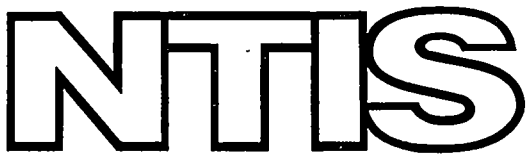
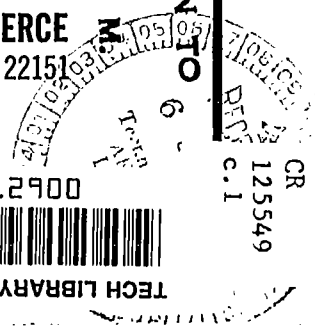
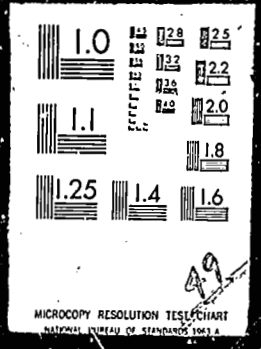
NT-16961

DISTRIBUTED BY:



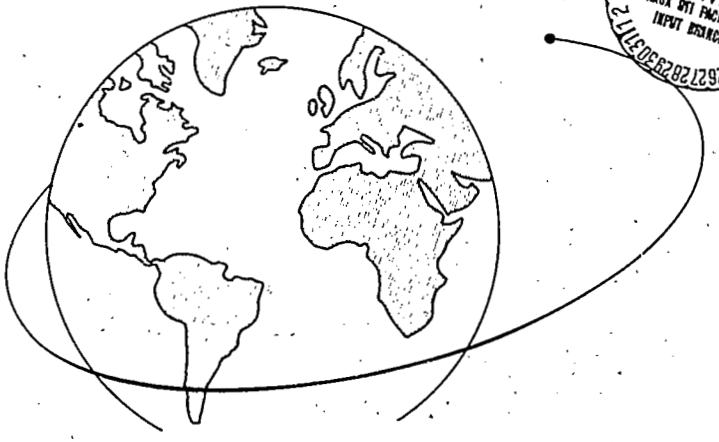
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151





FOUNDATIONS FOR ESTIMATION BY THE METHOD OF LEAST SQUARES

W. W. HAUCK, JR.



Smithsonian Astrophysical Observatory
SPECIAL REPORT 340

Research in Space Science
SAO Special Report No. 340

FOUNDATIONS FOR ESTIMATION BY THE METHOD OF LEAST SQUARES

Walter W. Hauck, Jr.

December 27, 1971

Smithsonian Institution
Astrophysical Observatory
Cambridge, Massachusetts 02138

009-113

TABLE OF CONTENTS

PRECEDING PAGE BLANK NOT FILMED

ABSTRACT

This paper discusses least-squares estimation from the point of view of a statistician. Much of the emphasis will be on problems encountered in application and, more specifically, on questions involving assumptions — what assumptions are needed, when are they needed, what happens if they are not valid, and if they are invalid, how can we detect that fact.

RÉSUMÉ

Cet article est une discussion de l'estimation des moindres carrés du point de vue du statisticien. On mettra surtout l'accent sur les problèmes rencontrés en pratique et plus spécialement sur les questions impliquant des suppositions - quelles suppositions sont nécessaires, quand sont-elles nécessaires, ce qui arrive si elles ne sont pas valides, et si elles sont invalides, comment peut-on déceler ce fait.

КОНСПЕКТ

В этом докладе обсуждается оценка по методу наименьших квадратов с точки зрения статистика. Много внимания обращается на задачи встречающиеся при применении и в особенности на вопросы затрагивающие предположения-какие предположения необходимы, когда они необходимы, что случается если они недействительны, и если они недействительны, как мы можем определить этот факт.

FOUNDATIONS FOR ESTIMATION BY THE METHOD OF LEAST SQUARES

Walter W. Hauck, Jr.

1. INTRODUCTION

This paper is the result of four seminars given to the Satellite Geophysics Group of the Smithsonian Astrophysical Observatory in August and September 1970. The purpose of the seminars was to consider methods of applying least-squares estimation to satellite tracking.

The method of least squares is widely used for estimation, although in many applications little consideration is given to its strengths and limitations. On the other hand, statisticians have done considerable work on the subject, under the heading of regression, although not always on those questions that are of the most interest in application.

A knowledge of basic probability and statistics is required. For review, the necessary concepts are explained in Section 2. The notation introduced there is used consistently throughout the paper. For reference, especially for those not reading the rest of the section, a glossary of notation is included at the end of Section 2. A knowledge of basic matrix theory will be assumed.

This work was supported in part by grant NGR 09-015-002 from the National Aeronautics and Space Administration.

2. PROBABILITY AND STATISTICS

A knowledge of some probabilistic and statistical concepts is necessary for an understanding of the discussion that follows. The level of this explanation will be that of a "quick refresher." For a more detailed explanation, refer to an introductory probability and statistics text, such as that by Hogg and Craig (1965).

2.1 Probability, Random Variables, and Distribution Theory

A natural first question is: What is probability? The currently popular approach is to treat probabilities as a particular class of mathematical measures. This approach is very rigorous and keeps mathematicians happy, but it does not answer the question of interest. To do that, we will use the relative frequency approach.*

First of all, it is necessary to have some group or aggregate to study. This group, whether of people, things, or events, will be called the population. Next, there is some property of this population that we are concerned with, and there must be something about this property that is undetermined. If everything is known about what is going on, there are no probabilities to determine.

This property must be able to be evaluated for each member of the population, and a numerical† value assigned to that evaluation. A random variable is a function of the members of the population; its value is the numerical evaluation of the property for that member. We will use capital

*An alternate approach, which I do not agree with, views probability theory as the study of human reasoning processes, and probabilities as subjective measures of degrees of certainty.

†The use of numerical here is meant to be very general.

letters to denote the random variable, the argument of which will never be explicitly stated, and small letters to denote the values taken on by a random variable.*

For example, take the population to be all flips of a coin, and the property to be whether it lands heads or tails. Assuming the coin does not have two heads or two tails, it is not known before the flip on which side the coin will land.

One possible random variable, denoted by X, is an indicator variable; that is,

X = 1 if heads, and

X = 0 if tails.

The set of all possible values the random variable may take is called the sample space, denoted by S.† In the example, $S = \{0, 1\}$.

To derive probabilities, it is necessary to distinguish between discrete and continuous sample spaces.

Discrete Case. Let $x_1, x_2, \ldots, x_N$ (where N may be infinity) denote the points of the sample space. Consider taking some n members of the population and recording the value of the random variable X for each member. For $i=1, \ldots, N$, let $f_i^{(n)}$ be the proportion of these n members for which $X = x_i$. Then, take more and more members of the population, record the values of X, and keep updating $\{f_i^{(n)}\}_{i=1}^N$. For a finite population, take all the members. For an infinite population, take the limit as $n \to \infty$. The final $\{f_i\}_{i=1}^N$ obtained by this method is the density function of the discrete random variable X. We can then say that the probability that the property in question will be evaluated as equal to x_i is f_i, or in shorthand, $P[X = x_i] = f_i$. Usually f_i will be written as $f(x_i)$.

*There will be exceptions to this rule in later sections.

†Strictly speaking, this is only one representation of the sample space, but the more general notion is not necessary for our purposes.

Continuous Case. Again consider the procedure of starting with n members of the population, recording the values of the random variable, and then taking more and more members to the limit of the entire population. This time, let

$$\cdots, x_{-2}, x_{-1}, x_0, x_1, x_2, \cdots$$

be a sequence of points such that

$$x_{i+1} - x_i = \Delta x \quad ,$$

where Δx is some positive constant. Then, let $f_n(x_i)\,\Delta x$ be the proportion of values falling in the half-open interval $(x_i - \Delta x/2,\ x_i + \Delta x/2]$. The limiting process is now two simultaneous processes: while taking $n \to \infty$, let $\Delta x \to 0$ in such a way as to avoid the occurrence of irregular frequencies. The problem is that if $\Delta x \to 0$ too quickly, there will be intervals where nothing has occurred simply because the number of members taken is not large enough.

In the limit,

$$P[x - \tfrac{1}{2}\,dx < X \leq x + \tfrac{1}{2}\,dx] = f(x)\,dx \quad .$$

That is, the probability of observing a value in an infinitesimal interval centered at x is given by f(x) dx. By taking the limit of sums, we have

$$P[a < X \leq b] = \int_a^b f(x)\,dx \quad ;$$

$f(\cdot)$ is the density function of the continuous random variable X.

For both cases, the distribution function F is defined by

$$F(x) = P[X \leq x] \quad ,$$

that is, the probability of observing a value $\leq x$.

In introductory texts, the three properties of probability that are presented as defining it are consequences of this derivation. These properties are the following:

If A and B are two subsets of S and if ϕ denotes the null set, then

1)* $P(A) \geq 0$.

2) If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$.

3) $P(S) = 1$.

The approach in terms of measures, referred to earlier, defines a specific class of mathematical measures as probability measures. These measures correspond to the distribution function as developed here. The measures are more general because the corresponding density function may not exist. Proofs based on this method can become quite complicated because of various measure theoretic problems that must be considered. It is my observation that a statistician will make as many measure theoretic assumptions as are necessary to prove a theorem (for example, that a function is measurable) since, in practice, they will be true.

The concept of probability can be extended to the case where two or more properties of the population are considered simultaneously. A derivation similar to that done here leads to bivariate and multivariate densities and distributions.

For example, consider the population of all men and their height (H) and weight (W). Both height and weight will have individual densities and, if considered together, a bivariate density. Now restrict the population to all men with a certain weight — say, $W = w_0$. The density of H derived for this restricted population is called the conditional density of height given weight and is denoted by $f(h|\,W = w_0)$ or $f(h|\,w_0)$.

*If the random variable is understood, the notation P(D) will sometimes be used as shorthand for $P[X \in D]$, the probability that the value of the random variable X will lie in the set D.

Two random variables X and Y are said to be (stochastically) independent if $f(x|y) = f(x)$ for all possible values y of Y. Intuitively, this says that knowing the value of Y does not provide any information about X. Two random variables that are not independent are said to be dependent.

Before we go on, it is important to note that the population under consideration may be "ideal" — that is, a hypothetical population that satisfies certain properties and is used to approximate a real population. Most distributions in use — for example, the normal and the Poisson — were first derived for this type of population.

2.2 Expectation

The expectation (or expected value) of any function $g(\cdot)$ of the random variable X is a weighted average of the value of the function over all possible values of X, the weights given by the density function. That is, for discrete X,

$$E[g(X)] = \sum_{i=1}^{N} f(x_i)\, g(x_i)$$

(remember that N may be infinity), and for continuous X,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx \quad ,$$

where $E(\cdot)$ denotes expectation. Mathematically, $E(\cdot)$ is a linear operator.

This concept can also be extended to multivariate and conditional cases by substituting the appropriate density into the above formulas. In the conditional case, the notation is $E[g(X)|y]$.

We will be concerned with three particular functions, the third an example of the bivariate case:

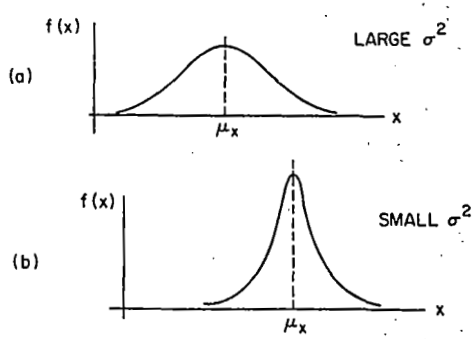
1) $g(X) = X$.

Then $E[g(X)]$ is the mean value of X, denoted by μ_X. This can be considered the average value of X. For a finite population, it is exactly equal to the average. For an infinite population, it is commonly referred to as the long-range average.

2) $g(X) = (X - \mu_X)^2$.

$E[g(X)]$, denoted by σ_X^2 or Var (X), is the variance of X (see diagram below). The variance is a measure of dispersion, that is, a measure of how close to the mean the values of X are. For example,

009-113



A commonly used quantity is the standard deviation σ_X, equal to $\sqrt{\sigma_X^2}$.

Related to the variance is a well-known and sometimes very useful result known as Tchebycheff's Inequality:

$$P[|X - \mu_X| \geq k\sigma_X] \leq \frac{1}{k^2}$$

for any $k > 0$ and for any distribution. The approximation is very poor for small k (for example, try any $k \leq 1$), but for large k ($k \geq 3$) the upper bound can be very useful.

3) $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$.

$E[g(X,Y)]$ is the <u>covariance</u> of X and Y, denoted by Cov (X, Y) or σ_{XY}. Two simple properties of the covariance are

$$\text{Cov}(X, Y) = \text{Cov}(Y, X), \text{ and}$$

$$\text{Cov}(X, Y) = E[XY] - \mu_X\mu_Y .$$

Covariance is a measure of association, but for that purpose it is not well suited, since it is not invariant under a change of scale; that is, $\text{Cov}(aX, Y) \neq \text{Cov}(X, Y)$ for any constant $a \neq 1$. What is used is the (product-moment) <u>correlation coefficient</u>: $\rho_{XY} = \text{Cov}(X, Y)/\sigma_X \sigma_Y$, which is scale invariant.

Both covariance and correlation originate from studies of the multivariate normal distribution, where they have a specific meaning; that is, the exact nature of the association being measured is clear. This is not true for other distributions.

Some understanding of the nature of the association measured by ρ_{XY} can be obtained by considering the following properties:

1) If $Y = aX + b$, where a and b are constants, then

$$\rho_{XY} = \text{sign}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -1 & \text{if } a < 0 \end{cases} ,$$

2) If X and Y are independent, then

$$\rho_{XY} = 0 ,$$

but the converse is not true unless both X and Y are normally distributed. If $\rho_{XY} = 0$, X and Y are said to be <u>uncorrelated</u>.

When a set of n random variables $X_1, \ldots, X_n$ is being considered, it is more convenient to work with the <u>covariance matrix</u>, Σ, defined by

$$\Sigma_{ij} = \begin{cases} \text{Cov}(X_i, X_j) & \text{if } i \neq j , \\ \text{Var}(X_i) & \text{if } i = j ; \end{cases}$$

Σ is a symmetric $n \times n$ matrix that will usually be positive definite.

If

$$X = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_n \end{pmatrix} \text{ and } \mu_X = \begin{pmatrix} \mu_{X_1} \\ \cdot \\ \cdot \\ \mu_{X_n} \end{pmatrix} ,$$

then

$$\Sigma_X = E[(X - \mu_X)(X - \mu_X)'] ,$$

where the subscript X now refers to the vector, and the prime denotes transpose. The expected value of a matrix is the matrix of expected values.

2.3 Distributions of Interest

2.3.1 Normal distribution

The normal distribution is the most important in least-squares theory. It can be derived for many different ideal populations. For example, Maxwell derived it as the distribution of velocities of molecules. It was also derived by Hagen as the distribution of errors under the following assumptions:

1) An error is the sum of a large number of infinitesimal errors, all of equal magnitude and all due to different causes.

2) The different components of errors are independent.

3) Each component of error has an equal chance of being positive or negative.

In our terminology and notation, this is saying

$$\text{Error} = X_1 + X_2 + \dots \quad ,$$

where the X_i's are independent random variables such that

$$P[X_i = +\epsilon] = P[X_i = -\epsilon] = \frac{1}{2}$$

for some infinitesimal ϵ and for all i.

