NATIONAL BUREAU OF STANDARDS REPORT

5699

Draft of
Part I, Section 3.1 (Linear Regression)
for

MANUAL ON EXPERIMENTAL STATISTICS
FOR ORDNANCE ENGINEERS

A Report to

OFFICE OF ORDNANCE RESEARCH
DEPARTMENT OF THE ARMY

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

THE NATIONAL BUREAU OF STANDARDS

Functions and Activities

The functions of the National Bureau of Standards are set forth in the Act of Congress, March 3, 1901, as amended by Congress in Public Law 619, 1950. These include the development and maintenance of the national standards of measurement and the provision of means and methods for making measurements consistent with these standards; the determination of physical constants and properties of materials; the development of methods and instruments for testing materials, devices, and structures; advisory services to Government Agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; and the development of standard practices, codes, and specifications. The work includes basic and applied research, development, engineering, instrumentation, testing, evaluation, calibration services, and various consultation and information services. A major portion of the Bureau's work is performed for other Government Agencies, particularly the Department of Defense and the Atomic Energy Commission. The scope of activities is suggested by the listing of divisions and sections on the inside of the back cover.

Reports and Publications

The results of the Bureau's work take the form of either actual equipment and devices or published papers and reports. Reports are issued to the sponsoring agency of a particular project or program. Published papers appear either in the Bureau's own series of publications or in the journals of professional and scientific societies. The Bureau itself publishes three monthly periodicals, available from the Government Printing Office: The Journal of Research, which presents complete papers reporting technical investigations; the Technical News Bulletin, which presents summary and preliminary reports on work in progress; and Basic Radio Propagation Predictions, which provides data for determining the best frequencies to use for radio communications throughout the world. There are also five series of nonperiodical publications: The Applied Mathematics Series, Circulars, Handbooks, Building Materials and Structures Reports, and Miscellaneous Publications.

Information on the Bureau's publications can be found in NBS Circular 460, Publications of the National Bureau of Standards ($1.25) and its Supplement ($0.75), available from the Superintendent of Documents, Government Printing Office, Washington 25, D. C.

Inquiries regarding the Bureau's reports should be addressed to the Office of Technical Information, National Bureau of Standards, Washington 25, D. C.

NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT		NBS REPORT
1103-40-5146	26 December 1957	5699

Draft of
Part I, Section 3.1 (Linear Regression)
for

MANUAL ON EXPERIMENTAL STATISTICS
FOR ORDNANCE ENGINEERS

Prepared By

Statistical Engineering Laboratory

A Report
to

OFFICE OF ORDNANCE RESEARCH
DEPARTMENT OF THE ARMY

U. S. DEPARTMENT OF COMMERCE

NATIONAL BUREAU OF STANDARDS

NOTICE

This report is a preliminary draft of Part I, section 3.1 (Linear Regression) for the Manual on Experimental Statistics for Ordnance Engineers.

At the time of printing, it has been noted that certain portions of the text should be revised in the final draft. No known inaccuracies exist in the present draft, but improvements in arrangement and exposition of some of the material may be made at a later time.

Certain figures referred to in the text have not been included, because the examples to which they apply will be replaced by other examples.

Table of Contents

Table of Contents (continued)

3. Description, Prediction and Correlation.

There are many situations in which we wish to know something about the relationships between two or more characteristics of a material (product or process). In some cases we may know or suspect from theoretical considerations that two properties are functionally related, and wish to know more about the structure of this relationship. In other cases, we may be interested in discovering whether there exists a degree of association between the two properties which could be used to our advantage. For example in specifying methods of test for a material, we might have 2 tests, both reflecting performance, but one of which is cheaper, simpler or quicker to run. If a high degree of association between the 2 tests is found, one might wish to run regularly only the simpler test.

In this section, we shall discuss the following situations in detail:

3.1) Linear relationships between two variables (fitting a straight line of the form $Y = \beta_0 + \beta_1 X$).

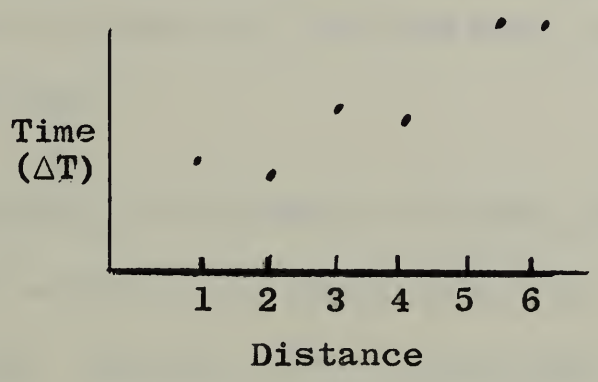
3.2) Relationships involving several variables.

3.3) Non linear relationships between two variables (fitting curves of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots$)

Where only two characteristics are involved, the natural first step in handling the experimental results is to plot the points on graph paper. Conventionally the "independent variable" X is plotted on the horizontal scale, and the "dependent variable" Y on the vertical scale.

There is no substitute for a plot of the data to give some idea of the general spread and shape of the results. Even though a line fitted by eye is usually not sufficient, a picture of the relationship is indispensable before we compute the line, and may sometimes save us from useless computing. When we are investigating an empirical association of two characteristics, a look at the plot will reveal whether such association is likely or whether we have only a patternless scatter of points. If we are investigating a structural relationship, the plotted data will show whether a hypothetical linear relationship is borne out or whether perhaps we must consider what grounds we have for fitting a curve of higher degree. In some cases, a plot will reveal unsuspected difficulties in our experimental set-up which must be ironed out before we fit any kind of line or do any kind of computing. An example of this occurred in measuring the time a drop of dye took in travelling between marked distances along a water channel. The channel was marked at equal distances and an observer

recorded the time at which it passed each marker. The device used for recording time consisted of two clocks hooked up so that when one was stopped, the other started, and therefore Clock 1 recorded times for Distance Markers 1, 3, 5 etc., and Clock 2 recorded for the even distances. When the data was plotted, it looked somewhat as follows:



It was obvious that there was a systematic difference between odd and even distances (presumably a lag in circuit between the two clocks). One could easily have fitted a straight line to the odd distances and a different one to the even distances, with approximately constant difference between the two lines. The effect was so persistent that the experimenter decided to find a better means of recording times before fitting any lines at all.

