge of the precision of measurement of instruments, i. e. their capability in repeatability, reproducibility and also accuracy, becomes rather critical indeed. In fact, any worthwhile experiment could be planned and carried out more appropriately with such continuing knowledge of instrument precision capability, as this would lead to improved analyses and predictions for the data taken. Moreover, we now see from the above that the matter of trying to find even some linear relationship between true values of the variables studied can get a bit complex or involved.

We have not exhausted in our account here the methods of estimating the slope, β. In fact, we should mention in passing that for the linear relation and error in both variates, grouping methods such as that of Wald-Bartlett (1949) might be used to advantage. Grouping methods were developed primarily for the case where the $\mu_i$ are random variables (discussed further later), but they may also be used for the case where they are varied systematically by the investigator over particular ranges of interest. The Wald-Bartlett method for estimating β involves dividing the data ordered in the x-direction into three approximately equal groups, computing the mean x's and y's of the two extreme groups, i. e. $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_3, \bar{y}_3)$, and estimating the slope β from

$$\hat{\beta} = (\bar{y}_3 - \bar{y}_1)/(\bar{x}_3 - \bar{x}_1). \quad (30)$$

(Of course, totals could be used in place of averages.) To illustrate for the measured energy versus striking energy fit, we will use the top five and bottom five points and compute for each point $m_R V_R^2/10^8$ and 27 $V_S^2/10^8$. This results in the following estimate of slope:

$$\hat{\beta} = \frac{(2.72+2.24+1.60+1.14+0.75)-(0.1288+0+0+0+0)}{(4.93+4.26+3.46+2.99+2.51)-(1.67+1.70+1.84+1.87+1.94)} = .91,$$

whereas from the linear least squares fit we obtained b = .88, indicating rather good agreement (although it does distribute the error to the independent variable, indicating the extreme sensitivity involved here!)

We will not discuss here the best methods of grouping and the various ramifications of the technique, but rather refer the reader to papers of Wald (1940), Bartlett (1949), Madansky (1959), and Neyman (1951).

For the case of error in both variables, we might finally mention an estimate of $\beta$ that seems intuitive on practical grounds. This involves finding the slope by least squares from the linear regression of the "dependent" variable x and averaging this with the reciprocal of the slope obtained by finding the regression of x and y, since both contain error. From the former, we have that $b_{y_x} = A_{xy}/A_{xx}$ and the latter that $b_{x_y} = A_{xy}/A_{yy}$. Using the preceding data, we obtain

$$b_{y_x} = 210.721/239.301 \qquad \text{and} \qquad b_{x_y} = \frac{210.721}{187.103}$$

$$= .8806 \qquad\qquad\qquad = 1.1262$$

so that
$$\hat{\beta} = (.8806 + 1/1.1262)/2 = .8843$$

Moran (1956) treats this type of estimate.

## IV. LINEAR LEAST SQUARES WITH BOTH VARIABLES SUBJECT TO ERROR, AND BOTH VARIABLES RANDOM

In this case, the model of formulae (21) and (22) still apply, but $\mu$ instead of being a controlled or fixed variable is now random. (There are some problems in the physical sciences or ballistics technology that fall into this category, but we believe the controlled variable case takes priority.) The errors, $d_i$ and $e_i$, are again considered to be normally distributed with zero means and variances $\sigma_d^2$ and $\sigma_e^2$ as before. It is easy to see that many of the formulas developed in Section III still apply to the case of $\mu$ being randomly distributed. In fact, (26), (27), (28) and (29) apply without alteration. Of course,it is very desirable for applications in the physical sciences that the variances in errors of measurement, $\sigma_d^2$ and $\sigma_e^2$, be small compared to the variance in $\mu$ or $\sigma_\mu^2$ to guarantee sufficient precision of measurement.

Although as mentioned we will not delve very deeply into this particular case - since the use of the controlled variable is widely practiced in the physical sciences, we will nevertheless establish a few principles of interest and record them here.

To begin with, if $\sigma_d^2$ and $\sigma_e^2$ are both known, then (29) becomes the maximum likelihood estimate of the slope $\beta$, for then equations (26) are the basic maximum likelihood estimates. We also see from (26) that if $\sigma_d^2$ and $\sigma_e^2$ are both known, then this case becomes an over-identified situation, since actually we need to know only the ratio $\lambda = \sigma_d^2/\sigma_e^2$. In fact, if the ratio $\lambda$ is known, then Madansky (1959) shows that the proper estimate of $\beta$ is given by

$$\hat{\beta} = \frac{A_{yy} - \lambda A_{xx} + [(A_{yy} - \lambda A_{xx})^2 + 4\lambda A_{xy}^2]^{1/2}}{2A_{xy}} \qquad (31)$$

This estimate of $\beta$ may be applied to the controlled independent variable case above also. For example, if we use the data for striking energy and measured residual energy above, and assume $\lambda$=1, we have

$$\hat{\beta} = \frac{187.103-239.301+[(187.103-239.301)^2+4(210.721)^2]^{1/2}}{2(210.721)}$$

$$= .884,$$

which is very nearly the same as the estimate from $(b_{y_x}+1/b_{x_y})/2$ above.

Madansky (1959) gives a rather detailed discussion of the case where the $\mu_i$ are random variables, including grouping methods for estimating $\beta$ et al.

For a case where the $\mu_i$ are random, and it is known that $\beta=1$, Grubbs (1948) gives methods of estimating $\sigma_d^2$ and $\sigma_e^2$ for two or more instruments which take simultaneous readings.