These assumptions are very restrictive but can be greatly relaxed. We will come back to this later.

The normal distribution is characterized by two parameters, μ and σ. If X is normally distributed, its density function is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \quad .$$

It is then possible to derive that

$$E[X] = \mu$$

and

$$\text{Var}[X] = \sigma^2$$

A shorthand notation* is

$$X \sim N(\mu, \sigma^2) \quad ,$$

which translates as "the distribution of X is normal with mean μ and variance σ^2."† A property of the normal distribution is that if $X \sim N(\mu, \sigma^2)$ and $Y = (X - \mu)/\sigma$, then

$$Y \sim N(0, 1) \quad .$$

*Throughout the paper, the symbol ~ will mean "is distributed as."

†When there would be no confusion, the subscript X on μ_X and σ_X^2 can be dropped.

N(0, 1) is referred to as the standard normal and is the one tabulated in tables of the normal distribution function. Because of the above property, it is possible to transform any normal random variable into the standard form.

Returning to the theory of errors, it is not Hagen's result that makes the normal distribution important. His assumptions are much too restrictive. The result that is usually cited is the Central Limit Theorem, which gives conditions for convergence to normality, but its assumptions can also be relaxed. The more useful results are theorems due to Liapunov and to Lindeberg and Feller.

Theorem 1 (Liapunov's Theorem)

Let $S_n = X_1 + \dots + X_n$ be the sum of n independent random variables with means $E[X_i] = \mu_i$, variances $\text{Var}(X_i) = \sigma_i^2 \neq 0$, and $\gamma_i = E[|X_i - \mu_i|^3]$. Let

$$Z_n = \frac{S_n - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

If

$$L_n = \frac{\sum_{i=1}^{n} \gamma_i}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} \xrightarrow[n \to \infty]{} 0 \quad ,$$

then

$$Z_n \xrightarrow[n \to \infty]{d} N(0, 1)$$

[$\xrightarrow[n \to \infty]{d}$ denotes convergence in distribution, that is, the distribution of Z_n converges to N(0, 1); and $\xrightarrow[n \to \infty]{}$, without the d, indicates the usual mathematical limit]. Then, under the conditions of the theorem, the distribution of S_n is approximately $N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$ for large n.

The assumption $L_n \xrightarrow[n \to \infty]{} 0$ is referred to as "negligibility in the limit." A heuristic condition for this assumption to be valid is that no X_i dominates the others – that is, the random variables do not differ too much in either magnitude or variance.

Theorem 2 (Lindeberg-Feller Theorem)

Let $S_n = X_1 + \ldots + X_n$ be the sum of n independent random variables with means $E[X_i] = \mu_i$, variances $Var(X_i) = \sigma_i^2 \neq 0$, and density functions $f_i(\cdot)$. Let

$$s_n = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

and

$$Z_n = \frac{S_n - \sum_{i=1}^{n} \mu_i}{s_n}$$

If for all $\epsilon > 0$,*

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|x-\mu_i| > \epsilon s_n} (x - \mu_i)^2 f_i(x)\, dx = 0 \quad , \qquad (1)$$

then

$$Z_n \xrightarrow[n \to \infty]{d} N(0,1) \quad .$$

Condition (1) is referred to as Lindeberg's condition. Roughly speaking, this condition requires that the variance σ_i^2 be due mainly to the density in an interval whose length is small in comparison with s_n. Feller (1966) shows that in a certain sense, Lindeberg's condition is necessary for convergence to normality. In addition, he provides examples of when it is satisfied.

Both Lindeberg's condition and the condition in Liapunov's Theorem that $L_n \xrightarrow[n \to \infty]{} 0$ are satisfied when the X_i's have the same distribution with finite variance. In that case, the statement of the theorem can be simplified as follows:

* $\int_{|x-\mu_i| > \epsilon s_n}$ indicates that the integral is taken over the set of x that satisfies $|x-\mu_i| > \epsilon s_n$.

Theorem 3 (Central Limit Theorem)*

Let $S_n \equiv X_1 + \ldots + X_n$ be the sum of n iid (= independent and identically distributed) random variables with mean μ and variance $0 < \sigma^2 < \infty$. Let

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \frac{\sqrt{n}\,(\frac{1}{n}S_n - \mu)}{\sigma} .$$

then

$$Z_n \xrightarrow[n \to \infty]{d} N(0,1)$$

As in the previous two theorems, the distribution of Z_n is approximately $N(0,1)$ for large n. How large n has to be for this approximation to be good depends on the distribution of the X_i's. For example, if $X_i \sim N(\mu, \sigma^2)$, then $Z_n \sim N(0,1)$ exactly for any n. For other distributions, $n \geq 20$ or 25 is usually large enough for the approximation to be good.

The crucial assumption in all three theorems is that the random variables be independent. Only in a few special cases has it been possible to prove convergence to normality when dependence is allowed.

The role that the correlation coefficient and covariance matrix play in normal distribution theory can be seen by examining the multivariate normal density function: Let

$$X = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ X_n \end{pmatrix} \quad , \quad \mu = \begin{pmatrix} \mu_{X_1} \\ \cdot \\ \cdot \\ \mu_{X_n} \end{pmatrix}$$

and $\Sigma = E[(X - \mu)(X - \mu)']$, the covariance matrix. Then,

$$f(X) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (X - \mu)' \Sigma^{-1} (X - \mu) \right] \qquad (2)$$

*The previous two theorems are also central limit theorems, but the capital letters on "Central Limit" are usually reserved for this result.

is the multivariate normal density function. For two random variables, X and Y, with a bivariate normal distribution, equation (2) reduces to

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} ,$$

where $\rho = \rho_{XY}$. As can be seen, Σ and ρ are parameters of the distribution. Also, in the bivariate case, ρ satisfies $\mathrm{Var}[X|y] = (1-\rho^2)\sigma_X^2$, which gives ρ, or more correctly ρ^2, a specific interpretation.

2.3.2 Other distributions

Three distributions will be needed for making tests of significance. A discussion of each follows.

1. Chi-Square Distribution

If $X_1, \ldots, X_n$ are iid $N(0, 1)$, then $\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$, where χ_n^2 denotes the chi-square distribution with n degrees of freedom; n is the parameter of the distribution.

Two theorems concerning this distribution are in order.

Theorem 4

If $X_1, \ldots, X_m$ are independent random variables such that $X_i \sim \chi_{n_i}^2$, then

$$\sum_{i=1}^{m} X_i \sim \chi_k^2 \quad , \quad \text{where } k = \sum_{i=1}^{m} n_i \quad .$$

Theorem 5

If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ and if

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \quad ,$$

then

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi_{n-1}^2$$

independent of $\bar{X}_n$.

Regardless of the distribution of the X_i's, as long as they are iid, $\bar{X}_n$ is called the sample mean and s_n^2 is called the sample variance. If present, the subscripts indicate the number of observations.

2. Student's t Distribution*

If $X \sim N(0, 1)$ and $Y \sim \chi_n^2$ independent of X, then

$$\frac{X}{\sqrt{Y/n}} \sim t_n \quad ,$$

where t_n denotes the t distribution with n degrees of freedom.

*It is called Student's t because William Gosset, who first derived this distribution, was prevented, by the brewery where he worked, from publishing the result under his own name. So he published it under the pseudonym "A Student."

Theorem 6

$$t_n \xrightarrow[n \to \infty]{} N(0,1) \quad ;$$

that is, the limit of the t distribution, as the number of degrees of freedom approaches infinity, is the standard normal distribution.

The following theorem is an immediate result of Theorem 5.

Theorem 7

If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{s_n} \sim t_{n-1} \quad .$$

3. F Distribution

If $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ independent of X, then

$$\frac{X/n}{Y/m} \sim F(n,m) \quad ,$$

the F distribution with n and m degrees of freedom.

Theorem 8

If $X \sim t_n$, then

$$X^2 \sim F(1,n) \quad .$$

The formula of the density functions of these three distributions is not necessary. Most statistics books contain tables of their distribution function, which is all that is needed

As presented in this section, the term degrees of freedom is used only to designate the parameters of these distributions. The reason for the terminology is related to estimation, especially of variances. In a very

general way, one degree of freedom is gained for every observation if the observations are independent, and one lost for every parameter estimated. We will return to this subject in Section 5.1.1, which should clarify all that is necessary for this paper.

2.4 Statistical Inference

2.4.1 Estimation

In almost all cases of interest, it is very difficult, if not impossible (as in the case of infinite populations), to determine exactly certain properties of the population under consideration. For example, an exact determination of the mean height or weight of the world population would be a somewhat difficult task.

The alternative is to take a sample (that is, some subset) of members of the population and determine the value of the property for these members. Some function of these observations is then used to approximate (estimate) the value of the property for the entire population. The very extensive problem of sampling theory – viz., how the members of the sample should be chosen – is extraneous to the purpose of this paper and so will not be discussed.

The questions that are of interest here and that keep many statisticians employed, are the following: Which functions of the observations should be used? Or, more specifically, what characterizes a good estimate, and are there general methods for finding them? Before we attack these questions, some notation is necessary.

Let the density of the random variable X be denoted by $f(x;\theta)$, where θ is the unknown parameter (corresponding to some property of the underlying population) that we wish to estimate. Suppose that the sample is of size n

and that the observed values are $x_1, \ldots, x_n$. Denote an estimate of θ by $g(x_1, \ldots, x_n)$, where g is some function. Note that, before the n observations are taken, $g(X_1, \ldots, X_n)$ can be treated as a random variable with its own distribution, which in theory can be derived from the distribution of X. Now $X_1, \ldots, X_n$ are n identically distributed random variables, though not necessarily independent.

Some properties that g may possess are the following:

1) Consistency.
g is consistent if

$$g(x_1, \ldots, x_n) \xrightarrow[n \to \infty]{} \theta \quad ;$$

that is, the estimate converges to the true value as the sample approaches the entire population. This is a minimum condition to be placed on an estimate.

2) Minimum mean-square error.
g has this property if it minimizes $E\{[h(X_1, \ldots, X_n) - \theta]^2\}$ over all possible functions of the observations h.

A problem here is that the quantity to be minimized depends on the unknown θ. It is gratifying when one function minimizes the mean-square error (MSE) for all θ. In practice, it is usually necessary to find the estimate that minimizes the MSE on some interval that is thought to contain θ.

3) Unbiasedness.
g is unbiased if

$$E[g(X_1, \ldots, X_n)] = \theta \quad .$$

When g is not unbiased, its <u>bias</u> is given by

$$E[g(X_1, \ldots, X_n)] - \theta \quad .$$

4) Minimum variance
g has this property if it minimizes

$$\mathrm{Var}\,[h(X_1, \ldots, X_n)]$$

over all functions h. This property is undesirable if g has a large bias, since the distribution of g would then be concentrated around the wrong value. A desirable estimate would be the minimum-variance unbiased (MVU) estimate. This property provides a criterion for choosing among unbiased estimates when more than one exists, although there is the problem, as with MSE, that the variance will usually depend on θ. In that case, an estimate is MVU if it is unbiased and has minimum-variance among all unbiased estimates for some value of θ.

It is generally desirable to find an unbiased or MVU estimate, but a word of caution is in order. Even when such an estimate exists, it does not always make sense This can be especially troublesome for the MVU case, since for a large class of problems the MVU estimate is unique.

As an example, suppose that $f(x;\lambda) = e^{-\lambda} \lambda^x / x!$, the Poisson density with mean λ ($\lambda > 0$), and that $\theta_1 = e^{-\lambda}$ is to be estimated on the basis of one observation. The only unbiased estimate, and hence the MVU estimate, is

$$g_1(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x = 1, 2, \ldots \end{cases}$$

If $\theta_2 = e^{-2\lambda}$ is to be estimated, the only MVU estimate is

$$g_2(x) = (-1)^x \quad .$$

g_1 may be acceptable in some cases, but g_2 is plainly nonsensical. Among other things, it does not make sense to use a negative estimate of a parameter that is known to be positive.

In most cases, it is desirable to have an estimate of the variance of an estimate. If s^2 is an unbiased estimate of the variance of $g(X_1, \ldots, X_n)$, then s is called the <u>standard error</u> of g.

Only two of the many methods for determining estimates will be discussed here. The first is the method of <u>least squares</u>.

In general, g is chosen to minimize

$$\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \quad ,$$

where $y_i = h(\theta)$ for some function h, $\hat{y}_i = h[g(x_1, \ldots, x_n)]$, and $\{w_i\}_{i=1}^n$ is a set of known constants or weights. This method is the subject of the remainder of the paper, so no more will be said about it here.

The second is the method of maximum likelihood. If $X_1, \ldots, X_n$ are iid with density $f(x;\theta)$, then

$$f(x_1, \ldots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

is called the likelihood function and denoted by $L(\theta;x_1, \ldots, x_n)$. It is a function of the unknown θ that treats the observations as known parameters. The maximum-likelihood estimate of θ, denoted by $\hat{\theta}$,* is the estimate of θ that maximizes $L(\theta;x_1, \ldots, x_n)$; that is,

$$L(\hat{\theta};x_1, \ldots, x_n) \geq L[h(x_1, \ldots, x_n); x_1, \ldots, x_n]$$

for all other functions of the observations.

The maximum-likelihood estimate is usually found by setting

$$\frac{d}{d\theta} L(\theta;x_1, \ldots, x_n) = 0$$

or, equivalently,†

$$\frac{d}{d\theta} \log L(\theta;x_1, \ldots, x_n) = 0 \quad .$$

This last equation is referred to as the likelihood equation.

*$\hat{\theta}$ will always denote an estimate of θ regardless of the method used to obtain it.

†All logs in this paper are natural or base e.

Theorem 9

If $\log L(\theta;x_1, \ldots, x_n)$ is differentiable in an interval including the true value, θ_0, let $\hat{\theta}$ be a root of the likelihood equation; that is, $(d/d\theta) \log L(\theta;x_1, \ldots, x_n)\big|_{\theta=\hat{\theta}} = 0$. Then, under certain conditions on $f(x;\theta)$,

1) $\hat{\theta}$ is a consistent estimate of θ, and

2) $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow[n \to \infty]{d} N(0, I^{-1})$,

where I, known as Fisher's information, equals $E\left[\left(\frac{d}{d\theta} \log f(x;\theta)\right)^2\right]$.

A result due to Cramér and Rao is that if $\bar{\theta}$ is an unbiased estimate of θ, then

$$\text{Var}(\bar{\theta}) \geq I^{-1} \frac{1}{n} \quad ,$$

where n is the number of observations. Combined with Theorem 9, this means that, asymptotically, the maximum-likelihood estimate is a minimum-variance unbiased estimate.

As with all asymptotic results, the question is: what happens for finite n? In this case, $n \geq 25$ is usually sufficient for $\hat{\theta}$ to be very close to a minimum-variance unbiased estimate.

If it is known that θ lies in some interval, $\hat{\theta}$ is chosen to maximize the likelihood function on that interval. It is possible that this maximum will not be a root of the likelihood equation. In that case, the results of Theorem 9 will not, in general, hold. For example, if $\theta \geq 0$ and all roots of the likelihood equation are negative, the two boundary points, $\theta = 0$ and $\theta = \infty$, must be checked to see which maximizes L.