If no obvious difficulties are revealed by the plot, we fit the line according to the procedures given in 3.1.

Fitting by eye is usually inadequate for the following reasons:

i) no two people would fit exactly the same line, and therefore the procedure is not objective.

ii) we need to have some measure of how well the line does fit the data, and this is only to be obtained by a mathematical procedure of fitting.

3.1 Linear Relationships Between Two Variables

In this section, we shall deal only with linear relationships. However, by a transformation, one can frequently make the relationship a linear one. For example, if the relationship is believed to be in the form $Y = c_1 X^{c_2}$, then $\log Y = \log c_1 + c_2 \log X$. Putting $v = \log Y$, $c_1' = \log c_1$, $u = \log X$, we have $v = c_1' + c_2 u$.

Before we give the detailed procedure for fitting a straight line, we should first think about the kinds of physical situations which can be described by a linear relationship between two variables. The methods of description and prediction may be different depending upon the underlying system. In general we recognize two different and important systems, which we shall call: STATISTICAL and FUNCTIONAL. Before giving any formulas, we

will describe both situations so that we can distinguish between them in a practical case. It is not possible to make this distinction from looking at the data. Before fitting the line - indeed before taking the measurements - we must stop and think, using whatever knowledge we possess about the physical system from which the measurements came.

Functional Relationships FI and FII

In this kind of relationship, we believe that there exists an exact mathematical formula relating the two variables, and the only reason that our observations do not fit this formula exactly is because of disturbances or errors of measurement in one or both variables. We shall discuss this type separately for the two cases:

Case I - Errors of measurement all in one variable Y.

Case II - Both variables (X,Y) subject to error.

Common situations which may be described by Functional relationships include calibration lines, comparisons of analytical procedures, and relationships in which time is the X variable. Such an example is shown in Figure 6.3

in Bennett and Franklin "Statistical Analysis in Chemistry and the Chemical Industry" (page 219); and involves the determination of carbon content of 36 samples of bail clay by two different methods.

Mathematical Description of Functional Relationships

Let X,Y represent two characteristics, items, products, or processes. Assume we have n pairs of observations (X_1,Y_1), (X_2,Y_2),..., (X_n,Y_n).

Case I (FI)

There is an underlying mathematical (functional) relationship $Y = \beta_0 + \beta_1 X$. We are able to measure X relatively accurately. However, our measurements of Y for a given X follow a normal distribution with mean $\beta_0 + \beta_1 X$ and variance $\sigma^2_{Y.X}$ which is independent of the value of X.

Solution

This case may be handled computationally in exactly the same manner as SII, but both the underlying assumptions and the interpretation of the end results are different.

Case II (FII)

There is an underlying mathematical (functional) relationship $Y = \beta_0 + \beta_1 X$. However, we can not observe either X or Y accurately.

Solution

The full treatment of this case depends on the assumptions we are willing to make about error distributions. For complete discussion of the problem, see [ ] F. S. Acton, "Analyzing Straight Line Data", John Wiley and Sons, in press).

However, there is a simple method for fitting the line which is generally applicable. (Methods for getting confidence limits on the line too. These are less satisfactory).

This method of getting the line is quoted from [ ] M. S. Bartlett, "Fitting a Straight Line When Both Variables are Subject to Error", Biometrics, Volume 5, Number 3, September 1949. (Similar methods had been used previously by other authors).

a) For the location of the fitted straight line use as one point the mean coordinates $\overline{X},\overline{Y}$ just as in the least-squares method.

b) For the slope, first divide the n plotted points into 3 groups, the equal numbers k in the two extreme groups being chosen to be as near 1/3n as possible (The 3 groups are non-overlapping when

considered, say, in the X direction). The slope β_1 is estimated by:

$$b_1 = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1}$$

Another treatment is possible if we can assume that we know the ratio of the errors. Attempts to observe the true point (X,Y) would result in a cluster of points (X_i,Y_i) about (X,Y). We might write $X_i = X + u_i$, $Y_i = Y + v_i$, where u_i and v_i are the deviations of the observations (X_i,Y_i) about (X,Y). We shall assume that u_i and v_i are each normally distributed with mean zero and variances σ_u^2 and σ_v^2 respectively. We shall further assume that the ratio of the two variances

$$\frac{\sigma_v^2}{\sigma_u^2} = \lambda \quad ,$$

which is known. We give a solution in Problem 3.1.2.1.

Statistical Relationships - Case I (SI)

In this case, we have first selected a random sample of items from some population (material, product, process, people), and have measured two characteristics on each item.

A classic example of this type is the relationship between height and weight of men. Any observant person knows that weight tends to vary with height, but also that individuals of the same height may vary widely in weight. It is obvious that the errors made in measuring height or weight are very small as compared to this inherent variation between individuals. One would surely not be willing to predict the exact weight of one individual from his height, but one might

be willing to estimate the average weight of all individuals of a given height.

The height-weight example is given as a homely one which is universally familiar. Such examples do exist also in physical science, particularly in cases involving comparison of two test methods. In many cases we have two tests, which are related to each other in some complicated way and which are both related to the performance of the material. The relationship may be obscured by inherent variations among sample units (due to varying density for example). We would be very interested in knowing whether the relationship between the two is sufficient to predict from one test an average value for the other - particularly if one test is considerably simpler or cheaper than the other.

Such an example is given in Figure 0, the results of a study of synthetic rubber [J. M. Buist and O. L. Davies, reported in Statistical Methods in Research by O. L. Davies]. Thirty samples of rubber were taken, and for each, a determination of abrasion loss (grams/h.p. hour) and hardness (degrees shore) were taken. The data seems to indicate (and this is probably not surprising) that there is a relationship between abrasion loss and hardness. While one might hesitate to predict the exact abrasion loss of a sample of rubber having a given hardness, one would be much more confident in predicting the average abrasion loss of a large number of

samples, each having the same hardness. The relationship is statistical rather than mathematical or exact and the error of observation, or experimental error is a negligible factor. Descriptions and predictions are applicable "on the average", and we can also give measures of the departure of individuals from the average.