## V. NON-LINEAR REGRESSION OR GENERALIZED LEAST SQUARES WITH ERROR IN BOTH VARIABLES

We have covered only the problem of linear least squares or regression so far along with some account of its relation to the use of physical laws. Our purpose has been to indicate a rather compact approach through the use of the $A_{uv}$ type computations or functions in the analysis and to show that in practice it is usually or in many cases highly desirable to work with physical relations or parameters, if at all possible, since such models are more informative and are lasting. It is nevertheless clear that we cannot begin to cover such an involved and wide field of interest in any depth here. In fact, the important objective of finding the most appropriate use or combination of statistical methods with models or laws in the physical sciences represents a field or area of interest that is undergoing continual development. The best gains will likely result in bridging the gap between the science of statistics on one hand and the field of physical application on the other. Non-linear, generalized least squares with error in both variables is therefore a wide-open field which depends critically on particular applications. However, once it has been decided to fit a hypothesized or developed model for the particular problem at hand, the scope of our account here will have to be limited to that of referring readers to the work of Deming (1943), and especially more recent accomplishments such as those of Britt and Luecke (1973), Celmins (1973), Chandler (1972) and similar treatments, for example.* For the non-linear, generalized treatment, iteration to the desired fit is generally required, and the modern computer is almost a necessary or most convenient aid. It is for these reasons that our account is severely curtailed here. We believe it desirable, nevertheless, to cover in this report some examples of multiple linear regression, the parabola and the use of orthogonal polynomials for the case of equally spaced values of the independent variable.

* The reader might also consult the recent paper of Gallant (1975), "Nonlinear Regression," The American Statistician, Vol. 29, No. 2, 73 - 81.

## VI. THE PLANE-ONE VARIABLE z (THE DEPENDENT VARIABLE) SUBJECT TO ERROR

In this case, we seek the relation between a dependent variable (subject to error) and two independent variables, x and y, which are relatively free of error (or we seek the regression of z on x and y) by the method of least-squares. Also, from the physical standpoint, we are very much interested in whether the fitted plane is unbiased, i. e. can be regarded as representing the functional or structural relation between the true values of z, and x and y. We will assume that the measured values of x and y are both "free of error", whereas the observed values of z are subject to a (random) error of measurement. Thus, the functional relation may be represented by

$$z = \alpha + \beta x + \gamma y \tag{32}$$

The model or assumption considered for the observed values $(x_i, y_i, z_i)$ is

$x_i$ (a variahle, free of error)

$y_i$ (a variable, free of error)

$z_i = [\alpha + \beta x_i + \gamma y_i]$, (subject to error $e_i \approx N(0, \sigma_e^2)$)

We propose to fit the equation

$$z = a + bx + cy \tag{33}$$

to the observed data by determining a, b and c (which will be estimates of $\alpha$, $\beta$ & $\gamma$) by the method of Least-Squares i. e. such that the sums of squares of the deviations (observed minus fitted values) is a minimum. We have

$$\phi = \sum_{i=1}^{n} (z_i - a - bx_i - cy_i)^2, \text{ to be a minimum.} \tag{34}$$

Note that $\bar{z} = a + b\bar{x} + c\bar{y}$. Hence, since the $A_{uv}$ are not origin dependent and to simplify the agebra, we make this substitution in $\phi$, obtaining

$\phi = \sum_{i=1}^{n} \{(z_i - \bar{z}) - b\,(x_i - \bar{x}) - c\,(y_i - \bar{y})\}^2$ which is to be a minimum.

$$\frac{\partial\phi}{\partial b} = -\,2\Sigma(x_i - \bar{x})\,\{(z_i - \bar{z}) - b(x_i - \bar{x}) - c\,(y_i - \bar{y})\} = 0$$

$$\frac{\partial\phi}{\partial c} = -\,2\Sigma(y_i - \bar{y})\,\{(z_i - \bar{z}) - b\,(x_i - \bar{x}) - c\,(y_i - \bar{y})\} = 0$$

Solving for b, c and a, we get

$$b = \frac{A_{xz}\,A_{yy} - A_{yz}\,A_{xy}}{A_{xx}A_{yy} - A_{xy}^2} \qquad (35)$$

$$c = \frac{A_{xx}\,A_{yz} - A_{xy}\,A_{xz}}{A_{xx}A_{yy} - A_{xy}^2} \qquad (36)$$

$$a = \bar{z} - b\bar{x} - c\bar{y} = \frac{1}{n}\{\Sigma z_i - b\Sigma x_i - c\Sigma y_i\} \qquad (37)$$

The variance of residuals will be given by

$$S^2 = \frac{1}{n-3} \sum_{i=1}^{n} \{(z_i - \bar{z}) - b\,(x_i - \bar{x}) - c(y_i - \bar{y})\}^2$$

or

$$\text{Est of } \sigma_e^2 = S^2 = \frac{1}{n(n-3)}\,\{A_{zz} - b\,A_{xz} - c\,A_{yz}\} \qquad (38)$$

Under the assumptions (32), it can be shown that the mean or expected values of a, b and c are respectively $\alpha$, $\beta$ and $\gamma$. Hence, for the model assumed,the method of Least-Squares gives an unbiased estimate (with minimum variance) of the functional or structural relation between the true values of z and the (fixed - i.e. "free of error") variates x and y if (32) is correct.