Example

This illustrates how these results can be extended to estimate more than one parameter. Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \theta)$, where both μ and θ are unknown:

$$L(\mu, \theta; x_1, \ldots, x_n) = \frac{1}{(2\pi\theta)^{n/2}} \exp\left[-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \mu)^2\right] ,$$

$$\log L(\mu, \theta; x_1, \ldots, x_n) = C - \frac{n}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{n} x_i^2 + \frac{n\mu}{\theta}\bar{X} - \frac{n\mu^2}{2\theta} ,$$

where C is a constant. Then,

$$\left.\frac{\partial \log L}{\partial \mu}\right|_{\substack{\mu=\hat{\mu}\\ \theta=\hat{\theta}}} = \frac{n\bar{X}}{\hat{\theta}} - \frac{n\hat{\mu}}{\hat{\theta}} = 0 .$$

Therefore, $\hat{\mu} = \bar{X}$. Note that $\hat{\mu}$ is an unbiased estimate of μ for any n. To obtain $\hat{\theta}$,

$$\left.\frac{\partial \log L}{\partial \theta}\right|_{\substack{\mu=\hat{\mu}\\ \theta=\hat{\theta}}} = -\frac{n}{2\hat{\theta}} + \frac{1}{2\hat{\theta}^2}\sum_{i=1}^{n} x_i^2 - \frac{n\bar{X}^2}{\hat{\theta}^2} + \frac{n\hat{\mu}^2}{2\hat{\theta}^2} = 0 .$$

Therefore,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 .$$

$\hat{\theta}$ is only asymptotically unbiased, since the unbiased estimate of θ for any n is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$

2.4.2 Significance tests

Suppose a hypothesis about the property in question (frequently called the null hypothesis, denoted by H_0) is to be checked as to whether H_0 is consistent with some observations that have been or will be taken. A test can be performed by constructing some function of the data, γ, called the test statistic, that has a known distribution if the hypothesis is true and such that, if H_0 is false, an "extreme value" of this distribution would be expected. Then, if an extreme (that is, unlikely if H_0 is true) value is observed, the hypothesis is "rejected." Otherwise, H_0 is "accepted."

The use of the terminology "accept" and "reject" can be misleading. A test of a hypothesis is very one-sided. It only subjects the hypothesis to a process of "disconfirmation." If H_0 is rejected, this may be taken as evidence against the hypothesis. If H_0 is accepted, all that can be said is that it could not be rejected. Acceptance is not evidence for a hypothesis.

Some clarification is needed as to what constitutes an extreme value. Let G denote the set of all possible values of γ, and let A and B be two disjoint subsets of G such that their union is G. The test is

accept H_0 if γ takes a value in A,

reject H_0 if γ takes a value in B.

B is determined, though not uniquely, by the requirement that, if H_0 is true,

$$P[\gamma \text{ takes a value in } B] = \alpha , \qquad (3)$$

where α, called the significance level or type I error, is a predetermined constant, usually 0.05 or 0.01. α is the probability of rejecting the hypothesis when it is, in fact, true. To choose B subject to equation (3), consideration is given to what would be true if H_0 were false. This is usually done by choosing B to minimize the type II error, β. β is the probability of accepting H_0 when it is false. $1 - \beta$ is called the power of the test.*

Alternative terminology, when γ takes a value in B, is to say that the result is significant at the α level. This translates as: the data differ significantly from those that would be expected if H_0 were true, where the significance level is α.

Tests are frequently named by the distribution of the test statistic when H_0 is true.

1 sample

$$X_1, \ldots, X_n \text{ iid } N(\mu, \sigma^2) ,$$

where μ and σ^2 are unknown.

*In general, β is a function of the various alternatives to H_0. Ideally, B is then chosen to minimize β for all the alternatives. If necessary, it is chosen to minimize β for a chosen class of alternatives.

Suppose H_0 states that $\mu = 0$. By Theorem 7, when H_0 is true, $\sqrt{n}\,\bar{X}/s \sim t_{n-1}$. Take $\gamma(x_1, \ldots, x_n) = \sqrt{n}\,\bar{X}/s$.

If $\mu < 0$, we would expect smaller values of γ than if $\mu = 0$. Similarly, if $\mu > 0$, we would expect larger values of γ.

If it is possible that μ may be any value, pick the constant C so that

$$P[|\gamma| \geq C] = \alpha \quad ;$$

then $A = (-C, C)$. If $\alpha = 0.05$ and $n = 10$, then $C = 2.634$, from a table of the t distribution with 9 degrees of freedom.

If it is known that $\mu \geq 0$, then pick C so that

$$P[\gamma \geq C] = \alpha \quad ;$$

Now, $A = (-\infty, C)$. For $\alpha = 0.05$ and $n = 10$, $C = 2.228$. C is chosen differently in this case, since negative values of γ cannot be due to $\mu < 0$.

This test is called the t test, two-sided if all values of μ are possible, and one-sided if only $\mu \geq 0$ (or $\mu \leq 0$) is possible.

As presented so far, the hypothesis testing procedure leads one either to accept or reject the null hypothesis. In practice, though, this is not what is usually done. Instead, the following is done: Collect the data, $x_1, \ldots, x_n$, and calculate the value of the test statistic $\gamma_0 = \gamma(x_1, \ldots, x_n)$. Then calculate the probability of observing a more extreme value of γ than γ_0, using the distribution of γ for the case when the null hypothesis is true. This probability, to be denoted here by p_0, is then a measure of the degree to which the data are inconsistent with the null hypothesis or, equivalently, of the strength of the evidence against the null hypothesis. The smaller the value of p_0, the greater the strength of the evidence against H_0. Therefore, it is usually best simply to state p_0 as the result of the test, although for large enough p_0 (for most purposes, $p_0 > 0.20$ or 0.25 is considered large), it is safe to say that there is no significant evidence against H_0. An alternative to stating p_0 would be to say that the test result is significant at the δ% level, where δ is usually taken as 0.5, 1, 5, 10, or 20 and is such that $\delta/100 \geq p_0$. A result significant at the δ% level is also significant at the δ'% level for any $\delta' > \delta$.

Example

In the previous example, suppose $n = 100$, $\bar{X} = 0.8$, and $s^2 = 16.0$. Then,

$$\gamma_0 = \gamma(x_1, \ldots, x_{100}) = \frac{\sqrt{n}\,\bar{X}}{s} = 2.0 \quad .$$

For the two-sided test,

$$p_0 = P[|\gamma| > 2.0] = 0.046 \quad ,$$

by use of the normal distribution to approximate Student's t distribution with 99 degrees of freedom. This result is significant at the 5% level. For the one-sided test,

$$p_0 = P[\gamma > 2.0] = 0.023 \quad ,$$

again by use of the normal approximation. This result is significant at the 3% level and, hence, also at the 5% level.

If an accept or reject decision is necessary, as when testing for outliers (section 8.2), then an α must be chosen. How does one go about picking an α? That choice depends on the purpose of the test and on the person doing the testing. To choose α, one must decide how much protection is desired against falsely rejecting H_0. The smaller the value of α is, the more protection, but for a given number of observations, reducing α means increasing the probability of falsely accepting H_0 (the type II error). The value of α must then be a personal decision for which there can be no general answer, except to say that $\alpha = 0.05$ and $\alpha = 0.01$ are the most common choices.

For the tests that appear in the remainder of this paper, we will indicate the null hypothesis (usually just referred to as the hypothesis), the test statistic, the distribution of that statistic when H_0 is true, and what constitutes an extreme value of the test statistic. To accomplish the latter, we will say, for example, reject the hypothesis when the test statistic, say Z, is too large. This means that if z_0 is the observed value of Z, then $p_0 = P[Z > z_0]$.

2.5 Glossary of Notation

Capital letters*	Random variables
Small letters*	Values taken on by random variables
S	Sample space
$f(x)$	Density function of X
$F(x)$	Distribution function of X
$P(A)$	Probability of A
$f(x\|y)$ or $f(x\|Y = y)$	Conditional density function of X given Y = y
$E[g(X)]$	Expected value of g(X)
$E[g(X)\|y]$	Expected value of g(X) given Y = y
μ_X	Mean of X
σ_X^2 or Var (X)	Variance of X
σ_X	Standard deviation of X
σ_{XY} or Cov (X, Y)	Covariance of X and Y
ρ_{XY}	Correlation between X and Y
Σ_X	Covariance matrix of the vector of random variables X
$\sim$	"is distributed as"
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\xrightarrow{d}$	Convergence in distribution
iid	"independent and identically distributed"
χ_n^2	Chi-square distribution with n degrees of freedom

*With exceptions.

$\bar{X}$	Sample mean
s^2	Sample variance
t_n	t distribution with n degrees of freedom
$F(n, m)$	F distribution with n and m degrees of freedom
$\hat{}$	Designates estimates

3. THE LEAST-SQUARES MODEL

The general model is

$$Y = f(Z_1, \ldots, Z_p, e)$$

for some function f, where Y is the "dependent variable," that is, the variable that is to be predicted; $Z_1, \ldots, Z_p$ are the "independent variables," that is, the variable that will be used to predict Y; and e is the error or residual term. This includes all errors — for example, in measurement — and all effects — that is, other variables — that are not included in the model. e is a random variable about which we want to make as few assumptions as possible. The form of the model is determined by physical considerations (when known), judgment, and trial and error.

In most of this paper, we will consider a special case called the linear-additive model:

$$Y = \beta_1 X_1 + \ldots + \beta_k X_k + e \quad ,$$

where the X_i's are known functions of the Z_i's, and the β_i's are constants, presumably unknown. The term linear refers to the condition that the model be linear in the coefficients and in the residual term.

Example

1) $Y = \beta_1 \cos \sqrt{3Z} + \beta_2 \exp[Z^2] + \beta_3 \frac{1}{\log Z} + \beta_4 + e$ is linear, with

$X_1 = \cos \sqrt{3Z}$,

$X_2 = \exp[Z^2]$,

$X_3 = 1/\log Z$,

$X_4 = 1$.

2) $Y = \beta_1 Z^{\beta_2} + e$ is not linear unless β_2 is known.

Occasionally it is possible to transform to a linear model. For example, $W = \exp[\beta Z + e]$ can be transformed to

$$Y = \log W = \beta Z + e \quad ,$$

which is linear.

Throughout this paper, assumptions will be made as needed. Once made, all assumptions are to be carried through unless it is otherwise stated.

Assumption 1. The model is correct.

Assumption 2. The X_i's can be observed without error.

4. THE PROBLEM AND ITS SOLUTION

Very simply, the problem is that $\beta_1, \ldots, \beta_k$ are unknown and some estimate of them is needed. Possible reasons for needing the estimates are

1) To test hypotheses about $\beta_1, \ldots, \beta_k$.

2) To be able to predict Y from some future observation on $X_1, \ldots, X_k$.

3) To test the correctness of the model.

To estimate $\beta_1, \ldots, \beta_k$, two things are needed. First, we must have some data. Let us assume that we have n ($n > k$) observations of the $(k + 1)$-component vector $(X_1, \ldots, X_k, Y)$. (If $n \leq k$, the problem is not statistical.) Lower case letters will be used to denote the observed quantities, and the subscript u will be used to denote the number of the observation. Note that the distribution of the residual variable e_u may be different for each u. Anything that is said about e (without a subscript) is to be interpreted as true for each e_u, where applicable.

Second, some criterion is needed. It should come as no surprise that the criterion to be considered here is to minimize the sum of the squared errors; that is, $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are chosen to minimize

$$\sum_{u=1}^{n} (y_u - \hat{y}_u)^2 \quad ,$$

where

$$\hat{y}_u = \sum_{i=1}^{k} \hat{\beta}_i x_{ui} \quad .$$

Hence, the term least squares.

Why this criterion? The two main reasons for using least squares follow:

1) The solution for the linear case we are working with is mathematically easy.

2) The estimates have some nice properties.

Unfortunately, these nice properties sometimes break down, and even when they do not, they are not always optimal.

To keep the notation manageable, we will use matrices:

$$\tilde{Y} = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{pmatrix}$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \cdot & & \\ \cdot & & \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \qquad \epsilon = \begin{pmatrix} e_1 \\ \cdot \\ \cdot \\ e_n \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \cdot \\ \cdot \\ \hat{\beta}_k \end{pmatrix} \qquad \hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \cdot \\ \cdot \\ \hat{y}_n \end{pmatrix} = X\hat{\beta} \quad .$$

Primes denote transposes.

In matrix notation, the model is

$$\tilde{Y} = X\beta + \epsilon \quad ,$$

and $\hat{\beta}$ is chosen to minimize

$$(\tilde{Y} - \hat{Y})'(\tilde{Y} - \hat{Y}) \quad .$$

By taking derivatives and setting them equal to zero, we obtain the normal equations

$$S\hat{\beta} = S_Y \quad ,$$

where S is a $k \times k$ symmetric matrix defined by

$$S_{ij} = \sum_{u=1}^{n} x_{ui}\, x_{uj} \quad \text{or} \quad S = X'X \quad ,$$

and S_Y is a k-component vector defined by

$$S_{Yi} = \sum_{u=1}^{n} x_{ui}\, y_u \quad \text{or} \quad S_Y = X'\tilde{Y} \quad .$$

Then, if S is invertible,

$$\hat{\beta} = S^{-1} S_Y$$

is the least-squares estimate of β. It is important to note that X, and hence S, have been treated as matrices of constants. This is the traditional approach. In effect, the problem is considered in terms of what can be said about Y for given values of the X_i's. Because of this, all expectations that follow are really conditional on X, although this will not be explicitly stated.

Assumption 3. S is nonsingular.

Assumption 4. $E[e] = 0$. As stated above, this assumption is really $E[e \mid X] = 0$. Since this should be true for any value of X, it is necessary that e be uncorrelated with the X_i's.

With Assumption 4, there is another way of looking at the model, since

$$E(Y \mid x_1, \ldots, x_k) = \sum_{i=1}^{k} \beta_i\, x_i \quad . \tag{4}$$

This means that for each $\{x_i\}_{i=1}^{k}$, Y has a distribution about the mean value $\sum_{i=1}^{k} x_i \beta_i$, the distribution being that of the random variable e. The curve

given by formula (4) is called the regression curve (hence the term linear regression), and it is this curve that we wish to estimate.

Theorem 10

If $E(e) = 0$, then $E[\hat{\beta}] = \beta$.

Assumption 5 $E[e_i e_j] = \sigma^2 \delta_{ij}$, where σ^2 is a constant and δ_{ij} is Kronecker's delta: $\delta_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$. That is, the residuals are uncorrelated and have constant variance

The assumption that the residuals are uncorrelated is less restrictive than is an assumption of independence, but there is little practical difference.

This assumption makes the following three theorems possible:

Theorem 11

Let $C = S^{-1}$; then $\Sigma_{\hat{\beta}} = \sigma^2 C$ That is, $\mathrm{Var}(\hat{\beta}_i) = \sigma^2 C_{ii}$, and $\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$.

Theorem 12

Let $d_u = y_u - \hat{y}_u$, the observed residuals, and $s^2 = \frac{1}{n-k} \sum_{u=1}^{n} d_u^2$
Then, $E[s^2] = \sigma^2$

Using these two theorems, we have

$$E\left[\left(Y - \sum_{i=1}^{k} x_i \hat{\beta}_i\right)^2\right] = E[(\Sigma x_i(\beta_i - \hat{\beta}_i) + e)^2] = \sigma^2(1 + X'CX) \quad ,$$

where $X = (x_1, \ldots, x_n)'$ is some future observation. This quantity is the variance of the prediction $\sum_{i=1}^{k} x_i \hat{\beta}_i$. The term $\sigma^2 X'CX$ is due to our not knowing β, and σ^2 is due to the residual term. The standard error of prediction is $s\sqrt{1 + X'CX}$.