Let us imagine our sample of rubber increasing in size considerably and let it be represented by Figure 0.2. If from Figure 0.2, one now plots the average abrasion losses corresponding to several fixed hardnesses they fall along a line YOY'. Similarly if one plots the average hardnesses corresponding to several fixed abrasion losses, they form a line XOX'. Clearly then, if one wished to predict the average (or even an individual) abrasion loss corresponding to a given hardness, he would use the first line. Likewise, if he wished to predict the average hardness corresponding to a given abrasion loss, he would use the second line. Note well that this is the only case of linear relationship in which we can fit two lines, one for predicting Y from X and one for predicting X from Y, and the only case in which the well known correlation coefficient has any meaning. This case is often called the bivariate normal case.

We have chosen to represent hardness by X and abrasion loss by Y, but the choice here is arbitrary - there are actually two regression lines. If we do find an empirical

relationship between the two, we will ordinarily choose as X the variable which is easier to measure.

Mathematical Description of Statistical Relationships - Case I

Let us represent the hardnesses by X and the abrasion losses by Y. Then we usually make the assumption that for any fixed value of X, the corresponding values of Y form a normal distribution with means $\beta_0 + \beta_1 X$ (The lines YOY' and $Y = \beta_0 + \beta_1 X$ are the same) and variance $\sigma^2_{Y.X}$ (read "variance of Y given X") which is constant for all values of X. Similarly, we usually assume that for any fixed value of Y, the corresponding values of X form a normal distribution with mean $\beta'_0 + \beta'_1 Y$ (The lines XOX' and $\beta'_0 + \beta'_1 Y$ are the same) and variance $\sigma^2_{X.Y}$, (variance of X given Y) which is constant for all values of Y. (Taken together, these two sets of assumptions imply that X and Y are jointly distributed according to the bivariate normal distribution). In most practical situations we have only a sample from all the possible pairs of values X and Y, and therefore we cannot determine the "true" line exactly. We must estimate the equation of one or both of the lines $\beta_0 + \beta_1 X$ or $\beta'_0 + \beta'_1 Y$. We have a random sample of n pairs of values (X_1,Y_1), (X_2,Y_2) ..., $(X_n.Y_n)$ and wish to estimate the line which will enable us to make the "best" predictions of the values of Y, corresponding to given values of X

(or the line to make the "best" predictions of X, given Y). Our method of fitting the line gives us "best" predictions in the sense that: for a given $X=X'$, our estimate of the corresponding value of $Y=Y'$ will (1) on the average equal the mean value of Y for that X (i.e., it will be on the "true" line $Y = \beta_0 + \beta_1 X$) and (2) will have a smaller variance than had we used any other method for fitting the line.

Statistical Relationships, Case II (SII)

The general case described above (SI) is the most familiar example of a statistical relationship, but we need to consider also a case of statistical relationship (SII), which must be treated a bit differently. In SII, one of our variables, although a random variable in the population, is sampled only within a limited range (or at selected preassigned values) for purposes of our experiment. In the height-weight example, suppose that our group of men included only those whose heights were between 5'4" and 5'8". We will now be able to fit a line predicting weight from height, but certainly not height from weight, since we know nothing from our experiment about men whose height is outside our selected range. A correlation coefficient computed from such data is not a measure of the true correlation among height and weight in the (unrestricted) population.

The restriction of the range of X, when it is considered as the independent variable, does not spoil our estimates of Y when we fit the line $Y = b_0 + b_1X$. The restriction of the range of the dependent variable however (as in fitting the line $X = b_0' + b_1'Y$) gives a seriously distorted estimate of the true relationship. A good illustration of this point is to be found in [ ] C. Eisenhart, "The interpretation of certain regression methods ...", and the following paragraphs and the diagram are quoted from there:

"To illustrate this point it will be sufficient for our purposes to consider Figure 1 which has been constructed from some artificial data which are especially suited to this purpose. We shall suppose that Y is the dependent variable and X the independent variable, and that the complete array of points shown arose from a sampling process in which neither X nor Y was restricted. It will be noticed that the observational points lie in a band sloping upward to the right and that as X increases by one unit the distribution of the corresponding Y's moves up by one-half a unit. We may consider the points of the entire band shown as portraying the relationship between X and Y in the large, that is, when a point (X,Y) is selected at random without restrictions on either X or Y. The slanting line

labelled (I) indicates the "average" relationship prevailing between Y and X, that is, for a given value of X the arithmetic mean of the corresponding observed values of Y is given by the point on this line with abscissa X.

"Let us now consider the situation in which the points have been selected with restriction on X. As the results of such a procedure of selection let us take only those points between the two vertical lines drawn just to the right of X=3 and just to the left of X=7. It will be seen that this does not upset the average Y for a given value of X within the prescribed limits, i.e., $\overline{Y}_X$ is unaltered for $3 < X < 7$. In other words, the introduction of a restriction with regard to X, the independent variable, has not spoiled the inferences with regard to Y, when Y is considered as the dependent variable - that is, when we are arguing from X to Y."

"Consider now the effect of restricting the observed Y in a sampling process and then trying to infer about $\overline{Y}_X$ in the population at large from given values of X. In Figure 1 this corresponds to considering, say, only those points that lie between the horizontal lines just above Y=3 and just below Y=7. It is seen immediately that in this case, i.e., between the horizontal lines, for every value of X the average of the observed Y values is Y=5, and consequently the relation of Y to X is portrayed by the line numbered (II). It is seen that in