Also, by methods indicated above for the line, it can be shown that

$$\text{Est of } \sigma_a^2 = \frac{n\ S^2\ [\Sigma x_i^2 \Sigma y_i^2 - (\Sigma x_i y_i)^2]}{A_{xx}A_{yy} - A_{xy}^2}\ , \qquad (39)$$

$$\text{Est of } \sigma_b^2 = \frac{n\ A_{yy}\ S^2}{A_{xx}\ A_{yy} - A_{xy}^2}\ , \qquad (40)$$

and

$$\text{Est of } \sigma_c^2 = \frac{n\ A_{xx}\ S^2}{A_{xx}\ A_{yy} - A_{xy}^2}\ . \qquad (41)$$

We now have all the information required to carry out the usual "Student's" t-tests to judge the hypotheses concerning whether the true parameters $\alpha$, $\beta$ and $\gamma$ can be regarded as being equal to zero or any selected constant values.

For example, to test whether the true slope $\beta$, in the functional or structural relation $z = \alpha + \beta x + \gamma y$, is equal to zero, we use "Student's" t-test

$$t = \frac{b-0}{\hat{\sigma}_b} = \frac{b\sqrt{A_{xx}\ A_{yy} - A_{xy}^2}}{S\sqrt{n\ A_{yy}}}$$

with n-3 degrees of freedom.

<u>Example.</u> The following data give the Ballistic Limits (BL) for various thicknesses and Brinell Hardness Numbers (BHN) of armor plate when tested with caliber .50 AP bullets. (The plates of armor were placed at an angle of obliquity of 42° from the line of fire). It is desired to find the linear regression equation of the Ballistic Limit (z) on the thickness (x) and BHN (y).

| z<br>BL<br>(f.s.) | x<br>Thickness<br>(in) | y<br>BHN |
|---|---|---|
| 927 | .253 | 317 |
| 978 | .258 | 321 |
| 1028 | .259 | 341 |
| 906 | .247 | 350 |
| 1159 | .256 | 352 |
| 1055 | .246 | 363 |
| 1335 | .257 | 365 |
| 1392 | .262 | 375 |
| 1362 | .255 | 373 |
| 1374 | .258 | 391 |
| 1393 | .253 | 407 |
| 1401 | .252 | 426 |
| 1436 | .246 | 432 |
| 1327 | .250 | 469 |
| 950 | .242 | 275 |
| 998 | .243 | 302 |
| 1144 | .239 | 331 |
| 1080 | .242 | 355 |
| 1276 | .244 | 385 |
| 1062 | .234 | 426 |

We have

$N = 20$

$A_{xx} = .022304$ $\qquad A_{xy} = 2.824$ $\qquad \bar{x} = .2498$

$A_{yy} = 882,664$ $\qquad A_{xz} = 184.392$ $\qquad \bar{y} = 367.8$

$A_{zz} = 13,211,771$ $\qquad A_{yz} = 2,439,192$ $\qquad \bar{z} = 1179.15$

To determine the coefficients a, b, c in $z = a + bx + cy$, we have from (35), (36) and (37) that

$$b = \frac{A_{xz}A_{yy} - A_{yz}A_{xy}}{A_{xx}A_{yy} - A_{xy}^2} = \frac{155867902.08}{19678.96288} = 7920.534$$

$$c = \frac{A_{xx}A_{yz} - A_{xy}A_{xz}}{A_{xx}A_{yy} - A_{xy}^2} = \frac{53883.0154}{19678.963} = 2.738102,$$

and $a = \bar{z} - b\bar{x} - c\bar{y} = - 1806.473.$

The tentative regression equation is

B.L. = -1806.473 + 7920.534 (Thickness) + 2.738 (BHN). The variance of residuals is

$$S^2 = \frac{1}{n(n-3)}\{A_{zz} - b\,A_{xz} - c\,A_{yz}\} = \frac{5,072,531.401}{340} =$$

14,919.21,

and $nS^2 = 298,384.2$. Then

$$\hat{\sigma}_c^2 = \frac{nS^2A_{xx}}{A_{xx}A_{yy} - A_{xy}^2} = .33819, \qquad \hat{\sigma}_c = .58154,$$

$$\hat{\sigma}_b^2 = \frac{n\,S^2\,A_{yy}}{A_{xx}A_{yy} - A_{xy}^2} = 13,383,479.17, \quad \hat{\sigma}_b = 3658.344,$$

and

$$\hat{\sigma}_a^2 = \frac{n\,S^2\,\{\Sigma x^2\Sigma y^2 - (\Sigma\,xy)^2\}}{A_{xx}A_{yy} - A_{xy}^2} = 873,756.2, \qquad \hat{\sigma}_a = 934.749.$$

Moreover,

$$t_a = \frac{a}{\sigma_a} = \frac{-1806.473}{934.749} = -1.933,$$

$$t_b = \frac{b}{\sigma_b} = \frac{7920.534}{3658.343} = 2.165,$$

and $t_c = \frac{c}{\sigma_c} = \frac{2.738102}{.58154} = 4.708.$

Now since $t_{.05} = 2.11$ for $\nu = 17$ d. f., the slope b is significantly different from zero at the 5% level. The coefficient of BHN is highly significant (P<.005). We would thus adopt the equation given above for predicting Ballistic Limit from thickness and BHN under conditions similar to those of the test carried out. (In this particular case,the thicknesses appear to vary randomly in character, as do the Brinell hardness numbers to some extent. If the thicknesses had varied over a wide range, the slope b might have been highly significant.)