Theorem 13 (Gauss-Markoff Theorem)

If we consider only estimates of linear functions of the β_i's that are

1) unbiased and

2) linear functions of the y_u's,

then the least-squares method gives the estimate with minimum variance (for all linear functions of the β_i's).

This last theorem details the nice properties that were promised earlier. It says that the least-squares estimate is best (that is, minimum variance unbiased) in the class of estimates that are linear functions of the y_u's. This is nice, but there is no reason for restricting oneself to this class if a better estimate can be found.

When the model contains a constant term, say β_k, then an alternative method is available. Since the least-squares solution for β_k is

$$\hat{\beta}_k = \overline{Y} - \sum_{i=1}^{k-1} \hat{\beta}_i \overline{X}_i \quad ,$$

the estimates of $\beta_1, \ldots, \beta_{k-1}$ can be obtained by considering the model rewritten as

$$y_u = \overline{Y} + \sum_{i=1}^{k-1} \beta_i (x_{ui} - \overline{X}_i) + e_u \quad ,$$

where

$$\overline{X}_i = \frac{1}{n} \sum_{u=1}^{n} x_{ui}$$

and

$$\overline{Y} = \frac{1}{n} \sum_{u=1}^{n} y_u \quad .$$

If S and S_y are modified to

$$S_{ij} = \sum_{u=1}^{n} (x_{ui} - \overline{X}_i)(x_{uj} - \overline{X}_j) = \sum_{u=1}^{n} x_{ui} x_{uj} - n \overline{X}_i \overline{X}_j \quad ,$$

and

$$S_{Yi} = \sum_{u=1}^{n} (x_{ui} - \overline{X}_i)(y_u - \overline{Y}) = \sum_{u=1}^{n} x_{ui} y_u - n \overline{X}_i \overline{Y} \quad ,$$

then

$$\begin{pmatrix} \hat{\beta}_1 \\ \cdot \\ \cdot \\ \hat{\beta}_{k-1} \end{pmatrix} = S^{-1} S_Y \quad .$$

The advantage of this procedure is that the matrix to be inverted is smaller. Its disadvantage is that it may be neither easy nor accurate to compute S and S_Y. It is necessary either to make one pass through the data to calculate the means and then another pass to calculate S and S_Y or to run the risk of taking the difference of two very large numbers, which could result in nonsense, that is, no remaining significant digits. These problems can be considerable, especially with large quantities of data.

This alternative procedure will not be mentioned again, but there are two points to note. First, the fitted curve goes through the point $(\overline{X}_1, \ldots, \overline{X}_{k-1}, \overline{Y})$. Second, the standard error of prediction is now of the form

$$s \sqrt{1 + \frac{1}{n} + (X - \overline{X})' C (X - \overline{X})} \quad ,$$

where $\overline{X}$ is the $(k-1)$-dimensional vector of means from the original sample. This shows very clearly the price paid for extrapolation in terms of large standard errors.

Note that it has not yet been necessary to assume a distribution for the e_u's. It has not even been assumed that all the e_u's have the same distribution. Some assumption is required for us to be able to make any probability statements about the solution, for example significance tests.

Usually, the normal distribution is assumed, for two main reasons: First, owing to results such as Theorems 1 to 3, the normal distribution is frequently a very good approximation to the residual distribution. Second, the normal distribution is the easiest to work with; that is, tests are available using test statistics with known and tabulated distributions. In theory, it is possible to find tests and the distributions of their test statistics for any assumed distribution of e. In practice, the effort is usually not worth it.

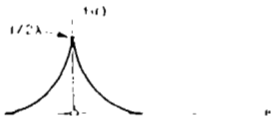
One additional point about normality is that if the e_u's are normally distributed, then the maximum-likelihood estimate is the same as the least-squares estimate. This has two implications. First, the least-squares estimate has, in this case, the additional nice property of being asymptotically minimum-variance unbiased among all estimates, not just among those in the restricted class considered earlier. Second, if the distributions of the e_u's are known and are not normal, it may be preferable to use maximum likelihood rather than least squares.

Example

If $f(e) = \frac{1}{2\lambda} \exp(-\lambda|e|) \quad (\lambda > 0)$

$$= \frac{1}{2\lambda} \exp\left(-\lambda\left|Y - \sum_{i=1}^{k} \beta_i X_i\right|\right) ,$$

the double exponential distribution,



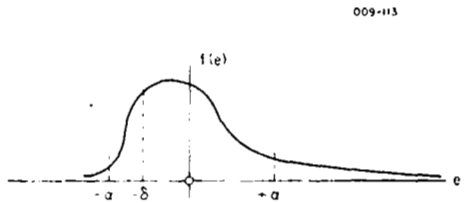
then, maximum likelihood says to choose $\hat{\beta}$ to minimize

$$\sum_{u=1}^{n} |d_u| = \sum_{u=1}^{n} \left|y_u - \sum_{i=1}^{k} \hat{\beta}_i x_{ui}\right|$$

Use of maximum likelihood should especially be considered if the distribution of e is not symmetric. Least squares treats a deviation (d_u) of $-a$ the same as one of $+a$. If the distribution of e were not symmetric, one of these deviations would be more unlikely than the other ($P[e \leq -a] < P[e \geq +a]$ would mean that $-a$ was the more unlikely)

Example

Suppose f(e) looks like:



In this case, $-a$ is more unlikely than $+a$

What is desired is that equally probable deviations be treated alike. In the above example, if $P[e \leq -\delta] = P[e \geq +a]$, then $-\delta$ and $+a$ are equally likely deviations. The quantity to be minimized should take this into account. Least squares does not.

Assumption 6. $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$.

This assumption will be used only for the significance tests that are to follow, unless otherwise stated. Results not connected to a test do not require this assumption.

Theorem 14

If $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$, then

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 C_{ii}) \quad .$$

Also, if P and R are two k-dimensional vectors of constants, then

$$P'\hat{\beta} \sim N(P'\beta, \; P'\Sigma_{\hat{\beta}} P)$$

and

$$\text{Cov}\,(P'\hat{\beta}, \; R'\hat{\beta}) = P'\Sigma_{\hat{\beta}} R \quad .$$

[Remember: $\Sigma_{\hat{\beta}} = \sigma^2 C$.]

PRECEDING PAGE BLANK NOT FILMED

5. A SECOND LOOK AT THE ASSUMPTIONS

The presentation so far has been of the standard statistical approach. Before completing this approach, let us go back and discuss the various assumptions. The questions to be considered are

1) How can it be detected whether the assumptions are valid?

2) What effect does a false assumption have?

3) Can the assumptions be avoided?

Because of their complexity, two topics will be left to the end of the paper: regression when the X_i's are observed with error, and nonlinear regression. In this section, we will work with the linear-additive model and assume that the X_i's are observed without error.

5.1 The Model

The model is $Y = \sum_{i=1}^{k} X_i \beta_i + e$, where e contains all errors in the measurement of Y and the effect of all the variables not included in the model. Assumption 1 was that this model is correct. All the results that have been given and that will follow depend on this assumption, so this assumption is an important one to check. Fortunately, however, all the results remain very nearly valid as long as the model is close to being correct.

What would it take for the model to be incorrect? An incorrect model can be characterized by a correlation between the d_u's and one of the variables. This correlation can be caused by the following:

1) The X_i's being used are the wrong functions of the Z_i's.

2) A variable, correlated with those being used, has been left out of the model.

3) The true regression equation is not linear.

A lack of fit due to the nonlinearity of the true regression can sometimes be eliminated by restricting the X_i's to a range on which the linear model is a better approximation. If this cannot be done, it is necessary to use a nonlinear procedure.

The two principal methods for checking this assumption follow.

5.1.1 A test for goodness of fit

This test assumes that for each $(x_{u1}, \ldots, x_{uk})$, n_u $(n_u > 1)$ observations of Y have been taken: $y_{u1}, \ldots, y_{un_u}$. Then, for each u

$$s_u^2 = \frac{1}{n_u - 1} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2 ,$$

where

$$\bar{y}_u = \frac{1}{n_u} \sum_{v=1}^{n_u} y_{uv} ,$$

will always be an unbiased estimate of σ^2, whether the model is correct or not. By combining over u,

$$s_2^2 = \frac{\sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2}{\sum_{u=1}^{n} (n_u - 1)}$$

will be an unbiased estimate of σ^2. By Theorem 12,

$$s_1^2 = \frac{\sum_{u=1}^{n} n_u (\bar{y}_u - \hat{y}_u)^2}{n - k}$$

will be an unbiased estimate of σ^2 if the model is correct. If the model is wrong, s_1^2 will be inflated by the difference between the fitted regression line and the true regression line. s_1^2 is called the lack-of-fit term and is obtained by treating $\bar{y}_u$ as the observation of Y corresponding to $(x_{u1}, \ldots, x_{uk})$ and then weighting inversely to the variance (which will be covered later). Also, s_1^2 is independent of s_2^2. This does not constitute a complete proof, but to see that s_1^2 and s_2^2 are independent, consider

$$\sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \hat{y}_u)^2 = \sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u + \bar{y}_u - \hat{y}_u)^2$$

$$= \sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2 - 2 \sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)(\bar{y}_u - \hat{y}_u)$$

$$- 2 \sum_{u=1}^{n} (\bar{y}_u - \hat{y}_u) \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u) + \sum_{u=1}^{n} n_u (\bar{y}_u - \hat{y}_u)^2 ,$$

$$= \sum_{u=1}^{n} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2 + \sum_{u=1}^{n} n_u (\bar{y}_u - \hat{y}_u)^2$$

and use the following lemma:

Lemma

If $g_1, \ldots, g_n$ iid $N(0, \sigma^2)$ and $v < n$, then

$$T_1 = \frac{1}{\sigma^2} \sum_{i=1}^{v} g_i^2 \sim \chi_v^2 ,$$

$$T_2 = \frac{1}{\sigma^2} \sum_{i=v+1}^{n} g_i^2 \sim \chi_{n-v}^2 ,$$

and T_1 and T_2 are independent.

Before the test is stated, a discussion of the degrees of freedom of variance estimates is necessary. In general, a variance estimate is of the form $S^2 = \frac{1}{f}\sum_{j=1}^{m} g_j^2$, where g_j is some function of the j^{th} observation and $m \geq f$. f is a constant equal to the number of independent observations minus the number of parameters estimated by those observations. As an example, for s_1^2, the $(y_u)_{u=1}^n$ are n independent observations and k parameters have been estimated $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$, so $f = n - k$. In that case, $g_j = \sqrt{n_j}\,(\bar{y}_j - \hat{y}_j)$. f is the number of degress of freedom of the variance estimate. If the distribution of g_j's is $N(0,\sigma^2)$, as is the case for s_1^2 when the model is correct and Assumption 6 holds, then

$$\frac{S^2}{\sigma^2} \cdot f = \frac{\sum_{j=1}^{m} g_j^2}{\sigma^2} \sim \chi_f^2$$

(The number of degrees of freedom of the χ^2 distribution is f not m, as might at first be expected, because the g_j's are dependent.)

By this result and the lemma,

$$\frac{s_1^2}{s_2^2} \sim F\left(n - k, \sum_{u=1}^{n}(n_u - 1)\right)$$

if the model is correct. If the model is wrong, a large value of the test statistic is expected. The test is to reject the hypothesis that the model is correct if

$$\frac{s_1^2}{s_2^2} \geq C \quad ,$$

where C is obtained from a table of the $F\left(n - k, \sum_{u=1}^{n}(n_u - 1)\right)$ distribution for the chosen significance level.

If the result of the test is not significant, there is little need to worry about a lack of fit. All results will at least be very close to being completely valid. If the result is significant, some other procedure must be used to determine the cause of the lack of fit so that it can be corrected.

If multiple observations are not available, an estimate of σ^2 from another set of data can be used as the denominator of the F test replacing $\sum_{u=1}^{n}(n_u - 1)$ by the appropriate number of degrees of freedom. The only requirements that this estimate must satisfy are that it be unbiased, whether the model is correct or not, and that it be independent of s_1^2.

5.1.2 Residual analysis

Residual analysis will be used to check goodness of fit if the F test cannot be performed or to try to discover the cause of the lack of fit if the test result was significant. Residual analysis has other uses, so it is best to start with a general overview of the procedure before going into the specifics of this application.

The basic idea behind residual analysis is that if the assumptions are correct, the $(e_u)_{u=1}^n$ are n uncorrelated random variables (possibly normally distributed) with mean 0 and variance σ^2. The $(d_u)_{u=1}^n$, being estimates of $(e_u)_{u=1}^n$, should then look like a sample with those properties. In fact, the d_u's have the covariance matrix $\sigma^2(I - X(X'X)^{-1}X')$ $(\neq \sigma^2 I)$ and so are correlated, but this effect is negligible unless the ratio $(n - k)/n$ is very small. Therefore, the d_u's should appear to be uncorrelated, to have constant variance, and to be uncorrelated with any of the variables in the model.

Usually, a residual analysis will give some idea of which assumptions, if any, are not valid and how, if necessary, the estimates can be corrected. The procedure is to examine plots of residuals, first overall (for example, as a histogram), and then against

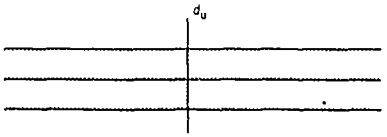
1) Time, if known.
2) $\hat{Y}$
3) X_j $(j = 1, \ldots, k)$.
4) Anything else that makes sense for a particular problem. For example, if the observations come from three different stations, the residuals for each station could be plotted separately.

When the residuals are being examined for goodness of fit, the following should be looked for:

1) Plot against $\hat{Y}$ or X_j $(j = 1, \ldots, k)$.

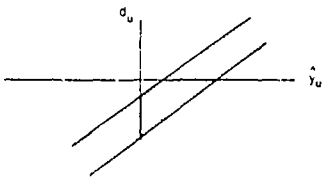
The residuals should lie in a horizontal band:

009-113



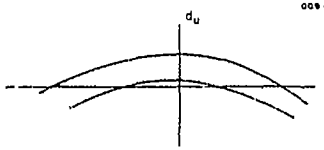
If they do not, something is wrong. For example,

009-113



If this occurs in a plot against $\hat{Y}$, it indicates that a constant term was left out. In a plot against some X_j, it indicates an error in the calculations.

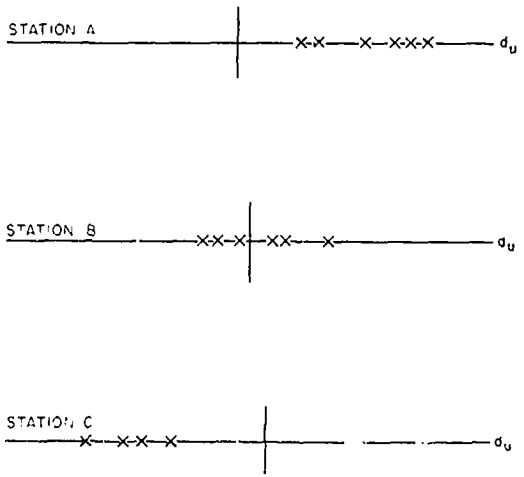
009-113



If this is a plot against X_j, it indicates that an X_j^2 term is needed; if against $\hat{Y}$, that some variable needs to be added to the model.