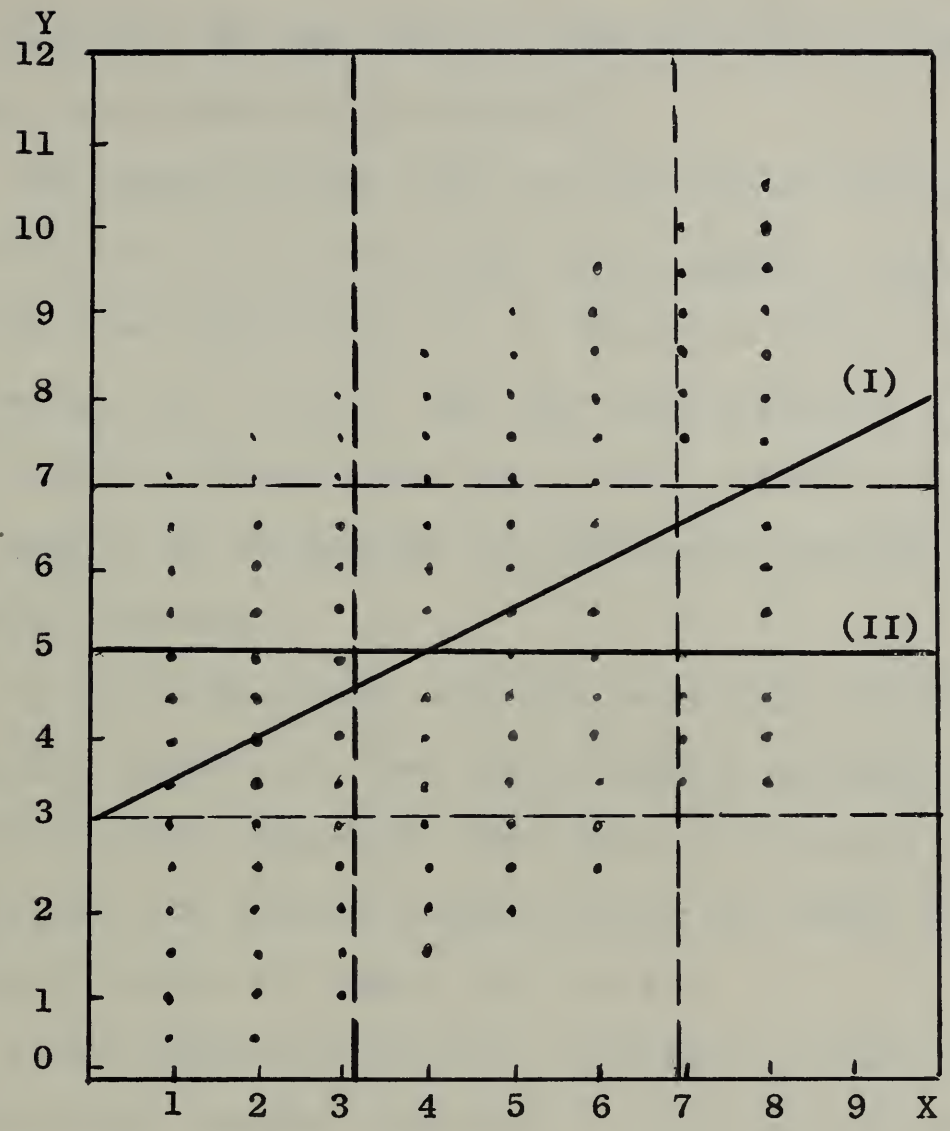


FIGURE 1

this case the "apparent" relation is not the correct one. Accordingly, we conclude that the restriction of the dependent variable is liable to seriously distort the relationship, so that what is observed is not representative of the true underlying situation.

The demonstration that we have chosen is simple and artifical but the conclusions drawn apply in general, namely, the restriction of X does not alter the regression of Y on X, but the restriction of Y does. For further illustrations and a very readable discussion see Chapter 20 of Methods of Correlation Analysis by Mordecai Ezekiel".

In any given case, consider carefully whether one is measuring samples as they come (thereby accepting the values of both properties that come with) which is SI, or whether one selects samples which are known to have a limited range of values of X (SII).

As an example of Case II, consider a study of watches to determine whether there was a relationship between the cost of a watch and its temperature coefficient. It was suggested that a correlation coefficient be computed. This was not possible because the watches had not been selected at random from the total watch production, but a deliberate effort had been made to get a group of low-priced, a group of medium priced, and a group of high-

priced watches.

Mathematical Description of Statistical Relationships (Case II)

We may write the conditions for Case II as follows:

Case II. For any given value of X, the corresponding values of Y have a normal distribution with mean $Y = \beta_0 + \beta_1 X$, and variance $\sigma^2_{Y.X}$ which is independent of the value of X. We have n pairs of values $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ in which X is the independent variable and Y is the dependent variable (i.e., the X values are selected, and the Y values thereby determined). We wish to describe the line which will enable us to make the "best" estimate of values of Y corresponding to given values of X (or vice versa).

We have already demonstrated that for Case I we require two lines, one for predicting Y from X and one for predicting X from Y. We shall now demonstrate why there is only one line in Case II.

For the sake of illustration, let us suppose that the points in Figure 0.3 represent all possible pairs (X,Y). Then YOY' and XOX' are the lines containing the mean

values of Y for each value of X, and the mean values of X for each value of Y, respectively. Suppose we select certain values of X, say the last four columns on the right. Now, for these columns, the same line YOY' contains the mean values of the Y's for given values of X. However, the mean X values for given values of Y are not on the line XOX'. Indeed, the mean X values will obviously depend on the values of X we observed and thus we may use only the line YOY', or estimates of it. Our "best" estimate of YOY' in the same sense as in Case I, is the same one we used in Case I for predicting Y from X.

Summary of 4 cases of Linear Relationships

	STATISTICAL (S)	
	SI	SII
Example	X = Height of Random Sample of Y = Weight Individuals	X = Height (preselected values of) Y = Weight of Individuals of Preselected Height.
Distinctive Features	X is not selected but "comes with" sample unit	X is measured beforehand; only selected values of X are used at which to measure Y
Errors of Measurement	[illegible] - Ordinarily negligible compared to variation between individuals	Same as in SI
Form of Line Fitted	$Y = \beta_0 + \beta_1 X$ $X = \beta_0' + \beta_1' X$	$Y = \beta_0 + \beta_1 X$
Procedure for Fitting	See section 3.1.3 and basic worksheet	See section 3.1.2 and basic worksheet
Correlation coefficient	$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$ See basic worksheet	Correlation exists but r cannot be computed from such an experiment

(Continued next page)

Summary of 4 cases of Linear Relationships
(Continued)