The variance of a value of z predicted from (33) is given by

$$\sigma_z^2 = \frac{\sigma_e^2}{n} + (x - \bar{x})^2 \sigma_b^2 + (y - \bar{y})^2 \sigma_c^2 + 2(x - \bar{x})(y - \bar{y}) \sigma_{bc}. \tag{42}$$

Estimates of $\sigma_e^2$, $\sigma_b^2$ and $\sigma_c^2$ are given by (39), (40) and (41), whereas an estimate of $\sigma_{bc}$ is given by

$$\text{Est of } \sigma_{bc} = \frac{-n\, A_{xy}\, S^2}{A_{xx}A_{yy} - A_{xy}^2} \tag{43}$$

## VII. THE PARABOLA - ONE VARIABLE z (THE DEPENDENT VARIABLE) SUBJECT TO ERROR

Here, we desire to fit a second degree curve or parabola to the observed data, i.e. we assume that the functional relation between the dependent variable z and the independent variable x is of the form

$$z = \alpha + \beta x + \gamma x^2 \tag{44}$$

Again, we postulate that the independent variable x is "free of error", whereas the dependent variable z is measured or obtained with error. Thus, the model considered for the observed values $x_i$, $z_i$ is

$$x_i = u_i \text{ (free of error)}$$

$$z_i = \alpha + \beta x_i + \gamma x_i^2 + e_i \text{ (contains error)} \tag{45}$$

We will fit the equation

$$z = a+bx+cx^2 \tag{46}$$

to the observed data by determining a, b and c (which will be estimates of α, β and γ respectively) in such a way that the sums of squares of the deviations of the observed values from the fitted values will be a minimum i.e. by the method of Least Squares. Actually, we do not have to go through the procedure of finding a, b and c so that

$d = \sum_{i=1}^{n} (z_i-a-bx-cx^2)^2$ is a minimum, since the method of

Least Squares is very general and we can, as a matter of fact, replace y in equation (33) for the plane by $x^2$. We thus have in a straight-forward manner that

$$b = \frac{A_{xz}A_{x^2x^2} - A_{x^2z}A_{xx^2}}{A_{xx}A_{x^2x^2} - A^2_{xx^2}} \tag{47}$$

$$c = \frac{A_{xx}A_{x^2z} - A_{xx^2}A_{xz}}{A_{xx}A_{x^2x^2} - A^2_{xx^2}} \tag{48}$$

$$a = \bar{z} - b\bar{x} - c\,\overline{x^2} = \frac{1}{n}\{\Sigma z_i - b\Sigma x_i - c\Sigma x_i^2\} \tag{49}$$

$$S^2 = \text{Est of } \sigma_e^2 = \frac{1}{n(n-3)}\{A_{zz} - bA_{xz} - cA_{x^2z}\} \tag{50}$$

$$\text{Est of } \sigma_a^2 = \frac{n\ S^2\ \{\Sigma x^2 \Sigma x^4 - (\Sigma x^3)^2\}}{A_{xx}\ A_{x^2x^2} - A^2_{xx^2}} \qquad (51)$$

$$\text{Est of } \sigma_b^2 = \frac{n\ S^2 A_{x^2x^2}}{A_{xx}A_{x^2x^2} - A^2_{xx^2}} \qquad (52)$$

$$\text{Est of } \sigma_c^2 = \frac{n\ S^2 A_{xx}}{A_{xx}A_{x^2x^2} - A^2_{xx^2}} \qquad (53)$$

The variance of a value of z predicted from the equation (46) is given by

$$\sigma_z^2 = \frac{\sigma_e^2}{n} + (x-\bar{x})^2\ \sigma_b^2 + (x^2-\overline{x^2})^2\ \sigma_c^2 + 2(x-\bar{x})(x^2-\overline{x^2})\ \sigma_{bc} \qquad (54)$$

Estimates of $\sigma_e^2$, $\sigma_b^2$ and $\sigma_c^2$ are given by (50), (52) and (53), whereas an estimate of $\sigma_{bc}$ is given by

$$\text{Est of } \sigma_{bc} = \frac{-n\ A_{xx^2}S^2}{A_{xx}A_{x^2x^2} - A^2_{xx^2}} \qquad (55)$$

Example. A test was conducted* to determine the effect of barrel length on muzzle velocity for a caliber .22 Long Rifle (Model 37 Remington). The observed data are given below and each average MV is based on 10 rounds.

| Barrel Length (in.)<br>x | Ave. Vel. (f/s)<br>z |
|---|---|
| 28 | 1084 |
| 26 | 1075 |
| 24 | 1091 |
| 22 | 1096 |
| 20 | 1100 |
| 18 | 1098 |
| 16 | 1085 |
| 14 | 1088 |
| 12 | 1085 |
| 10 | 1079 |
| 8 | 1067 |
| 6 | 1040 |

*by W. O. L. F. Moore - See APG Firing Record Misc. 017

We have

$n = 12$

$\Sigma x = 204$

$\Sigma z = 12988$

$\Sigma x^2 = 4040$

$\Sigma x^3 = 88,128$

$\Sigma x^4 = 2,042,720$

$\Sigma xz = 221,540$

$\Sigma x^2 z = 4,392,064$

$\Sigma z^2 = 14,060,306$

$A_{xx} = 6864$

$A_{zz} = 35528$

$A_{xz} = 8928$

$A_{x^2 x^2} = 8,191,040$

$A_{x^2 z} = 233,248$

$A_{xx^2} = 233,376$

Using formulae (47) - (53), we find

$$b = 10.6286, \quad c = -.27435, \quad a = 994.0115, \quad S^2 = 42.8464$$

$$\hat{\sigma}_b = 1.547, \quad \hat{\sigma}_c = .0448, \quad \hat{\sigma}_a = 11.920$$

Herrce,

$$t_b = \frac{b}{\sigma_b} = 6.87 \quad (p < .01)$$

$$t_c = 6.12 \quad (p < .01)$$

$$t_a = 83.39 \quad (p < .01)$$

Since a,b and c are highly significant values statistically, we adopt the equation

$$MV = 994.01 + 10.629\ (BL) - .2744\ (BL)^2,$$

where

MV = Muzzle Velocity and BL = Barrel Length (in.)