2) As an example of other possible plots, for the plot of the residuals by station, all three plots should look alike. If something like the following happens,

009-113



it would indicate systematic differences between the observations from the three stations. This could be corrected by introducing, by use of indicator variables, an additional constant term for each station:

$$X_{uA} = \begin{cases} 1 & u^{th} \text{ observation from station A} \\ 0 & \text{otherwise} \end{cases},$$

and similarly for X_{uB} and X_{uC}.

3) Plot against time.

Again, the residuals should lie in a horizontal band. If they do not, something not in the model is changing over time.

It is possible to test the randomness of the pattern of the sign of the residuals. The test is called the sign test and does not require the normality assumption.

Start by counting the number of runs. For example,

+ + - + + + - - -

has four runs. The test is to reject the hypothesis of randomness if there are too few runs. For small n, a special table (such as in Draper and Smith, 1966) must be referred to for the distribution of the test statistic. For large n, the following normal approximation can be used. Let

n_1 = number of positive signs,

n_2 = number of negative signs and

W = number of runs.

Then, if $n_1 > 10$ and $n_2 > 10$,

$$Z = \frac{W - \mu + 1/2}{\sigma} \sim N(0, 1) \text{ (approximately)} ,$$

where

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

and

$$\sigma^2 = \frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} .$$

Reject the hypothesis if Z is too small.

A significant result could possibly be due to some uncontrolled variable changing values. In particular, the magnitude of a systematic error may be changing.

5.2 The Solution

The assumption that S is nonsingular is not generally a problem. S will be singular if there are some linear relations among the X_i's. In that case, the normal equations will have an infinite number of solutions, all of which are equivalent in the sense that they give exactly the same predicting equation. There are two ways of handling this problem.

First, if there are ℓ linear relations among the X_i's, you can either drop ℓ of the X_i's or introduce ℓ constraints on the β_i's.

Example (Cochran, 1969)

Let n = 4, suppose $x_{u3} = x_{u2} - x_{u1}$, u = 1, ..., 4, and let

$$S = \begin{pmatrix} 14 & 39 & 25 \\ 39 & 158 & 119 \\ 25 & 119 & 94 \end{pmatrix}$$

and

$$S_Y = \begin{pmatrix} 72 \\ 234 \\ 162 \end{pmatrix} .$$

Solution 1: Put $\hat{\beta}_3 = 0$ and solve for $\hat{\beta}_1$ and $\hat{\beta}_2$ (this is equivalent to dropping X_3).

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 14 & 39 \\ 39 & 158 \end{pmatrix}^{-1} \begin{pmatrix} 72 \\ 234 \end{pmatrix} = \begin{pmatrix} \frac{2250}{691} \\ \frac{468}{691} \end{pmatrix} .$$

Then,

$$\hat{y}_u = \frac{1}{691}(2250\, x_{u1} + 468\, x_{u2})$$

Solution 2: Put $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 0$. Eliminate any one of them from the normal equations and solve:

$$\hat{y}_u = \frac{1}{691}(-468\, x_{u1} + 3186\, x_{u2} - 2716\, x_{u3})$$

$$= \frac{1}{691}(2250\, x_{u1} + 468\, x_{u2})$$

The second method is to find a generalized inverse of S, that is, a matrix, S^g, that satisfies

$$S\, S^g\, S = S$$

At least one generalized inverse exists for any matrix. If S^g can be found,

$$\hat{\beta} = S^g\, S_y$$

is the least-squares solution. Rao (1965) and Graybill (1969) discuss methods for finding S^g. Note that $S^g = S^{-1}$ if S^{-1} exists.

The function $P'\beta$ of the β_i's is said to be <u>estimable</u> if there exists an n-component vector R such that

$$E[R'\,\tilde{Y}] = P'\beta \quad ,$$

that is, if there exists a linear combination of the y_u's that is an unbiased estimate of $P'\beta$. If there are linear relations among the X_i's, not all linear combinations of the β_i's are estimable, in contrast to the case of no linear relations, where all linear combinations of the β_i's are estimable.

<u>Theorem 15</u>

1) $P'\beta$ is estimable if and only if $P'(I - S^g S) = 0$, where S^g is any generalized inverse of S.
2) If there exists an $\ell \times k$ ($\ell < k$) matrix G such that $G X' = 0$, then $P'\beta$ is estimable if and only if $G P = 0$. (G is the matrix of the coefficients of the ℓ linear relations among the X_i's.)
3) If $\hat{\beta}$ is any solution of the normal equations and $P'\beta$ is estimable, then its unique estimate is $P'\hat{\beta}$. This means that if $P'\beta$ is estimable, there is exactly one linear combination of the y_u's that is an unbiased estimate of $P'\beta$, namely $P'\hat{\beta}$.

<u>Example</u>

In the previous example, $G = (1, -1, 1)$, so $P_1\beta_1 + P_2\beta_2 + P_3\beta_3$ is estimable if and only if $P_3 = P_2 - P_1$. For instance, $\beta_1 + \beta_2$ and $\beta_1 - \beta_3$ are estimable, but $\beta_1 - \beta_2$ is not. Note also that, in this case, neither β_1, β_2, nor β_3 is estimable!

A more common problem is that S may be ill-conditioned, that is, nearly singular. Finding S^{-1} is then a problem in numerical analysis.

One possible method, due to Householder, uses orthogonal transformations. The problem is to find an $n \times n$ orthogonal matrix Q such that

$$QX = R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \quad ,$$

where $\tilde{R}$ is a $k \times k$ upper triangular matrix. Then,

$$\hat{\beta} = \tilde{R}^{-1}\, \underline{Y} \quad ,$$

where $\underline{Y}$ is the vector consisting of the first k components of $Q\tilde{Y}$. An iterative procedure for finding $\hat{\beta}$ based on this method is detailed in two papers (Golub, 1965; Golub and Businger, 1965) Golub claims that his procedure will also find $\hat{\beta}$ when S is singular

Since S is a symmetric, positive-definite matrix, another possible method would be to use the Cholesky decomposition of S:

$$S = R'\, R \quad ,$$

where R is an upper triangular, $k \times k$ matrix. Then,

$$S^{-1} = R^{-1}(R^{-1})' \quad .$$

Once R is found, S^{-1} is easy to find since it is simple to invert triangular matrices. Golub (1969) discusses this method and others, including the modified Gram-Schmidt orthogonalization procedure. Golub also provides a good bibliography on this topic.

Some work has been done on comparing various methods for inverting S (see Jordan, 1968; Rice, 1966, Wampler, 1969).

5.3 The Residuals

Assumption 4 was that $E[\epsilon] = 0$. If the model contains a constant term, this assumption will always be effectively valid. Suppose that $E[\epsilon] = \mu_e \neq 0$ and that β_k is the constant term. Then,

$$Y = \sum_{i=1}^{k-1} X_i \beta_i + \beta_k + \epsilon \qquad (5)$$

$$= \sum_{i=1}^{k-1} X_i \beta_i + (\beta_k + \mu_e) + (\epsilon - \mu_e)$$

$$= \sum_{i=1}^{k-1} X_i \beta_i + \beta_k^* + \epsilon^* \quad , \qquad (6)$$

where $\beta_k^* = \beta_k + \mu_e$ and $\epsilon^* = \epsilon - \mu_e$. Since $E[\epsilon^*] = 0$, the model has been transformed so as to satisfy the required condition. The least-squares procedure will go ahead as if (6) were the correct model instead of (5), and all the estimates except for the constant term will be unbiased. This could be taken as an argument for always including a constant term in the model, since if there is no constant term and if $\mu_e \neq 0$, all the estimates will be biased by some unknown amount. The estimate of β_k^* will have mean

$$E[\hat{\beta}_k^*] = \beta_k + \mu_e \quad .$$

The assumption that $E[\epsilon_i \epsilon_j] = \sigma^2 \delta_{ij}$ is unnecessary. Suppose that Σ_e is the covariance matrix of the ϵ_u's:

$$\Sigma_e = E[\epsilon\epsilon'] \quad .$$

Then,

$$\tilde{Y} = X\beta + \epsilon$$

can be transformed to*

$$\Sigma_e^{-1/2} \tilde{Y} = \Sigma_e^{-1/2} X\beta + \Sigma_e^{-1/2} \epsilon$$

or

$$W = Z\beta + g \quad .$$

Then,

$$\Sigma_g = E[g g'] = E[\Sigma_e^{-1/2} \epsilon\epsilon' \Sigma_e^{-1/2}] = I \quad .$$

So,

$$E[g_i g_j] = \delta_{ij} \quad ,$$

and the g_u's satisfy Assumption 5 with $\sigma^2 = 1$. Then, by minimizing $(W - \hat{W})' (W - \hat{W})$,

$$\hat{\beta} = (Z'Z)^{-1} Z'W$$

$$= (X' \Sigma_e^{-1} X)^{-1} X' \Sigma_e^{-1} \tilde{Y} \quad ,$$

$$\Sigma_{\hat{\beta}} = (X' \Sigma_e^{-1} X)^{-1} \quad ,$$

and

$$E[s^2] = 1 \quad ,$$

where

$$s^2 = \frac{1}{n-k} (W - \hat{W})' (W - \hat{W}) = \frac{1}{n-k} (Y - \hat{Y})' \Sigma_e^{-1} (Y - \hat{Y}) \quad .$$

The Gauss-Markoff Theorem applies to $\hat{\beta}$ obtained this way. In the case where Σ is diagonal (that is, where there are uncorrelated errors), this is called weighting inversely to the variance. Since if $\mathrm{Var}(\epsilon_u) = \sigma_u^2$, then

*$\Sigma_e^{1/2}$ is the unique positive definite square root of Σ_e, and $\Sigma_e^{-1/2}$ is its inverse.

$$\Sigma_e = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

and the quantity to be minimized is

$$\sum_{u=1}^{n} \frac{1}{\sigma_u^2} (y_u - \hat{y}_u)^2 \quad .$$

Unfortunately, this requires either that Σ_e be known or that an estimate of it is available. Even when not known, the relative magnitudes may be known, especially in the diagonal case. Then,

$$\Sigma_e = \sigma^2 G \quad ,$$

where σ^2 is an unknown constant and G is a known symmetric positive-definite matrix. By transforming with $G^{-1/2}$, we obtain

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

$$\Sigma_{\hat{\beta}} = \sigma^2 (X' G^{-1} X)^{-1} \quad ,$$

and

$$E\{s^2\} = \sigma^2 \quad ,$$

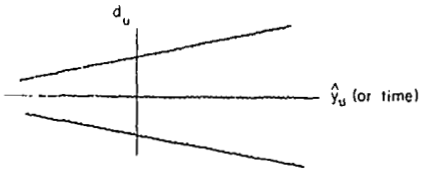
where

$$s^2 = \frac{1}{n - k} (Y - \hat{Y})' G^{-1} (Y - \hat{Y}) \quad .$$

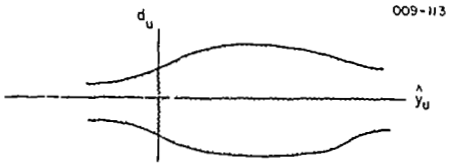
What happens if the weight matrix is not used when it should be? Consistent unbiased estimates of the β_i's will still be obtained, but they will not be the minimum-variance estimates and s^2 will not be an unbiased estimate of σ^2. This effect is small, however, if the correlations between the e_u's are very small and if the σ_u^2's do not vary greatly.

Two methods can be used to detect that the variances are not constant. The first is residual analysis. Deviations from constant variance are characterized by deviations from the horizontal band that are symmetric about the $\hat{y}_u$ (or time) axis. For example,

009-112



This indicates that σ_u^2 increases with $\hat{y}_u$. The following is also evidence of nonconstant variance.



In many cases, a nonconstant variance indicates that another variable should be included in the model. If it is possible to determine what that variable should be, it would be preferable to include it in the model rather than trying to estimate the Σ_e matrix.

An approximation that is available if needed is

$$\sigma_u^2 \propto g(\hat{y}_u) \text{ (approximately)},$$

or

$$\sigma_u^2 \propto g(\text{time}) \quad ,$$

where g is some function and $\propto$ means proportional to. In the first example above,

$$\sigma_u^2 \propto \hat{y}_u \quad .$$

This would require two least-squares solutions: the first to determine $\hat{Y}$ from unweighted estimates of β, and the second to determine the weighted estimates of β using this approximation to the G matrix.

If multiple observations on Y are available for at least some of the $(x_{u1}, \ldots, x_{uk})$, it is possible to make Bartlett's test for homogeneity of variances.

Suppose that for $u = 1, \ldots, m \leq n$, there are $n_u > 1$ observations on Y corresponding to $(x_{u1}, \ldots, x_{uk})$. Let

$$s_u^2 = \frac{1}{n_u - 1} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2 \quad ,$$

$$\bar{s}^2 = \frac{\sum_{u=1}^{m} \sum_{v=1}^{n_u} (y_{uv} - \bar{y}_u)^2}{\sum_{u=1}^{m} (n_u - 1)} = \frac{\sum_{u=1}^{m} (n_u - 1)\, s_u^2}{\sum_{u=1}^{m} (n_u - 1)} \quad ,$$

$$M = \left[\sum_{u=1}^{m} (n_u - 1) \right] \log \bar{s}^2 - \sum_{u=1}^{m} \left[(n_u - 1) \log s_u^2 \right] \quad ,$$

and

$$C = 1 + \frac{1}{3(m-1)} \left[\sum_{u=1}^{m} \frac{1}{n_u - 1} - \frac{1}{\sum_{u=1}^{m} (n_u - 1)} \right] .$$

Then, if the hypothesis of equal variances is correct,

$$\frac{M}{C} \sim \chi^2_{m-1} \text{ (approximately)} \quad ,$$

and the test is to reject the hypothesis if M/C is too large. This will test only the equality of variances at those points for which the multiple observations were available.

There are two problems with this test:

1) It is very sensitive to departures from normality.

2) The χ^2 approximation is not very good if $n_u \leq 6$, although special tables of the distribution do exist for that case.

5.4 Normality

The effect of any departure from normality is that the actual significance levels of any tests used are different from the stated values. For the F and t tests, if the departure from normality is not large, when a 5% test is stated, the real significance level will be on the order of 7 to 10%. As a general rule, F and two-sided t tests are less affected by departures from normality than is the one-sided t test. The one-sided test is strongly affected by any skewness (that is, departure from symmetry) of the e distribution.

There are two reasonable methods for checking normality. The first is to construct a histogram of the observed residuals d_u. This is then compared to the histogram that would be expected if the e_u were normally distributed, with the estimated variance s^2 being used. There are two methods of making this comparison. The first is the χ^2 goodness-of-fit test. Let

$$G = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \quad ,$$

where m is the number of classes in the histogram, O_i is the number observed in the i^{th} class, and E_i is the number expected in the i^{th} class. Then, approximately,

$$G \sim \chi^2_{m-2} \quad ,$$

and the test is to reject the hypothesis of normality if G is too large.* For the χ^2 approximation to be good, the distribution of e should be near normal and the classes should be chosen so that $E_i \geq 1$ $(i = 1, \ldots, m)$.