	FUNCTIONAL (F)	
Example and Distinctive Features	X and Y are related by a mathematical formula, which is not observed exactly because of disturbances or errors in one or both variables. Example: Determination of elastic constant of a spring which obeys Hooke's law. X = Known weight applied, Y = elongation.	
	FI	FII
Errors of Measurement	All measurement error in Y	X and Y both subject to error
Form of Line Fitted	$Y = \beta_0 + \beta_1 X$	See sections 3.1.2.
Procedure for Fitting	See section 3.1.1 and basic worksheet	Procedure depends on what assumptions can be made. See sections 3.1.2.
Correlation coefficient	Not applicable	Not applicable

Basic Worksheet for all types of Linear Relationships

X denotes _____ Y denotes _____

ΣX = _____ ΣY = _____

$\bar{X}$ = _____ $\bar{Y}$ = _____

Number of points: n = _____

ΣXY = _____

$(\Sigma X)(\Sigma Y)/n$ = _____

Σxy = _____

ΣX^2 = _____ ΣY^2 _____

$(\Sigma X)^2/n$ = _____ $(\Sigma Y)^2/n$ = _____

Σx^2 = _____ Σy^2 = _____

$b_1 = \frac{\Sigma xy}{\Sigma x^2}$ = _____ $\frac{(\Sigma xy)^2}{\Sigma x^2}$ = _____

$\bar{Y}$ = _____ $(n-2)s_Y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$ = ___

$b_1\bar{X}$ = _____ s_Y^2 = _____

$b_0 = Y - b_1\bar{X}$ = _____ s_Y = _____

Equation of the line:

$Y = b_0 + b_1X$

_____

s_{b_1} = _____

s_{b_0} = _____

Estimated variance of the slope:

$s_{b_1}^2 = \frac{s_Y^2}{\Sigma x^2}$ = _____

Estimated variance of intercept:

$s_{b_0}^2 = s_Y^2 \left\{\frac{1}{n} + \frac{\bar{X}^2}{\Sigma x^2}\right\}$ = ____

NOTES: (See next page)

NOTES for Basic Worksheet:

$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$

$\Sigma y^2 = \Sigma Y^2 - (\Sigma Y)^2/n$

$\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/n$

$(n-2)s_Y^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$

In computing ΣX^2, ΣY^2, ΣXY, carry all decimal places that are obtained using a computing machine - i.e., if data are recorded to 2 decimal places, carry 4 decimal places in squares. Carry this same number of decimal places in $\frac{(\Sigma X)^2}{n}$, $\frac{(\Sigma Y)^2}{n}$, $\frac{\Sigma X \Sigma Y}{n}$. Otherwise, one may lose too many significant figures in subtraction.

3.1.1. Problems for FI Relationships

[illegible]

Problem 3.1.1.1. What is the best line to be used for estimating Y from given values of X? (Caution - Extrapolation, i.e., use of the line for prediction outside the range of data from which the line was computed, is extremely hazardous).

Solution

i) Using Basic Worksheet, compute the line: $Y = b_0 + b_1X$. (This is an estimate of the true regression line $Y = \beta_0 + \beta_1X$).

ii) The above equation may be used:

a) to estimate a single value of Y(Y') corresponding to a single new value of X(X') or

b) to estimate the average value of Y associated with any X.

Either a) or b) gives the same numerical answer, since Y is obtained merely by substituting a given value of X in the equation of the fitted line. However, the limits of uncertainty on the estimated Y depend on whether one wishes to estimate a single Y value or an average value of Y.

In our situation we have n pairs of values from one experiment and we have fitted a regression line. If we wish to predict Y for a given X, we use that value of Y obtained from the fitted regression equation for the given X. The question then arises - how good a prediction of Y do we have? That is, if we could possibly repeat the experiment many times, each time obtaining n pairs of (X,Y) values, what range of Y values for that X would really be obtained?

The individual Y values (given X) for all the sets of data will spread over a larger range than will the collection consisting of $\bar{Y}$'s (given X). In some cases, we may be satisfied to say that the average of all Y (for given X) lies in a certain interval, and at other times we may need to know how large (or small) a single Y value we'd be likely to get. Our decision is similar to the case when we must decide, in writing a specification, whether we are satisfied to specify average strength, or whether we must specify minimum strength. The only difference here is that we are dealing with relationships between two variables, and therefore are talking about Y values for fixed X. A good example is given in [ ] Bowker and Lieberman (page 895).

"For example, tensile strength of cement is related to the curing time (t) by: strength = $\alpha e^{-\beta/t}$. This function is transformed by taking logarithms:

$$\ln \text{ strength} = \ln \alpha - \beta/t = \ln \alpha - \beta z$$

(where $z = 1/t$) which is now linear in z and can be estimated by the method of least squares. The cement manufacturer is interested in the average tensile strength of his cement after a particular period of time i.e., a confidence statement [for mean Y given X], whereas a builder is interested in the tensile strength of his

particular batch of cement to determine whether it will carry the required load. After a specified period of time, the builder would like to have such a statement as the probability is $1-\alpha$ that the tensile strength of his batch of cement will lie in a specified interval."

If we wish to estimate a single value of Y' given a value X' we use the line:

$$Y' = b_0 + b_1 X' ,$$

and the variance of estimate of a single Y'

$$\text{Var } Y' = \sigma^2_{Y.X} \left[1 + \frac{1}{n} + \frac{(X'-\bar{X})^2}{\Sigma x^2}\right]$$

If we wish to predict the average Y associated with a value X', we use the same line:

$$Y' = b_0 + b_1 X'$$

and the variance of estimate of mean Y

$$\text{Var } Y' = \sigma^2_{Y.X} \left[\frac{1}{n} + \frac{(X'-\bar{X})^2}{\Sigma x^2}\right]$$

(This latter variance is the variance of estimate of a point on the line). When both single Y's and the average Y are known to be normally distributed, we can make [illegible] interval estimates about them also (see problem 3.1.1.2).

Problem 3.1.1.2 Statement of Problem

Give a $(1-\alpha)$ confidence interval estimate for:

a) the line as a whole. (This is equivalent to giving a confidence interval estimate for the (population) mean value of Y corresponding to a value of X, simultaneously for all values of X.

b) a single point on the line (i.e., the population mean value of Y corresponding to a value of $X = X'$).

c) a single (future) value $Y = Y'$ corresponding to a new value of $X = X'$.