Transformation of Original Data. In view of the fact that it may be desirable to make linear transformations on the original variables (in order to reduce effectively the size of numbers in the calculations), the pertinent formulae given below may prove of value. Suppose we change the original variables x and z as follows:

$$u_i = c(x_i - h) \qquad v_i = d(z_i - k) \tag{56}$$

where c, d, h and k are constants. Then, it can be shown that

$$A_{xx^2} = \frac{1}{c^3} A_{uu^2} + \frac{2h}{c^2} A_{uu} \tag{57}$$

$$A_{x^2x^2} = \frac{1}{c^4} A_{u^2u^2} + \frac{4h}{c^3} A_{uu^2} + \frac{4h^2}{c^2} A_{uu} \tag{58}$$

$$A_{x^2z} = \frac{1}{c^2d} A_{u^2v} + \frac{2h}{cd} A_{uv} \tag{59}$$

We had previously shown that

$$A_{xx} = \frac{1}{c^2} A_{uu}$$

## VIII. THE REGRESSION OF A DEPENDENT VARIABLE (z) ON THREE INDEPENDENT VARIABLES (x, y and u)

Here we use the model

$$z_i = \alpha + \beta(x_i - \bar{x}) + \gamma(y_i - \bar{y}) + \delta(u_i - \bar{u}) + e_i , \tag{60}$$

and we will estimate z from the equation,

$$z = a + b(x - \bar{x}) + c(y - \bar{y}) + d(u - \bar{u}) = (a - b\bar{x} - c\bar{y} - d\bar{u}) + bx + cy + du, \tag{61}$$

where a, b, c and d are to be determined by the Method of Least Squares.

If we let

$$\Delta = \frac{1}{n^2}\begin{vmatrix} A_{xx} & A_{xy} & A_{xu} \\ A_{yx} & A_{yy} & A_{yu} \\ A_{ux} & A_{uy} & A_{uu} \end{vmatrix} = \frac{\Delta_1}{n^2}\text{, say,} \qquad (62)$$

then from the Method of Least Squares, we find straightforwardly that

$$a = \frac{1}{n}\Sigma z_i \quad \text{[The constant term of (61) is } \bar{z}-b\bar{x}-c\bar{y}-d\bar{u}\text{]} \qquad (63)$$

$$b = \frac{1}{\Delta_1}\begin{vmatrix} A_{xz} & A_{xy} & A_{xu} \\ A_{yz} & A_{yy} & A_{yu} \\ A_{uz} & A_{uy} & A_{uu} \end{vmatrix}, \qquad (64)$$

$$c = \frac{1}{\Delta_1}\begin{vmatrix} A_{xx} & A_{xz} & A_{xu} \\ A_{yx} & A_{yz} & A_{yu} \\ A_{ux} & A_{uz} & A_{uu} \end{vmatrix}, \qquad (65)$$

$$d = \frac{1}{\Delta_1}\begin{vmatrix} A_{xx} & A_{xy} & A_{xz} \\ A_{yx} & A_{yy} & A_{yz} \\ A_{ux} & A_{uy} & A_{uz} \end{vmatrix}. \qquad (66)$$

$$S^2 = \frac{1}{n(n-4)}\{A_{zz} - b\,A_{xz} - c\,A_{yz} - d\,A_{uz}\}, \qquad (67)$$

$$\text{Est of } \sigma_a^2 = \frac{S^2}{n}, \qquad (68)$$

$$\text{Est of } \sigma_b^2 = \frac{n\,S^2\,(A_{yy}\,A_{uu} - A_{yu}^2)}{\Delta_1}, \qquad (69)$$

$$\text{Est of } \sigma_c^2 = \frac{n\,S^2\,(A_{xx}\,A_{uu} - A_{xu}^2)}{\Delta_1}, \qquad (70)$$

and

$$\text{Est of } \sigma_d^2 = \frac{n\,S^2\,(A_{xx}\,A_{yy} - A_{xy}^2)}{\Delta_1}, \qquad (71)$$

[Note: If we wanted to fit the cubic $z = a+b(x-\bar{x})+c(x^2-\overline{x^2})+d\cdot(x^3-\overline{x^3})$ we could simply replace $y_i$ and $u_i$ above by $x_i^2$ and $x_i^3$ respectively.]

## IX. FITTING OF POLYNOMIALS FOR THE CASE WHERE THE OBSERVED VALUES OF THE INDEPENDENT VARIABLE ARE AT EQUALLY SPACED INTERVALS

If we are interested in fitting a polynomial of the form

$$z = a_0 + a_1x + a_2x^2 + \dots + a_rx^r \qquad (72)$$

for the relation between the variables z and x, and the independent variable x is equally spaced, i. e.

$$x_i = e+(i-1)f;\ i=1,\ 2,\ \dots,\ n\ ,$$

then the computations for a Least-Square fit can be simplified considerably by the use of Orthogonal Polynomials. Following Fisher and Yates (1943) consider polynomials, defined as follows:

$$P_r\,(t_i) = b_0 + b_1t_i + b_2t_i^2 + \dots + b_rt_i^r \qquad (73)$$

where i=1, 2, ...,n represents the number of points, r the degree of the polynomial (r=0, 1, 2, ...) and the b's are constants to be determined. The variable $t_i$ will be a linear transformation or function of the observed values of the independent variables $x_i$ which are equally spaced (free of error). Polynomials of the form (73) will be called orthogonal if

$$\sum_{i=1}^{n} P_r\,(t_i) \cdot P_s\,(t_i) = 0 \quad \text{for } r \neq s. \qquad (74)$$

Our procedure will be to fit

$$z_i = A_oP_o(t_i) + A_1P_1(t_i) + A_2P_2(t_i)+\dots + A_rP_r(t_i) \tag{75}$$

by the method of Least Squares. Hence, we determine $A_o$, $A_1$, etc., so that

$$\phi = \sum_{i=1}^{n} \{z_i - A_o{}^{P}_{o}(t_i) - A_1P_1(t_i) - \dots - A_rP_r(t_i)\}^2 \tag{76}$$

is a minimum.