The second method of comparison is the Freeman-Tukey test. Let

$$V = \sum_{i=1}^{m} (\sqrt{O_i} + \sqrt{O_i + 1} - \sqrt{4E_i + 1})^2 \quad ,$$

using the same notation as above. The approximate distribution of V is the same as for G. The test is to reject if V is too large. This test does not require that the distribution of e be near normal and is also much less sensitive to small values of E_i. (When any E_i is small, a small change in that E_i can result in a very large change in G.)

The second method for checking normality is a normal plot. Let Φ denote the standard normal distribution function; that is, if $W \sim N(0, 1)$, then $P[W \leq w] = \Phi(w)$. Suppose the residuals $d_1, \ldots, d_n$ are ordered from smallest to largest. Then, plot d_u versus $\Phi^{-1}(u/(n + 1))$.† A sample from $N(\mu, \sigma^2)$ will lie on the line through $(\mu, 0)$ with slope $1/\sigma$. Special paper is available (for example, from Keuffel & Esser) that takes care of the Φ^{-1}, so that only d_u versus $u/(n + 1)$ (or $(3u - 1)/(3n + 1)$) need be plotted.

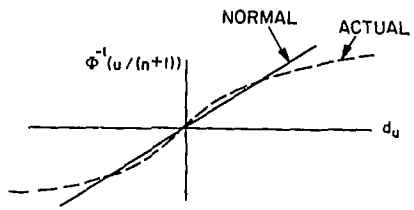
Unfortunately, this plot is not very sensitive to small departures from normality, but it should show if something really horrible is happening, as in the following sketches:

*In general, $G \sim \chi^2_f$, where f = (number of classes) - (number of estimated parameters) - (number of constraints on the E_i). In this case, σ^2 is estimated and the E_i are constrained by $\sum_{i=1}^{m} E_i = n$.

†Anscombe and Tukey (1963) recommend the use of $\Phi^{-1}[(3u - 1)/(3n + 1)]$.

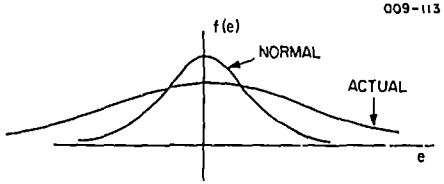
20

009-113

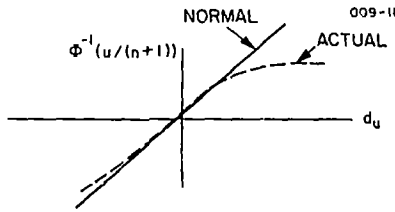


This implies large tails, that is,

009-113

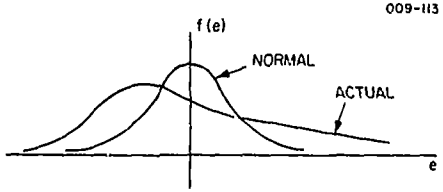


009-113



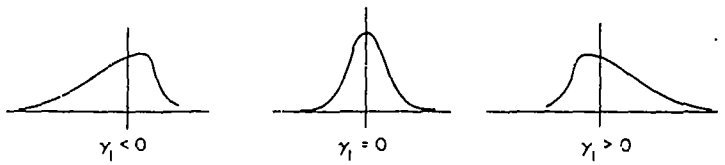
This implies skewness:

009-113



Another possible test for normality, which is not so easy, is to test for skewness and kurtosis. Suppose W is a random variable with mean μ and standard deviation σ. Let $\gamma_1 = (1/\sigma^3)E[(W-\mu)^3]$ and $\gamma_2 = (1/\sigma^4)E[(W-\mu)^4] - 3$. Then γ_1 is a measure of the skewness, that is, the departure from symmetry, of the distribution of W:

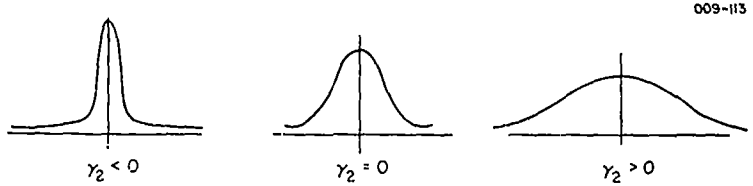
009-113



FOR EXAMPLE,
NORMAL DISTRIBUTION

γ_2 is a measure of the kurtosis, that is, the flatness or peakedness, of the distribution of W:

009-113



FOR EXAMPLE,
NORMAL DISTRIBUTION

With the sample estimates of these measures, a test for equality to zero could be made.* A significant result would imply nonnormality. The calculations, especially of the standard errors, are quite complicated. See Anscombe (1961) for details.

*The reader should be warned that there are two definitions of kurtosis. They differ by the constant 3, so some references may give 3 as the kurtosis for the normal distribution.

6. TESTING HYPOTHESES ABOUT THE REGRESSION COEFFICIENTS

Now that the problems with the assumptions have been considered, we will again make all six assumptions; that is,

$$y_u = \sum_{i=1}^{k} x_{ui}\beta_i + e_u \quad (u = 1, \ldots, n) \quad ,$$

$$e_1, \ldots, e_n \text{ are iid } N(0, \sigma^2) \quad .$$

The three most frequent hypotheses are

1) $\beta_i = W$ (for some i) .

2) $\beta_{r+1} = W_{r+1}, \ldots, \beta_k = W_k$.

3) $w_{i1}\beta_1 + \ldots + w_{ik}\beta_k = W_i \quad (i = 1, \ldots, g \leq k)$,

where the w_{ij}'s and W_i's are known.

The procedure for testing 3), and hence 1) and 2), is the following:

1) Fit the regression without the conditions and calculate $R_k = \sum_{u=1}^{n} \hat{y}_u^2$ and s^2.

2) Refit the regression subject to the conditions to be tested and calculate $R_{k-g} = \sum_{u=1}^{n} \hat{\hat{y}}_u^2$. ($\hat{\hat{y}}_u = \sum_{i=1}^{k} x_{ui}\hat{\hat{\beta}}_i$, where the $\hat{\hat{\beta}}_i$'s are the estimates of the β_i's subject to the g conditions.) Then if the hypothesis is true,

$$H = \frac{R_k - R_{k-g}}{g\, s^2} \sim F(g, n - k),$$

and the test is to reject the null hypothesis if H is too large. The distribution of H is the result of the following theorem (by redefining the β_i's):

Theorem 16

If $\beta_{k-g+1} = \ldots = \beta_k = 0$, then,

1) $(1/\sigma^2)(R_k - R_{k-g}) \sim \chi^2_g$,

2) $(1/\sigma^2)(\Sigma y_u^2 - R_k) = (n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$,

3) they are independent.

Therefore,

$$\frac{R_k - R_{k-g}}{g\, s^2} \sim F(g, n - k) \quad .$$

The statistic R_k is called the reduction in the sum of squares due to regression. The subscript indicates the number of independent parameters estimated. A useful identity is $R_k = \sum_{u=1}^{n} y_u^2 - \sum_{u=1}^{n} d_u^2$.

An equivalent method for testing a hypothesis of the form 3) when g = 1 is the following. Let the hypothesis be that $P'\beta = W$. By Theorem 14,

$$P'\hat{\beta} \sim N(P'\beta, \sigma^2 P'CP) \quad .$$

So, if the hypothesis is true,

$$A = \frac{P'\hat{\beta} - W}{s\sqrt{P'CP}} \sim t_{n-k} \quad .$$

The test is to reject the hypothesis if $|A|$ is too large (a two-sided test).

This is exactly equivalent to the F test, as can be shown by proving that $A^2 = (R_k - R_{k-1})/s^2$ and using Theorem 8.

A test we will be using later is that of $\beta_i = 0$ for some i. In that case,

$$A^2 = \frac{\hat{\beta}_i^2}{s^2 C_{ii}} = \frac{R_k - R_{k-1}}{s^2} \quad .$$

If the null hypothesis $\beta_i = 0$ is rejected by this test, the corresponding variable, X_i, will be said to be significant.

7. CHOOSING A REGRESSION EQUATION

The two questions to be considered in this section are

1) Which of two or more competing equations (models) is best?

2) If the model must be simplified because of limitations on cost and space, which variables are to be dropped?

The criteria to be considered are conflicting:

1) As many variables as possible are wanted so that the predictions are good.

2) As few variables as possible are wanted so that cost and space problems can be avoided.

It is easier to answer the second question first, so we will begin there. Assume that there exists a list of variables $X_1, \ldots, X_q$ from which some number (not necessarily decided on in advance) need to be selected for use in a regression equation. A number of procedures have been developed to solve this problem. Unfortunately, they do not always lead to the same solution.

Before these methods are discussed, two more statistics must be introduced. Let

$$r^2 = \frac{\sum_{u=1}^{n} \hat{y}_u^2 - n\bar{Y}^2}{\sum_{u=1}^{n} y_u^2 - n\bar{Y}^2} = \frac{\sum_{u=1}^{n} (\hat{y}_u - \bar{Y})^2}{\sum_{u=1}^{n} (y_u - \bar{Y})^2} .$$

The square of the sample multiple-correlation coefficient is r^2, which is equal to the square of the sample correlation between Y and $X\hat{\beta}$. It is interpreted as the percentage of the variation in the sample that is explained by the fitted regression. In general, it is desirable to maximize r^2. [The sample correlation between two variables, V and W, is given by

$$r_{VW} = \frac{\sum_{u=1}^{n} (v_u - \bar{V})(w_u - \bar{W})}{\left[\sum_{u=1}^{n} (v_u - \bar{V})^2 \sum_{u=1}^{n} (w_u - \bar{W})^2\right]^{1/2}} ,$$

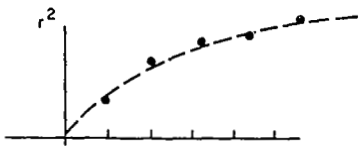
where n is the number of observations.]

The methods are as follows:

1) All possible regressions.

Compute all possible regressions. For each number of variables used in the regression, pick the one that maximizes r^2. This gives the following curve:

009-113



You must then decide where on this curve you would like to be, a nonstatistical decision.

This method can be too much work to be feasible, especially for a large number of variables. (k variables imply 2^k different regressions.) On the other hand, it is the only method guaranteed to give the best regression (in terms of maximum r^2) for the number of variables used.

2) Backward elimination.

Start by computing the regression with all variables and then successively eliminate them, calculating the new regression after each elimination. The criterion for elimination is to pick the one with the smallest value of $\hat{\beta}_i^2 / s^2 C_{ii}$ (the test statistic for the F test that $\beta_i = 0$). Stop when all remaining variables test as being significant at some chosen significance level.

This is a good procedure if the regression with all variables can be handled.

3) Forward selection.

Start with the variable that is the most highly correlated (in absolute value) with Y in the sample. Then insert other variables in turn. The criterion for choosing the next variable to be entered is a bit complicated. Suppose that $X_1, \ldots, X_j$ have already been entered. The next variable to be entered is the one that maximizes the square of the partial-correlation coefficient with Y while controlling for $X_1, \ldots, X_j$ (we will return to this). This is equivalent to finding the variable that maximizes $R_{j+1} - R_j$ and hence causes the largest increase in r^2. This procedure is stopped when the last variable entered tests as not being significant, or when a satisfactory value of r^2 is obtained.

This procedure is usually more economical than backward elimination, but it can be improved upon.

4) Stepwise regression.

This is the same as forward selection except that after fitting a new regression, look back at the variables that had been included. If any of them test as not being significant, throw out the one that is least significant (smallest F value).

The partial-correlation coefficient mentioned earlier can be very difficult to compute, especially for $j > 1$.* Draper and Smith (1966) present an algorithm for stepwise regression that greatly simplifies the computational problems.

For most problems, this is the best method. It is an improvement over forward selection since it does not retain variables that are no longer significant.

*For $j = 1$, the formula for the partial correlation between X_2 and Y, controlling for X_1, is

$$r_{X_2, Y \cdot X_1} = \frac{r_{X_2 Y} - r_{Y X_1} r_{X_2 X_1}}{\sqrt{(1 - r^2_{X_1 Y})(1 - r^2_{X_1 X_2})}} ,$$

For whatever method employed, it is useful first to compute the regression with all the variables (if possible). This will tell how large r^2 can become. It is also a good idea to use a large α for the tests. This forces more variables into the equation and hence leaves some leeway to throw out particularly bothersome variables.

Turning now to the first question, it is not one that can be answered entirely by statistics. The only thing that can be said statistically is to pick the equation that maximizes r^2. However, this does not take into account the number of variables used. A better procedure, especially for small n, would be to pick the one that maximizes

$$\hat{R}^2 = 1 - (1 - r^2) \frac{n - 1}{n - k - 1} ,$$

where k is the number of variables. $\hat{R}^2$ is the unbiased estimate of the population multiple-regression coefficient R^2. R^2 is the portion of the variation in Y that can be explained by the true regression and is equal to the square of the correlation between Y and $\sum_{i=1}^{k} X_i \beta_i$. Of course, owing to sampling variation, maximizing $\hat{R}^2$ does not necessarily maximize R^2, but there is nothing that can be done about that.

Still, maximizing r^2 or $\hat{R}^2$ does not take into account various costs, such as that of obtaining data. Tradeoff between cost and the number of variables must be decided by the user.

8. OTHER TOPICS

8.1 Constraints

In many cases, it may be known that the β_i's must satisfy certain constraints. For example, if the k β_i's are functions of ℓ γ_j's ($\ell < k$), which are the quantities of interest, there will be $k - \ell$ constraints on the β_i's.

If the $\hat{\beta}_i$'s are to satisfy the same constraints as the β_i's, or if constraints are to be imposed on the $\hat{\beta}_i$'s in order to test a hypothesis, the ordinary least-squares solution will not work. Suppose the constraints are consistent and linear; that is, they can be written in the form $G\beta = D$, where G is an $r \times k$ matrix of rank r, D is an $r \times 1$ vector, and both G and D are known. The assumption that G has rank r eliminates redundant constraints. Two methods of handling this problem will be considered. If the constraints are not linear, some nonlinear procedure must be used.

1) Lagrange multipliers.

This method finds the $\hat{\beta}$ that minimizes $\sum_{u=1}^{n} (y_u - \hat{y}_u)^2$ subject to $G\hat{\beta} = D$. The solution, assuming without loss of generality that $\Sigma_e = I$, is obtained by minimizing

$$(\tilde{Y} - X\beta)'(\tilde{Y} - X\beta) + 2(G\beta - D)'\lambda \quad ,$$

where λ is the $r \times 1$ vector of Lagrange multipliers. The factor of 2 is used only to simplify the calculations. The normal equations then become

$$\begin{pmatrix} S & G' \\ G & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} S_Y \\ D \end{pmatrix} \quad ,$$

and the solution is

$$\hat{\beta}_L = \hat{\beta}_0 - S^{-1} G' \hat{\lambda} \quad ,$$

where

$$\hat{\beta}_0 = S^{-1} S_Y \text{ (the usual least-squares estimate)} \quad ,$$

$$\hat{\lambda} = (G S^{-1} G')^{-1} \delta \quad ,$$

and

$$\delta = G\hat{\beta}_0 - D \quad ,$$

assuming that $(G S^{-1} G')^{-1}$ exists. This assumption is reasonable since $G S^{-1} G'$ is an $r \times r$ matrix, G is of rank r, and S is assumed to be of rank $k > r$. If the inverse does not exist, a generalized inverse, $(G S^{-1} G')^g$, can be used.