Solution:

a) Interval estimate for the whole line.

i) Choose $1-\alpha$, the confidence with which we wish to make our estimate.

ii) Compute s_Y^2 (our best estimate of σ_Y^2) using Basic Worksheet.

iii) Look up $F_{1-\alpha}$ for 2, $n-2$ degrees of freedom in Table III.

iv) Compute $w_1 = \sqrt{2F}\, s_Y \left[\frac{1}{n} + \frac{(X-\bar{X})^2}{\Sigma x^2} \right]^{1/2}$

v) A $(1-\alpha)$ confidence interval for the whole line is:

$$\bar{Y} + b_1 (X-\bar{X}) \pm w_1$$

b) Interval estimate of mean value of Y corresponding to a value of $X = X'$ (i.e., a single point on the line).

i) Choose $1-\alpha$, the confidence with which we wish to make our estimate.

ii) Compute s_Y^2 (our "best" estimate of $\sigma_{Y\cdot X}^2$) using Basic Worksheet.

iii) Look up $t_{1-\alpha/2}$ for $n-2$ degrees of freedom in Table II.

iv) Compute: $w_2 = t_{1-\alpha/2}\, s_Y \left[\frac{1}{n} + \frac{(X'-\bar{X})^2}{\Sigma x^2} \right]^{1/2}$

v) A $(1-\alpha)$ confidence interval estimate for the mean value of Y corresponding to $X = X'$ is:

$$\bar{Y} + b_1 (X'-\bar{X}) \pm w_2$$

(Note that an interval estimate of the intercept is obtained by setting $X' = 0$ in the above).

c) Interval estimate of an individual value Y' corresponding to a single value of $X = X'$.

i) Choose $1-\alpha$, the confidence with which we wish to make our estimate.

ii) Compute s_Y^2 (our "best" estimate of $\sigma_{Y\cdot X}^2$) using Basic Worksheet.

iii) Look up $t_{1-\alpha/2}$ for $n-2$ degrees of freedom in Table II.

iv) Compute $w_3 = t_{1-\alpha/2}\, s_Y \left[1 + \frac{1}{n} + \frac{(X'-\bar{X})^2}{\Sigma x^2} \right]^{1/2}$

v) A $(1-\alpha)$ confidence interval estimate for Y' (the single value of Y corresponding to $X = X'$) is

$$\bar{Y} + b_1 (X' - \bar{X}) \pm w_3$$

Once we have fitted the line, we will want to make predictions from it, and we will want to know how good our predictions are. Often these will be given in the form of an interval together with a confidence coefficient associated with the interval - i.e., confidence interval estimates. We may want to make one of several kinds of confidence interval estimates:

a) A confidence interval for the line as a whole.

b) A confidence interval for a single point on the line - i.e., a confidence interval for the mean value of Y corresponding to a single value of $X = X'$.

If the fitted line is, say, a calibration line which will be used over and over again, we will want to make the interval estimate described in (a). In other cases, the line as such may not be so important. The line may have been fitted only to investigate or check the structure of the relationship. The interest of the experimenter may be centered at one or two values of the variables.

Another kind of interval estimate is sometimes required:

c) A single value of Y corresponding to a new value of $X = X'$.

The 3 different kinds of confidence interval statements (a), (b), and (c) have somewhat different interpretations.

The confidence interval for (b) is interpreted as follows:

Suppose that we repeated our experiment a large number of times. Each time we obtain n pairs of values

(X_i, Y_i), fit the line, and compute a confidence interval estimate of the mean value of $Y(\bar{Y}_1)$ corresponding to some one value of $X(X_1)$. Such interval estimates (of Y associated with X_1) are expected to be correct a proportion $(1-\alpha)$ of the time. If we were to make an interval estimate of $Y(\bar{Y}_2)$ corresponding to another value of $X(X_2)$, these interval estimates would also be expected to be correct the same proportion $(1-\alpha)$ of the time. However, taken together, these intervals do not constitute a confidence statement which would be expected to be correct a proportion $(1-\alpha)$ of the time. Nor is the effective level of confidence $(1-\alpha)^2$ because the two statements are not independent but are correlated in a manner intimately dependent on the values X_1 and X_2 for which predictions are to be made.

The confidence interval for the whole line (a) implies the same sort of repetition of the experiment except that our confidence statements are not now limited to one X at a time, but we can talk about any number of X values simultaneously. The confidence intervals for $\bar{Y}$ corresponding to all the chosen X values will be simultaneously correct a proportion $(1-\alpha)$ of the time, and therefore our confidence statement applies to the line as a whole. It will be noted that the intervals in (a) are larger than the intervals in (b) by the ratio $\sqrt{2F}/t$. This wider interval is the "price" one pays for making joint statements

about Y for any or all of the X values, rather than the Y for a single X.

Another caution is in order. We cannot use the same computed line in (b) and (c) to make a large number of predictions and claim that $100(1-\alpha)\%$ of the predictions will be correct. The estimated line may be very close to the "true line", in which case nearly all of the interval predictions may be correct, or the line may be considerably different from the "true line" in which case very few may be correct. In practice, provided our situation is "in control", we should always revise our estimate of the line to include additional information in the way of new points.

Problem 3.1.1.3 Give a confidence interval estimate for β_1, the slope of the "true" line $y = \beta_0 + \beta_1 X$.

Solution.

i) Choose $1-\alpha$, the confidence with which we want to make our interval estimate.

ii) Look up $t_{1-\alpha/2}$ for $n-2$ degrees of freedom in Table II.

iii) Compute s_Y^2 using Basic Worksheet.

iv) Compute $w_4 = t_{1-\alpha/2} \dfrac{s_Y}{\sqrt{\Sigma x^2}}$

v) A $(1-\alpha)$ confidence interval estimate of β_1 is

$b_1 \pm w_4$

Problem 3.1.1.4 Give a $(1-\alpha)$ confidence interval estimate of the value of $X = X'$ (independent variable) that yielded n' new

values of Y having as their average $\bar{Y}'$.