Differentiating (76) with respect to $A_o$, $A_1$, ... $A_r$ and setting the derivatives equal to zero, we find the Normal Equations:

$$A_o \sum_{i=1}^{n} P_o^2(t_i) + A_1 \sum_{i=1}^{n} P_o(t_i)P_1(t_i) + \dots$$

$$\dots A_r \sum_{i=1}^{n} P_o(t_i)P_r(t_i) = \sum_{i=1}^{n} P_o(t_i)\, z_i$$

$$A_o \sum_{i=1}^{n} P_o(t_i)P_1(t_i) + A_1 \sum_{i=1}^{n} P_1^2(t_i) + \dots \tag{77}$$

$$\dots + A_r \sum_{i=1}^{n} P_1(t_i)P_r(t_i) = \sum_{i=1}^{n} P_1(t_i)z_i$$

.

.

$$A_o \sum_{i=1}^{n} P_o(t_i)P_r(t_i) + A_1 \sum_{i=1}^{n} P_1(t_i)P_r(t_i) = \dots$$

$$\dots + A_r \sum_{i=1}^{n} P_r^2(t_i) = \sum_{i=1}^{n} P_r(t_i)\, z_i$$

Note that the cross-product terms not on the principal diagonal are of the type

$\sum_{i=1}^{n} P_r(t_i)\, P_s(t_i)$, where $r \neq s$.

But these are zero if the polynomials are orthogonal. Thus, for orthogonal polynomials we would have solutions for the A's immediately, i. e.

$$A_o = \frac{\sum_{i=1}^{n} P_o(t_i)z_i}{\sum_{i=1}^{n} P_o^2(t_i)},$$

$$A_1 = \frac{\sum_{i=1}^{n} P_1(t_i)z_i}{\sum_{i=1}^{n} P_1^2(t_i)}, \qquad (78)$$

$$\vdots$$

and

$$A_r = \frac{\sum_{i=1}^{n} P_r(t_i)\ z_i}{\sum_{i=1}^{n} P_r^2(t_i)}.$$

The problem then is to find the polynomials $P_r(t_i)$ which result in orthogonality. This can be done if we put $t_i = (x_i - \bar{x})/f$ (where f is width of interval between the observations, $x_i$), and choose the $P_r(t_i)$ as follows:

$$P_o(t_i) = 1 = \xi_o'\text{(in the attached tables)}$$

$$P_1(t_i) = \lambda_1 t_i = \xi_1',$$

$$P_2(t_i) = \lambda_2[t_i^2 - \frac{(n^2-1)}{12}] = \xi_2', \qquad (79)$$

$$P_3(t_i) = \lambda_3 [t_i^3 - \frac{(3n^2-7)}{20} t_i] = \xi_3',$$

$$P_4(t_i) = \lambda_4 [t_i^4 - \frac{3n^2-13}{14} t_i^2 + \frac{3(n^2-1)(n^2-9)}{560}] = \xi_4',$$

etc.

The $\lambda$'s are constants which depend on the number of points n and are chosen so that for values of $t_i$ (which are positive or negative integers or 0), the above polynomials in the brackets turn out to be whole numbers. The general recurrence formulae for the

$P_r(t_i)$ or $\xi_r'$ are

$$\xi_{r+1} = \xi_1 \xi_r - \frac{r^2(n^2-r^2)}{4(4r^2-1)} \xi_{r-1} \qquad r = 1, 2, \ldots \qquad (80)$$

where

$$\xi_r' = \lambda_r \xi_r .$$

Table XXIII, Orthogonal Polynomials, pp. 62 - 68, of the Fisher and Yates tables (1943) give the required values of the Orthogonal Polynomials $P_r(t)$ or $\xi_r'$ for $r = 1, 2, \ldots, 5$ (i. e., through the fifth degree) and for the number of points, n, up through n = 52. The values of $\xi_2'$ and $\xi_4'$ are symmetrical about their middle values, and the $\xi_1'$, $\xi_3$ and $\xi_5'$ are also symmetrical except that the first half of the sequences are the negatives of those in the last half. For this reason, only half of the values (i.e., the upper ones) are tabulated for $n \geq 9$. The first two rows under each table give values of the sum of the squares of $\xi_r'$ and the third or last row just below each table gives values of the $\lambda_r$.