Theorem 17

If $G\beta = D$ is true, then

1) $E[\hat{\beta}_L] = \beta$,

2) $E\left[\frac{1}{n - k + r}\left(\sum_{u=1}^{n} d_u^2 - \hat{\lambda}' D\right)\right] = \sigma^2$,

3) $\Sigma_{\hat{\beta}_L} = \sigma^2 S^{-1}[I - G'(GS^{-1}G')^g G S^{-1}]$.

2) Weights.

This method does not give an estimate that satisfies $G\hat{\beta} = D$. What it does is make it possible to "give up" a little on the constraints in order to obtain a smaller sum-of-squares of residuals. The method is widely used because it can be handled without modifying an existing least-squares program.

The procedure is to treat the constraint as an "observation equation"; that is, the model is considered as being of the form

$$\begin{pmatrix} \tilde{Y} \\ D \end{pmatrix} = \begin{pmatrix} X \\ G \end{pmatrix} \beta + \begin{pmatrix} \epsilon \\ f \end{pmatrix} \quad ,$$

where f is the "residual variable" for the constraint equations. Let

$$W = \begin{pmatrix} \tilde{Y} \\ D \end{pmatrix} \quad , \quad Z = \begin{pmatrix} X \\ G \end{pmatrix} \quad , \text{ and } \quad h = \begin{pmatrix} \epsilon \\ f \end{pmatrix} \quad ;$$

then

$$W = Z\beta + h \quad ,$$

and the usual least-squares procedure can be followed. Find $\hat{\beta}$ to minimize

$$(W - Z\hat{\beta})' \Sigma_h^{-1} (W - Z\hat{\beta}) \quad ,$$

where

$$\Sigma_h = \begin{pmatrix} \Sigma_e & 0 \\ 0 & H \end{pmatrix} \quad .$$

The off-diagonal matrices are taken to be 0 because it does not make sense to talk of a "covariance" between a random variable and a constraint. H is an r×r positive-definite diagonal matrix with very small diagonal elements, and H^{-1}, also diagonal, is the matrix of weights. Σ_e will again be taken equal to I without loss of generality. The solution is then

$$\hat{\beta}_W = \hat{\beta}_0 - (S + G'H^{-1}G)^{-1} G'H^{-1} \delta \quad ,$$

where $\hat{\beta}_0$ and δ are defined as for $\hat{\beta}_L$ and where it is assumed that $(S+G'H^{-1}G)^{-1}$ exists, which will be the case when S^{-1} exists.

Theorem 18

1) $\lim_{H \to 0} \hat{\beta}_W = \hat{\beta}_L$.
2) If $G\beta = D$ is true, then $E[\hat{\beta}_W] = \beta$.

The first statement of this theorem is a result that does not seem to be available in the literature, so some explanation is in order. In effect, it says that by use of large enough weights, a good approximation to the Lagrange estimate can be obtained. The proof is a straightforward application of a matrix identity known as the matrix inversion lemma or Woodbury's Theorem. For this case, it gives

$$(S + G'H^{-1}G)^{-1} = S^{-1} - S^{-1}G'(H + GS^{-1}G')^{-1} GS^{-1}$$

as long as S^{-1} exists.

8.2 Outliers

An outlier is an observation whose residual is far larger than the others, that is, 4 or 5 standard deviations from the mean. It may be due to gross errors, for example, a mistake in recording an observation, in which case, it is desirable to remove that observation from the data. On the other hand, the outlier may be due to an unusual combination of circumstances and is therefore providing information that the other observations do not. Automatic rejection of outliers – that is, removal of the corresponding observations from the data – is not very wise, because of the risk of losing this information. Rejection of points that are not gross errors leads to an underestimate of σ^2. In general, it is valuable to investigate outliers carefully to determine their cause. Any outliers that are rejected should be reported on separately.

The most general rule for rejecting outliers is the following: Pick the largest residual (in absolute value), remove the corresponding point – say, (x_0, y_0) – from the data, and then redo the analysis. (As used here, x_0 is a k component vector.) Using s^2, C, and $\hat{\beta}$ from the redone analysis, let

$$W = \frac{y_0 - x_0' \hat{\beta}}{s\sqrt{1 + x_0' C x_0}} \quad .$$

The test is to reject the hypothesis that (x_0, y_0) is <u>not</u> a gross error if W^2 is too large ($W^2 \sim F(1, n - k - 1)$ if the hypothesis is true). By Theorems 7 and 8, a good approximation for large n is

$$W \sim N(0, 1) \quad .$$

It is a good idea to use a very small significance level in order to minimize the possibility of rejecting a point that is not a gross error.

It is also necessary to take into account the fact that the largest residual has been picked. If the actual significance level of the test is to be α_0, then $\alpha = \alpha_0/n$ should be used in the following manner.* Reject the hypothesis if $W^2 > d$, where d is determined by $P[V > d] = \alpha$, with $V \sim F(1, n-k-1)$. Then, $P[W^2 > d] = \alpha_0$.

Example

Let n = 20 and suppose $\alpha = 0.05$ is used. Then, $\alpha_0 = 0.64$ $(= 1 - (0.95)^{20})$, so it should not be a surprise if the largest residual is rejected. If $\alpha_0 = 0.05$ is wanted, than $\alpha = 0.05/20 = 0.0025$ would have to be used. For n = 1000, with the normal approximation, rejecting the hypothesis for $|W| > 4.05\, s\sqrt{1 + x_0' C x_0}$ gives a 5% test; for $|W| > 4.4\, s\sqrt{1 + x_0' C x_0}$ gives a 1% test.

*This is an approximation valid for small α_0. The exact relationship is $\alpha_0 = 1 - (1 - \alpha)^n$.

9. REGRESSION WHEN ALL VARIABLES ARE SUBJECT TO ERROR

In this section, it will be necessary to distinguish between two types of relations, regression and functional:

1) A regression expresses a relation between the expected value of one variable, Y, and another set of variables, $X_1, \ldots, X_k$. For example,

$$E[Y] = \alpha + \beta X \quad \text{or} \quad Y = \alpha + \beta X + \epsilon \ ,$$

where ϵ is the usual residual term. Note that the relation $X = (Y - \alpha)/\beta$ does not make any sense.

2) A functional relationship expresses an exact relationship among a set of variables. In this case, if the variables could be observed without error, there would be no statistical problem and the unknown coefficients could be calculated directly. For example,

$$Y = \alpha + \beta X \ ,$$

or, equivalently,

$$X = \frac{Y - \alpha}{\beta} \ ,$$

which now makes sense.

These two types are not mutually exclusive. A functional relation with one and only one variable subject to error is the same as a regression relation with the residual term being the error.

The notation for this section will be the following: The model is:

$$Y = X\beta + \epsilon \ , \quad X = (X_1, \ldots, X_k) \ ,$$

where e is the other-effects term, which will be identically zero for a functional relation. The observed variables are

$$V = X + h$$

and

$$W = Y + f \quad ,$$

where $h = (h_1, \ldots, h_k)$ and f are random variables representing the errors of observation. For n observations, the model in matrix notation becomes

$$\tilde{W} = (\tilde{V} - H)\beta + e + \tilde{f} \quad ,$$

where

$$\tilde{W} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}, \quad \tilde{V} = \begin{pmatrix} v_{11} \cdots v_{1k} \\ \vdots \quad\quad \vdots \\ v_{n1} \cdots v_{nk} \end{pmatrix}, \quad H = \begin{pmatrix} h_{11} \cdots h_{1k} \\ \vdots \quad\quad \vdots \\ h_{n1} \cdots h_{nk} \end{pmatrix}, \quad \text{and} \quad \tilde{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} .$$

Suppose W is to be predicted from some future observation on V. For this case, the least-squares solution ($\hat{\beta} = (\tilde{V}'\tilde{V})^{-1}\tilde{V}'\tilde{W}$) for the regression of W on V works, since W and V are observed without error. The model is

$$W = V\beta + g \quad ,$$

where

$$g = e + f - h\beta \quad .$$

It is necessary, though, to assume that $E[g] = 0$. If all the variances and covariances are constant, then

$$\sigma_g^2 = \sigma_e^2 + \sigma_f^2 + \beta' \Sigma_h \beta + 2\sigma_{ef} - 2\beta'(\sigma_{eh} + \sigma_{fh}) \quad ;$$

σ_{eh} and σ_{fh} are k-dimensional vectors where

$$(\sigma_{eh})_i = \sigma_{eh_i} \quad ,$$

and

$$(\sigma_{fh})_i = \sigma_{fh_i}$$

($\sigma_e^2 = \sigma_{ef} = \sigma_{eh_i} = 0$ for the case of a functional relation between X and Y.) The effect of the errors of observation in the "independent" variables is an increase in the residual variance.

Before we go on, there are three points to consider. First, $\hat{\beta}$ is not necessarily unbiased, nor for that matter, even consistent. For predictions, this should not matter. However, there are conditions for unbiasedness that will be mentioned later. Also, s^2 is not always an unbiased estimate of σ_g^2. It will be, though, if the same conditions hold that are necessary to guarantee the unbiasedness of $\hat{\beta}$ and if $E[g_u g_v] = 0$ for $u \neq v$.

Second, since σ_g^2 depends on β, weighting inversely to the variance by minimizing

$$\frac{1}{\sigma_g^2} \sum_u (v_u - W_u \hat{\beta})^2$$

leads to a different solution [$W_u = (w_{u1}, \ldots, w_{uk})$]. That solution can be obtained by solving k simultaneous cubic equations or by using a direct minimization technique. Even if it is feasible to find this solution, there is no guarantee that the solution will be unbiased under any conditions, or even unique.

The real problem is the assumption that $E[g] = 0$. As explained in Section 5.3, $E[e] \neq 0$ and $E[f] \neq 0$ can be taken care of if we include a constant term in the model. This does not generally work for $E[h] \neq 0$. To see where the problem is, remember that the X_i's are really functions of some other set of variables $Z_1, \ldots, Z_p$, which are observed with error. If the X's are not linear functions of the Z's, then except in very special cases, $E[h] \neq 0$

even if the errors on the Z_i's have mean zero, and $E[h_u]$ may be different for each u.

Example

Suppose you make n observations on Z and observe $z_u + k_u$ for $u = 1, \ldots, n$, where k_u is the error term. If $x_u = e^{z_u}$, then

$$e^{z_u+k_u} = x_u e^{k_u} ;$$

therefore,

$$h_u = x_u(e^{k_u} - 1)$$

and

$$E[h_u] = x_u(E[e^{k_u}] - 1) .$$

Unless $E[e^{k_u}] = 1$, $E[h_u]$ will not be zero and will be different for each u, since it depends on x_u.

The only way I know to handle this problem is to avoid it; that is, whenever possible, observe directly those variables, or linear functions of them, to be used in the regression equation.

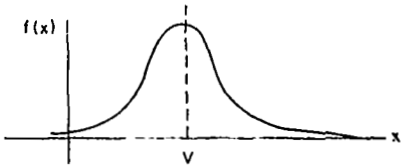
Another difficulty is that $E[g] = 0$ requires that g be uncorrelated with the V_i's (or X_i's, depending on which is taken as fixed – cf. Section 4). This, in turn, will generally require that the V_i's (X_i's) be uncorrelated with e, f, and the h_i's. Of these, the most unreasonable is that V_j (X_j) be uncorrelated with h_j, i.e., that the observation (true value) be uncorrelated with the error in that observation.

Because of these difficulties, there is no doubt but that the assumption $E[g] = 0$ is, at least, questionable. Unfortunately, it is a necessary assumption if anything is able to be said about what happens to least squares in the presence of observation errors in the X_i's. So, from now on, we will ignore the difficulties and make the assumption.

To determine when $\hat{\beta}$ will be unbiased, consideration must be given to how the data were obtained. As a clarifying example, suppose that observations of some sort are made on Y at different values of X (one independent variable), that the equipment can be adjusted to obtain different, but unknown, values of X, and that there is at hand a meter from which the values of V(= X + h) are read. Then, the data can be obtained in two ways:

1) Controlled experiment. The values of V at which observations are to be made are set beforehand; that is, when the experiment is being run, the equipment is adjusted until the meter reading agrees with the values chosen. The result is that V can be considered as fixed – that is, not random – and X, the true value, is a random variable. With $E[h] = 0$, this means that X (= V - h) has a distribution with $E[X] = V$ as shown:

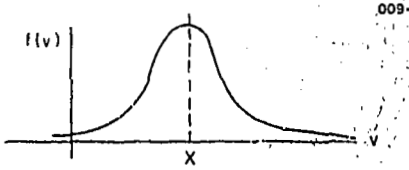
009-113



If the experiment is repeated, the true value may be different, but the expected value of both true values will be the same, namely, equal to V.

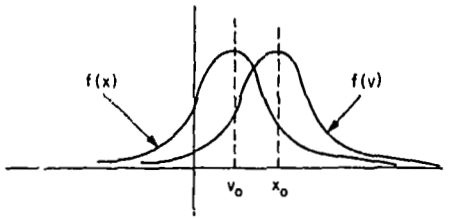
2) Random experiment. The values of V are not set beforehand. They are determined by "spinning the dial," that is, in some random manner. Here, X is fixed and V is random. If $E[h] = 0$, then V has the following distribution with $E[V] = X$:

009-113



If the experiment is repeated – that is, if a duplicate observation for the same value of V is made – the distribution will be different, as shown:

009-113



When the experiment is first run, the true value is x_0, say, but v_0 is observed. When the experiment is repeated, v_0 is observed, but the expected value of the second true value is not x_0, except in very rare circumstances.

Returning to the general case of k X_i's, consider $\hat{\beta}$:

$$\hat{\beta} = (\tilde{V}'\tilde{V})^{-1}\tilde{V}'\tilde{W} \quad ,$$

$$= \beta + (\tilde{V}'\tilde{V})^{-1}\tilde{V}'g \quad ,$$

where

$$\tilde{W} = \tilde{Y} + g \quad .$$

If V is fixed, $E[\hat{\beta}] = \beta$ since $E[g] = 0$. If V is not fixed,

$$E[\hat{\beta}] = \beta + E[\{(X + H)'(X + H)\}^{-1}(X + H)'g] \quad .$$

The last term is definitely not equal to zero unless H = 0, since it contains terms in h_{ui}^2. The situation is also complicated by the fact that X is unknown, so $\hat{\beta}$ has an unknown bias.

Related to this is the following interesting result. If the model is

$$Y = \beta_0 + \beta_1 X + \ldots + \beta_k X^k + e \quad ,$$

with V = X + h, E[h] = 0, and V considered to be fixed as in the controlled-experiment case, then $\hat{\beta}_k$ and $\hat{\beta}_{k-1}$ will be the only unbiased estimates.

If the parameters of a functional relationship are to be estimated with data from a random experiment, there are some additional problems. Consider the functional relationship

$$Y = \alpha + \beta X$$

with k = 1 and let the observed variables be W(= Y + f) and V(= X + h), where E[f] = E[h] = 0. Now suppose α and β are to be estimated from n observations.