Solution:

i) Look up $t_{1-\alpha/2}$ for n-2 degrees of freedom in Table II.

ii) Compute b_1 from Basic Worksheet.

iii) Compute s_Y^2 from Basic Worksheet.

iv) Compute

$$C = b_1^2 - \frac{t_{1-\alpha/2}^2 \, s_Y^2}{\Sigma x^2}$$

v) Put

$$X' = \bar{X} + \frac{b_1(\bar{Y}'-\bar{Y})}{C} \pm \frac{t_{1-\alpha/2} \, s_Y}{C} \sqrt{\frac{(Y'-\bar{Y})^2}{\Sigma x^2} + \left(\frac{1}{n} + \frac{1}{n'}\right) C}$$

This is a $1-\alpha$ confidence interval estimate of the value of $X = X'$ corresponding to $Y = \bar{Y}'$.

Problem 3.1.1.5 Give a value of $X = X'$ which you expect with confidence $(1-\alpha)$ will correspond to a value of Y not less than Q.

Solution:

i) Look up $t_{1-\alpha}$ for n-2 degrees of freedom in Table II.

ii) Compute b_1 from Basic Worksheet.

iii) Compute s_Y^2 from Basic Worksheet

iv) Compute $C = b_1^2 - \frac{t_{1-\alpha}^2 \, s_Y^2}{\Sigma x^2}$

v) Compute

$$X' = \bar{X} + b_1\left(\frac{(Y-\bar{Y})}{C}\right) + \frac{t_{1-\alpha} s_Y}{C} \sqrt{\frac{(Q-\bar{Y})^2}{\Sigma x^2} + \left(\frac{n+1}{n}\right) C}$$

where the sign before the last term is + if b_1 is positive or - if b_1 is negative. We have confidence $(1-\alpha)$ that a value of $X = X'$

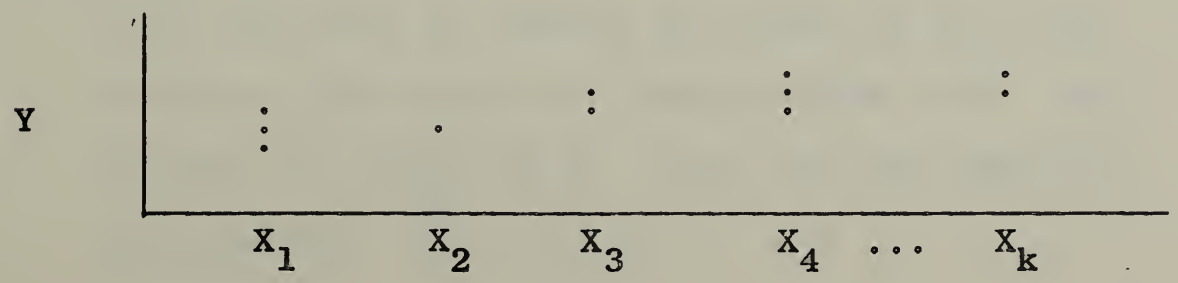
will correspond to (produce) a value of Y not less than Q.

(See discussion of "confidence" in straight line prediction in narrative after Problem 3.1.1.2).

Problem 3.1.1.6. Is the assumption of linear regression justified? This involves a test of the assumption that the mean Y values for given X values do lie on a straight line (we assume that for any given value of X, the corresponding Y values are normally distributed with variance $\sigma^2_{Y.X}$, which is independent of the value of X).

Solution:

A test is available provided that we have more than one observation on Y at one or more values of X. Assume that we have n pairs of values (X_i, Y_i) and that among these pairs there occur only k values of $X (k < n)$. For example, the plot might look as follows:



Our observations would be recorded in a table like the following: (i.e., we have n_i observations at each X and we have k different values of $X (k < n$, and $\Sigma n_i = n)$.

	X_1	X_2	X_3	...	X_k	$X = \frac{\Sigma n_i X_i}{n}$
	Y_{11}	Y_{21}	Y_{31}	...	Y_{k1}	
	Y_{12}		Y_{32}	...	Y_{k2}	
	Y_{13}					
Total $T_{Y.X}$						Grand Total T_Y =
Sum of Squares						
No. of observations n_i	-	-	-	...	-	

Compute:

i) $T_{Y.X}$ = the total of Y for each X

T_Y = the total of all Y

$\overline{Y} = \frac{T_Y}{n}$

$\overline{X}$ = the weighted average of X, i.e., $\frac{\Sigma n_i X_i}{n}$

ii) Compute: $$S_1 = \sum_{i=1}^{k} \frac{(T_{Y.X})^2}{n_i} - \frac{T_Y^2}{n}$$

i.e., for each X, square the value of $T_{Y.X}$ and divide by the number of observations of Y. Sum across all values of X. From this sum subtract the quantity $\frac{T_Y^2}{n}$

iii) Compute $$b = \frac{\Sigma XY - \frac{(\Sigma n_i X_i)(\Sigma Y)}{n}}{\Sigma n_i X_i - \frac{(\Sigma n_i X_i)^2}{n}}$$

iv) Compute $S_2 = b\left[\Sigma XY - \frac{(\Sigma n_i X_i)(\Sigma Y)}{n}\right]$

v) Compute $S_3 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$

vi) Look up $F^*_{1-\alpha}$ for k-2, n-k degrees of freedom in Table III.

vii) Compute $F = \left(\frac{S_1 - S_2}{S_3 - S_1}\right)\left(\frac{n-k}{k-2}\right)$

viii) If $F > F^*_{1-\alpha}$, decide that the mean values of Y for given X do not lie on a straight line.

If $F < F^*$, the hypothesis of a linearity is not disproved.

3.1.2 Problems for FII Relationships

Problem 3.1.2.1 What is our "best" estimate of the true relationship?