It can be shown that an ordinary polynomial

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k$$

can always be expressed in terms of orthogonal polynomials for any specified set of values of x. For example,

$$y = -35 + 59x - 21x^2 + 2x^3, \text{when } x=1, 2, 3, \ldots 7$$

can be written in the form

$$y = 5 + (-4+x) + 3\,(12-8x+x^2) + 12(-6+\frac{41}{6}x-2x^2 + \frac{1}{6}x^3),$$

where the polynomials in parentheses are orthogonal, as seen from the table below:

| x | $P_1$ | $P_2$ | $P_3$ | $P_1P_2$ | $P_1P_3$ | $P_2P_3$ |
|---|---|---|---|---|---|---|
|  | $(-4+x)$ | $(12-8x+x^2)$ | $(-6+\frac{41}{6}x$<br>$-2x^2+\frac{1}{6}x^3)$ |  |  |  |
| 1 | -3 | 5 | -1 | -15 | 3 | -5 |
| 2 | -2 | 0 | 1 | 0 | -2 | 0 |
| 3 | -1 | -3 | 1 | 3 | -1 | -3 |
| 4 | 0 | -4 | 0 | 0 | 0 | 0 |
| 5 | 1 | -3 | -1 | -3 | -1 | 3 |
| 6 | 2 | 0 | -1 | 0 | -2 | 0 |
| 7 | 3 | 5 | 1 | 15 | 3 | 5 |
| Total | 0 | 0 | 0 | 0 | 0 | 0 |

The above exhibits the required properties of the orthogonal polynomials.

Example

Using the data of Section VII for length of barrel of the caliber .22 Long Rifle versus average muzzle velocity,we arrange the computations as in the following table,where the values of $\xi_r'$ are taken from Fisher and Yates Tables (1943), Page 62.

| Barrel Length (in.) | Sum of Velocities (f/s) $s_i$ | Diff. of Vel. (f/s) $d_i$ | $\xi_1^1$ | $\xi_2^1$ | $\xi_3^1$ |
|---|---|---|---|---|---|
| 18, 16 | 2183 | 13 | 1 | -35 | -7 |
| 20, 14 | 2188 | 12 | 3 | -29 | -19 |
| 22, 12 | 2181 | 11 | 5 | -17 | -25 |
| 24, 10 | 2170 | 12 | 7 | 1 | -21 |
| 26, 8 | 2142 | 8 | 9 | 25 | - 3 |
| 28, 6 | 2124 | 44 | 11 | 55 | 33 |

$t_i = \xi_1 = (x_i-\bar{x})/f = (x_i - 17)/2$ ; $\xi_1^1 = \lambda_1 t_i = 2t_i$,

$\xi_2^1 = \lambda_2[t_i^2 - (n^2-1)/12] = 3[t_i^2-143/12]$,

etc., as in (79).

Mean Velocity, $\bar{z} = (2183 + 2188 + 2181 + 2170 + 2142 + 2124)/12$

$= 1082.33$

$z = a + b\xi_1^1 + c\xi_2^1 + d\xi_3^1$, where $a = \bar{z} = 1082.33$

$b = \Sigma\xi_{1i}^1 d_i/572 = 744/572 = 1.3007$

$c = \Sigma\xi_{2i}^1 s_i/12012 = -4394/12012 = -.3658$

$d = \Sigma\xi_{3i}^1 d_i/5148 = 582/5148 = .1131$

The analysis of variance table is as follows:

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | F Ratio |
|---|---|---|---|---|
| Linear Regression | 1 | 967.72 | 967.72 | 24.21 |
| Quadratic Regression | 1 | 1,607.33 | 1,607.33 | 40.20 |
| Cubic Regression | 1 | 65.80 | 65.80 | 1.65 |
| Residual Error | 8 | 319.82 | 39.98 | |
| Total | 11 | 2,960.67 | | |

Note that the sum of squares are found simply from

$$(744)^2/572 = 967.72,\ (-4394)^2/12012 = 1607.33, \text{ etc.}$$

Since quadratic regression is highly significant but cubic regression is not, we fit the quadratic, which in terms of the original $x_i$ is

$$z = 1082.33 + 1.3007(2)(x-17)/2 - .3658(3) \cdot$$

$$[(x-17)^2/4 - (143/12)]$$

$$= 994 + 10.63x - .2744x^2 \text{ as before.}$$

The advantageous use of orthogonal polynomials in least squares curve fitting for numerous applied problems is thus clearly seen, especially along with significance tests for the coefficients.

# REFERENCES

1. ASTM Manual on Fitting Straight Lines (1962), (ASTM Special Technical Publication No. 313.) Published by the American Society for Testing and Materials, 1916 Race Street, Philadelphia, Penna. 19103.

2. Bartlett, M. S., (1949), "Fitting a Straight Line When Both Variables are Subject to Error," *Biometrics, Vol. 5,207-12.*

3. Berkson, J: (1950), "Are There Two Regressions?", *Journal of the American Statistical Association, Vol. 45, 164-80.*

4. Britt, H. I, & R. H. Luecke (1973), "The Estimation of Parameters in Non Linear, Implicit Models," *Technometrics, Vol. 15, No. 2, 233-47.*

5. Chandler, J. P. (1972), "On an Iterative Procedure for Estimating Functions When Both Variables are Subject to Error," *Technometrics, Vol. 14, No. 1, 71-76.*

6. Deming, W. E. (1943), "Statistical Adjustment of Data," John Wiley & Sons, New York.

7. Celmins, A. (1973), "Least Squares Adjustment with Finite Residuals for Non-Linear Constraints and Partially Correlated Data," *U. S. Army Ballistic Research Laboratories Report No. 1658, Aberdeen Proving Ground, Maryland 21005.* (AD #766283)

8. Fisher, R. A. & F, Yates, (1943) "Statistical Tables for Biological, Agricultural and Medical Research", Oliver and Boyd, London.