Solution:

Assuming that we know the ratio of the two variances $\lambda = \sigma_v^2/\sigma_u^2$ we shall estimate the relationship by the following formula:

$$Y = \bar{Y} + (\theta + \sqrt{\theta^2 + \lambda}\,)(X - \bar{X})$$

where

$$\theta = \frac{s_Y^2 - s_X^2}{2s_{XY}}$$

$$s_Y^2 = \frac{1}{n-1}(\Sigma Y_i^2 - n\bar{Y}^2),\quad s_X^2 = \frac{1}{n-1}(\Sigma X_i^2 - n\bar{X}^2)$$

$$s_{XY} = \frac{1}{n-1}(\Sigma X_i Y_i - n\overline{XY}).$$

It is worth noting that we can make estimates of σ_u^2 , and σ_v^2 as follows

$$s_u^2 = \frac{s_Y^2 - (\theta + \sqrt{\theta^2 + \lambda})\, s_{XY}}{\lambda}$$

$$s_v^2 = \lambda s_u^2 .$$

3.1.3 Problems for SI Relationships

Problem 3.1.3.1 What is the best line to be used for estimating Y from given values of X?

The solution is identical to that of problem 3.1.1.1

Problem 3.1.3.2 Give confidence intervals for:

the line as a whole;

a point on the line;

a single Y corresponding to a new value of X

The solution is identical to that of problem 3.1.1.2

Problem 3.1.3.3 Give a confidence interval estimate for the slope of the "true" line.

The solution is identical to that of problem 3.1.1.3

Problem 3.1.3.4 What is the best line for predicting X from given values of Y?

Solution:

For this problem we fit a line $X = b_0' + b_1' Y$ (an estimate of the true line $X = \beta_0' + \beta_1' Y$). To fit this line we need to interchange the roles of the X and Y variables in the computations outlined in the Basic Worksheet and proceed as in problem 3.1.1.1.

Problem 3.1.3.5 What is the degree of relationship of the two variables X and Y as measured by r, the coefficient of correlation.

Solution: Compute $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$

Compute Σx^2, Σy^2, Σxy from Basic Worksheet

The sample product moment correlation coefficient is an estimate of $\rho = \sqrt{\beta_1 \beta_1'}$, the "true" product correlation coefficient. A .95 confidence interval estimate for ρ can be obtained by choosing the belt corresponding to n in Table VII, and reading the limits for the computed value of r.

If $\rho = \pm 1$ all the points lie on a line, and $Y = \beta_0 + \beta_1 X$ and $X = \beta_0' + \beta' Y$ coincide. If $\rho = +1$, the slope is positive, and if $\rho = -1$, the slope is negative. If $\rho = 0$ then X and Y are said to be uncorrelated. If the confidence interval from Table VII excludes $\rho = 0$, then we may state that the data give reason to believe that there is a relationship between the two variables. The confidence coefficient associated with this statement is .95.

3.1.4 Problems for SII Relationships

3.1.4.1 What is the best line to be used for estimating Y from given values of X?

The solution is identical to that of problem 3.1.1.1.

3.1.4.2 Give confidence intervals for:

the line as a whole;

a point on the line;

A single Y corresponding to a new value of X.

The solution is identical to that of problem 3.1.1.2

3.1.4.3 Give a confidence interval estimate for the slope of the "true" line.

The solution is identical to that of problem 3.1.1.3.

U. S. DEPARTMENT OF COMMERCE

Sinclair Weeks, *Secretary*

NATIONAL BUREAU OF STANDARDS

A. V. Astin, *Director*

THE NATIONAL BUREAU OF STANDARDS

The scope of activities of the National Bureau of Standards at its headquarters in Washington, D. C., and its major field laboratories in Boulder, Colorado, is suggested in the following listing of the divisions and sections engaged in technical work. In general, each section carries out specialized research, development, and engineering in the field indicated by its title. A brief description of the activities, and of the resultant reports and publications, appears on the inside front cover of this report.

WASHINGTON, D. C.

Electricity and Electronics. Resistance and Reactance. Electron Tubes. Electrical Instruments. Magnetic Measurements. Dielectrics. Engineering Electronics. Electronic Instrumentation. Electrochemistry.

Optics and Metrology. Photometry and Colorimetry. Optical Instruments. Photographic Technology. Length. Engineering Metrology.

Heat and Power. Temperature Physics. Thermodynamics. Cryogenic Physics. Rheology and Lubrication. Engine Fuels.

Atomic and Radiation Physics. Spectroscopy. Radiometry. Mass Spectrometry. Solid State Physics. Electron Physics. Atomic Physics. Nuclear Physics. Radioactivity. X-rays. Betatron. Nucleonic Instrumentation. Radiological Equipment. AEC Radiation Instruments.

Chemistry. Organic Coatings. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Gas Chemistry. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

Mechanics. Sound. Mechanical Instruments. Fluid Mechanics. Engineering Mechanics. Mass and Scale. Capacity, Density, and Fluid Meters. Combustion Controls.

Organic and Fibrous Materials. Rubber. Textiles. Paper. Leather. Testing and Specifications. Polymer Structure. Organic Plastics. Dental Research.

Metallurgy. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion. Metal Physics.

Mineral Products. Engineering Ceramics. Glass. Refractories. Enameled Metals. Concreting Materials. Constitution and Microstructure.

Building Technology. Structural Engineering. Fire Protection. Heating and Air Conditioning. Floor, Roof, and Wall Coverings. Codes and Specifications.

Applied Mathematics. Numerical Analysis. Computation. Statistical Engineering. Mathematical Physics.

Data Processing Systems. SEAC Engineering Group. Components and Techniques. Digital Circuitry. Digital Systems. Analogue Systems. Application Engineering.

- Office of Basic Instrumentation
- Office of Weights and Measures

BOULDER, COLORADO

Cryogenic Engineering. Cryogenic Equipment. Cryogenic Processes. Properties of Materials. Gas Liquefaction.

Radio Propagation Physics. Upper Atmosphere Research. Ionospheric Research. Regular Propagation Services. Sun-Earth Relationships.

Radio Propagation Engineering. Data Reduction Instrumentation. Modulation Systems. Navigation Systems. Radio Noise. Tropospheric Measurements. Tropospheric Analysis. Radio Systems Application Engineering.

Radio Standards. Radio Frequencies. Microwave Frequencies. High Frequency Electrical Standards. Radio Broadcast Service. High Frequency Impedance Standards. Calibration Center. Microwave Physics. Microwave Circuit Standards.


NBS