9. Grubbs, Frank E. (1948),"On Estimating Precision of Measuring Instruments and Product Variability", *Journal of the American Statistical Association, Vol. 43, 243-264.*

10. Madansky, A. (1959),"The Fitting of Straight Lines When Both Variables are Subject to Error,"*Journal of the American Statistical Association, Vol. 54, 173-205.*

11. Moran, P. A. P. (1956), "A Test of Significance for an Identifiable Relation," *Journal of the Royal Statistical Society Series B, Vol. 18, 62-64.*

12. Neyman, J. (1951), "Existence of Consistent Estimate of the Directional Parameter in a Linear Structural Relation Between Two Variables," *Annals of Mathematical Statistics,Vol 22, 496-512.*

# REFERENCES

13. Scheffe', H. (1961), "The Analysis of Variance," John Wiley and Sons, New York, Section 3.5.

14. Wald, A. (1940), "Fitting of Straight Lines If Both Variables Are Subject to Error," *The Annals of Mathematical Statistics, Vol. 11, 284-300.*

## DISTRIBUTION LIST

No. of Copies Organization

12 Commander
Defense Documentation Center
ATTN: DDC-TCA
Cameron Station
Alexandria, VA 22314

1 Commander
US Army Materiel Command
ATTN: AMCDMA-ST
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commander
US Army Materiel Command
ATTN: AMCRD-T
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commander
US Army Materiel Command
ATTN: AMCRD-R
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commander
US Army Aviation Systems
Command
12th and Spruce Streets
St. Louis, MO 63166

1 Director
US Army Air Mobility Research
and Development Laboratory
Ames Research Center
Moffett Field, CA 94035

1 Commander
US Army Electronics Command
ATTN: AMSEL-RD
Fort Monmouth, NJ 07703

1 Commander
US Army Missile Command
ATTN: AMSMI-R
Redstone Arsenal, AL 35809

No. of Copies Organization

1 Commander
US Army Tank Automotive Command
ATTN: AMSTA-RHFL
Warren, MI 48090

2 Commander
US Army Mobility Equipment
Research & Development Center
ATTN: Tech Docu Cen, Bldg 315
AMSME-RZT
Fort Belvoir, VA 22060

1 Commander
US Army Armament Command
Rock Island, IL 61202

1 Commander
US Army Frankford Arsenal
ATTN: SARFA-J7200, C.M. Dickey
Philadelphia, PA 19137

1 Commander
US Army Harry Diamond Labs
ATTN: AMXDO-TI
2800 Powder Mill Road
Adelphi, MD 20783

1 Commander
US Army Natick Laboratories
ATTN: Operations Research and
Analysis Office
Natick, MA 01762

1 Director
US Army TRADOC Systems Analysis
Activity
ATTN: ATAA-SA
White Sands Missile Range
New Mexico 88002

1 Commander
US Army Armor and Engineer Board
Fort Knox, KY 40121

## DISTRIBUTION LIST

No. of Copies — Organization

2 Commandant
US Army Logistics Center
Fort Lee, VA 23801

1 Commandant
US Army Artillery and
Missile School
Fort Sill, OK 73504

1 Commandant
US Army Air Defense School
ATTN: USADS-ATSAD-D-S
Fort Bliss, TX 79916

1 Commander
US Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

1 US Military Academy
Dept of Engineering
Mahan Hall
West Point, NY 10996

1 Chief, US Army Strategy and
Tactics Analysis Group
8120 Woodmont Avenue
Bethesda, MD 20014

3 Commander
US Naval Ordnance Systems
Command
ATTN: ORD-9132
Washington, DC 20360

1 Superintendent
US Naval Postgraduate School
Monterey, CA 93940

1 AFSC (SDW, R. Hartmeyer)
Andrews AFB
Washington, DC 20331

No. of Copies — Organization

1 ADTC (ADBPS-12)
Eglin AFB, FL 32542

1 SAC (DITW)
Offutt AFB, NB 68113

1 Director
Operations Analysis Office
Tactical Air Rec Center
Shaw Air Force Base, SC 29152

1 ASD (ASBEE-20)
Wright-Patterson AFB, OH 45433

1 California Institute of Tech
Department of Aeronautics
ATTN: Prof Homer J. Stewart
Pasadena, CA 91109

1 Prof & Coordinator for Army
Graduate Students
School of Industrial & System
Engineering
Georgia Institute of Technology
ATTN: Dr. L. C. Callahan
Atlanta, GA 30332

1 Johns Hopkins University
Dept of Operations Research
and Industrial Engineering
ATTN: Dr. A. J. Duncan
Committee E-11, ASTM
Baltimore, MD 21218

1 Princeton University
Dept of Aeronautical Engineering
ATTN: Prof. Martin Summerfield
Princeton, NJ 08540

1 Southwest Research Institute
ATTN: LTG Austin W. Betts
USA (Retired)
P. O. Drawer 28510
San Antonio, TX 78284

## DISTRIBUTION LIST

No. of Copies — Organization

1 University of California
Department of Chemistry
ATTN: Prof Joseph E. Mayer
School of Science & Engineering
La Jolla, CA 93038

1 University of Illinois
College of Engineering
ATTN: Dr. Daniel C. Drucker
Urbana-Champaign Campus
Urbana, IL 61801

1 University of Pennsylvania
The Moore School of
Electrical Engineering
ATTN: Prof Morris Rubinoff
Philadelphia, PA 19104

No. of Copies — Organization

1 LT General Leslie E. Simon
USA (Retired)
1761 Pine Tree Road
Winter Park, FL 32789

**Aberdeen Proving Ground**

Marine Corps Ln Ofc
Cdr, USAEA
ATTN: Biophysics Lab
Dir, USAMSAA