One set of estimates would be that obtained by least squares for the regression of W on V, that is, for

$$W = \alpha_1 + \beta_1 V + g_1 \quad ,$$

where $g_1 = f - \beta_1 h$. Then,

$$\hat{\beta}_1 = \frac{\sum_{u=1}^{n} (v_u - \overline{V})(w_u - \overline{W})}{\sum_{u=1}^{n} (v_u - \overline{V})^2}$$

and

$$\hat{\alpha}_1 = \overline{W} - \hat{\beta}_1 \overline{V} \quad .$$

Another possible set of estimates could be found by considering the regression of V on W, that is,

$$V = \alpha_2 + \beta_2 W + g_2 \quad ,$$

where $g_2 = h - \beta_2 f$. Then,

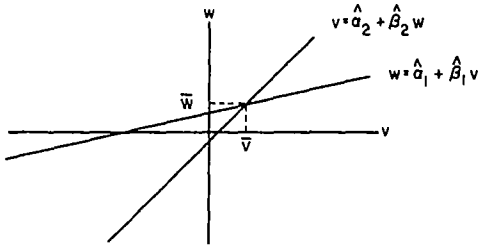
$$\hat{\beta}_2 = \frac{\sum_{u=1}^{n} (v_u - \overline{V})(w_u - \overline{W})}{\sum_{u=1}^{n} (v_u - \overline{V})^2}$$

and

$$\hat{\alpha}_2 = \overline{V} - \hat{\beta}_2 \overline{W} \quad .$$

The four unknown parameters are related by $\beta_1\beta_2 = 1$ and $\alpha_1 = -\alpha_2\beta_1$, but the estimates do not satisfy these relations. The estimates lead to two different lines:

009-113



Is either one of these lines the estimate of $Y = \alpha + \beta X$? In general, no; the true line lies somewhere between the two. The special case, as mentioned earlier, is where one of the variables can be considered fixed. Then, there is only one regression line to be estimated and $E[\hat{\beta}] = \beta$. For the random-experiment case, least squares breaks down because it considers errors only in one direction, while there are errors in both directions that must be taken into account. Lindley (1947) and Madansky (1959) claim that minimizing $(1/\sigma_{g_1}^2) \sum_{u=1}^{n} (w_u - \hat{\alpha} - \hat{\beta} v_u)^2$ takes both errors into account. (Remember that σ_g^2 depends on β.) This procedure requires knowledge of, or estimates of, all the variances and covariances. Also, the solution is not necessarily

unique. A good justification for using this estimate in the case $\sigma_f^2 = \sigma_h^2 = 1$ and $\sigma_{fh} = 0$ is that minimizing this quantity, now $[1/(1 + \hat{\beta}^2)] \sum_{u=1}^{n} (w_u - \hat{\alpha} - \hat{\beta} v_u)^2$, minimizes the sum of squares of the distances between the observed points (v_u, w_u) and the fitted line.

An alternative to least squares would be to use the maximum-likelihood estimate. This would require a knowledge of the joint distribution of all the errors. As an example, let k = 1 and suppose f and h have a bivariate normal distribution. It turns out that for there to be a solution to the maximum-likelihood equation, some assumptions must be made about σ_f^2, σ_h^2, and σ_{fh}, but if they are all assumed to be known, there will be more equations than unknowns and the solution will not be unique. If no assumptions are made, the solution will depend on the unknown variances and covariances. Examples of assumptions that result in a unique solution are

1) $\sigma_{fh} = 0$ and $\sigma_f^2/\sigma_h^2 = \lambda$ known.

2) σ_h^2 and σ_f^2 known; σ_{fh} unknown.

Madansky (1959) gives the estimates in these two and several other cases.

Maximum likelihood can be a problem, since it may not always be desirable to assume that the errors have a multivariate normal distribution, or any other distribution for that matter. Also, enough may not be known about the variances and covariances for it to be possible to get a solution.

Malinvaud (1966) and Madansky discuss some other methods that do not require knowledge of the various covariances and variances. Each involves some particular assumptions about the true values and errors.

One of these is the method of instrumental variables. This requires knowledge of other variables, which are observed without error, that are uncorrelated with the h_i's and, ideally, highly correlated with the V_i's. For the case k = 1, β is estimated by

$$\hat{\beta} = \frac{\sum_{u=1}^{n} (w_u - \overline{W})(r_u - \overline{R})}{\sum_{u=1}^{n} (v_u - \overline{V})(r_u - \overline{R})} ,$$

where R is the instrumental variable. Given the assumption that R is observed without error, $\hat{\beta}$ will be unbiased. The problem is that it is not yet known what happens when the assumption is not true but is still a reasonable approximation, the most likely situation to be encountered in practice.

Another method is that of grouping. This involves classifying the observations into groups and fitting the group means. This method yields consistent estimates under certain rather stringent assumptions about the observations and errors.

In summary, if what is wanted are estimates for prediction purposes, least squares can be used without worrying about the problems. If unbiased estimates of the parameters are wanted, then the independent variables should be controlled. If that is not possible, try to use maximum-likelihood estimates that are at least asymptotically unbiased, or use Lindley's method.

10. NONLINEAR REGRESSION

In practice, it is not always possible to use the linear additive model. The application of least squares then almost always implies the use of an iterative minimization technique.

For nonlinear regression, our model is

$$Y = f(Z_1, \ldots, Z_p; \beta_1, \ldots, \beta_k) + e ,$$

where e is the residual term and $\beta_1, \ldots, \beta_p$ are the coefficients to be estimated.* Note that even with the nonlinear model, we must assume that the residual term is additive. Given the observations $(z_{u1}, \ldots, z_{up}, y_u)$ for $u = 1, \ldots, n$, the problem is to find $\hat{\beta}_1, \ldots, \hat{\beta}_k$ to minimize

$$S(\hat{\beta}) = \sum_{u=1}^{n} (y_u - \hat{y}_u)^2 ,$$

where $\hat{y}_u = f(z_{u1}, \ldots, z_{up}; \hat{\beta}_1, \ldots, \hat{\beta}_k) = f(\tilde{Z}_u; \hat{\beta})$, and $\tilde{Z}_u = (z_{u1}, \ldots, z_{up})$. The assumptions needed here are essentially the same as in the linear case:

1) The model is correct.

2) The Z_i's are observed without error.

3) $E[e_u] = 0$.

4) $E[e_u e_v] = \sigma^2 \delta_{uv}$, for u, v = 1, ..., n, which can be satisfied by using Σ_e if necessary, as in Section 5.3.

*The use of Z_i's instead of X_i's is deliberate. There is no longer any reason to use functions of the Z_i's in the model.

Before we consider the methods for solving this problem, three points that are true in general must be stated:

1) There is no guarantee that the $\hat{\beta}_i$'s will be unbiased.

2) In general, $E[s^2] \neq \sigma^2$.

3) If $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$, then the least-squares estimates are also the maximum-likelihood estimates.

The most commonly used methods are discussed in the following subsections:

10.1 Linearization

By Taylor's theorem,

$$f(\tilde{Z}_u;\beta) = f(\tilde{Z}_u;\beta_0) + \sum_{i=1}^{k} \frac{\partial f(\tilde{Z}_u;\beta)}{\partial \beta_i}\bigg|_{\beta=\beta_0} (\beta_i - \beta_{0i})$$

is approximately true for β_0 near β. Then, the model is approximately

$$y_u - f(\tilde{Z}_u;\beta_0) = \sum_{i=1}^{r} \delta\beta_i w_{ui} + e_u \quad ,$$

where

$$\delta\beta_i = \beta_i - \beta_{0i} \quad ,$$

and

$$w_{ui} = \frac{\partial f(\tilde{Z}_u;\beta)}{\partial \beta_i}\bigg|_{\beta=\beta_0} \quad .$$

This is a linear model, and least squares can be used to estimate $\delta\beta_1, \ldots, \delta\beta_k$, where β_0 is an initial guess at the value of β. β_0 is then replaced by $\beta_0 + \hat{\delta\beta}$, where $\hat{\delta\beta} = (\hat{\delta\beta}_1, \ldots, \hat{\delta\beta}_k)'$, and the process is repeated. This continues until the improvement, as measured by the decrease in S, becomes small.

Three problems may be encountered when this method is used:

1) It may converge very slowly.

2) It may oscillate wildly.

3) It may diverge.

To minimize these problems, use $\hat{\delta\beta}/2$ instead of $\hat{\delta\beta}$ if

$$S(\beta_0 + \hat{\delta\beta}) > S(\beta_0) \quad .$$

It is always a good idea to calculate $S(\beta_0 + \hat{\delta\beta})$ after each step to be able to keep track of what is happening.

When this method is used, approximate tests of significance can be obtained by assuming that the linearized form of the model is valid around $\hat{\beta}$, the final estimate of β. Then s^2 can be used as an approximation for σ^2, although it is not unbiased, and the final $(W'W)^{-1}$ matrix can be used for the standard errors of $\hat{\beta}$.

10.2 Steepest Descent

The basic idea of all the steepest descent (gradient) methods is that from any point β_0, the vector $-\nabla S(\beta)\big|_{\beta=\beta_0}$ points in the direction of the greatest decrease in S. Many modifications of this idea have been developed, the best of which seems to be that by Fletcher and Powell (1963). The basic steps of their procedure are the following, where subscripts denote the iteration number. At the n^{th} step, you begin with $\hat{\beta}^{(n)}$, $g_n = \nabla S(\hat{\beta}^{(n)})$, and H_n. (H_1 is chosen to be positive definite, and $\hat{\beta}^{(1)}$ is any initial estimate.) Then,

1) $p_n = -H_n g_n$.

2) Find a_n to minimize $S(\hat{\beta}^{(n)} + a p_n)$ with respect to a.

3) $\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + a_n p_n$.

4) $f_n = g_{n+1} - g_n$.

5) $H_{n+1} = H_n + A_n + B_n$, where

$$A_n = \frac{a_n p_n p_n'}{p_n' f_n} \quad ,$$

and

$$B_n = -\frac{H_n f_n f_n' H_n'}{f_n' H_n f_n} .$$

Straeter and Hogge (1970) say that this procedure is easy to implement for general problems, but it does require an accurate linear minimization technique (step 2). For least-squares problems, they prefer the following method, which has a much faster rate of convergence in terms of function evaluations.

10.3 Marquardt's Compromise*

Marquardt (1963) found that for a number of the least-squares problems he worked with, the directions of improvement (in the k-dimensional parameter space) obtained by linearization and steepest descent were nearly 90° apart. His algorithm provides a method for interpolating between the two directions.

Beginning with an estimate $\hat{\beta}$, let

$$d(\hat{\beta}) = (y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n)' \quad \text{and} \quad \delta\beta = (\delta\beta_1, \ldots, \delta\beta_k)' .$$

Then define $\underset{\sim}{A}d(\hat{\beta})$ to be a $k \times k$ matrix given by

$$[\underset{\sim}{A}d(\hat{\beta})]_{ij} = \frac{\partial(y_i - \hat{y}_i)}{\partial\hat{\beta}_j} .$$

The basic idea is to find $\delta\beta$ to minimize

$$C = [d(\hat{\beta}) + \underset{\sim}{A}d(\hat{\beta})\,\delta\beta]'\,[d(\hat{\beta}) + \underset{\sim}{A}d(\hat{\beta})\,\delta\beta] \cong S(\hat{\beta} + \delta\beta) - S(\hat{\beta})$$

*Sometimes referred to as the Levenberg-Marquardt or Armstrong-Marquardt algorithm.

subject to the constraints that $(\delta\beta)'(\delta\beta) < \gamma$, for some γ, and that $C < 0$. The correction vector is given by

$$\delta\beta = \{[\underset{\sim}{A}d(\hat{\beta})]'\,[\underset{\sim}{A}d(\hat{\beta})] + \lambda I\}^{-1}[\underset{\sim}{A}d(\hat{\beta})]'\,d(\hat{\beta}) ,$$

where, in practice, λ is chosen so that $S(\hat{\beta} + \delta\beta) < S(\hat{\beta})$ and the matrix $\{[\underset{\sim}{A}d(\hat{\beta})]'\,[\underset{\sim}{A}d(\hat{\beta})] + \lambda I\}$ is invertible. Armstrong discusses approaches for specifying λ (Armstrong, 1970) and has shown (Armstrong, 1968) that as $\lambda \to \infty$, $\delta\beta$ goes toward the direction of the negative gradient and that as $\lambda \to 0$, $\delta\beta$ approaches the correction vector that would be obtained using linearization.

Brown and Dennis (1970) have derived a derivative-free analogue of this method. According to their paper, this analogue compares favorably with the original algorithm.

11. REFERENCES AND BIBLIOGRAPHY

ACTON, F. S.

1959. Analysis of Straight-Line Data. John Wiley & Sons, New York, 267 pp.

ANSCOMBE, F. J

1961. Examination of residuals. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. I, ed. by J. Neyman, University of California Press, Berkeley, Calif., pp. 1-36.

ANSCOMBE, F. J., and TUKEY, J. W.

1963. The examination and analysis of residuals. Technometrics, vol. 5, pp. 141-159.

ARMSTRONG, E. S.

1968. A combined Newton-Raphson and gradient parameter correction technique for solution of optimal-control problems. NASA Publ. TRR-293, 61 pp.

1970. On the specification of Levenberg parameters for improving convergence in least-squares programming. NASA, unpublished, 23 pp.

BERKSON, J.

1950. Are there two regressions? Journ. Amer. Statistical Assoc., vol. 45, pp. 164-180.

BROWN, K. M., and DENNIS, J. E.

1970. Derivative free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares approximation. IBM-Philadelphia Scientific Center, Tech. Rep. No. 320-2994, 21 pp.

COCHRAN, W. G.

1969. Unpublished notes for Statistics 139, Analysis of variance. Department of Statistics, Harvard University, 67 pp.

DRAPER, N. R., and SMITH, H.

1966. Applied Regression Analysis. John Wiley & Sons, New York, 407 pp.

FELLER, W.

1966. An Introduction to Probability Theory and Its Applications. Vol. II, John Wiley & Sons, New York, 626 pp.

1968. An Introduction to Probability Theory and Its Applications. Vol. I, 3rd ed., John Wiley & Sons, New York, 510 pp.

FLETCHER, R., and POWELL, M. J. D.

1963. A rapidly converging descent method for minimization. Comp. Journ., vol. 6, pp. 163-168.

GOLUB, G.

1965. Numerical methods for solving linear least squares problems. Numerische Mathematik, vol. 7, pp. 206-216.

1969. Matrix decompositions and statistical calculations. In Statistical Computation, ed. by R. C. Milton and J. A. Nelder, Academic Press, New York, pp. 365-397.

GOLUB, G., and BUSINGER, P.

1965. Linear least squares solutions by Householder's transformations. Numerische Mathematik, vol. 7, pp. 269-276.

GRAYBILL, F. A.

1969. Introduction to Matrices with Applications in Statistics. Wadsworth Publ. Co., Inc., Belmont, Calif., 372 pp.

HOGG, R. V., and CRAIG, A. T.

1965. Introduction to Mathematical Statistics. 2nd edition, Macmillan Co., New York, 383 pp.

JORDAN, T. L.

1968. Experiments on error growth associated with linear least-squares procedures. Math. Comp., vol. 22, pp. 579-588.

KENDALL, M. G., and STUART, A.

1961. The Advanced Theory of Statistics. Vol. II, Hafner Publ. Co., New York, 676 pp.

1963. The Advanced Theory of Statistics. Vol. I, 2nd ed., Hafner Publ. Co., New York, 433 pp.

BIOGRAPHICAL NOTE

WALTER W. HAUCK, JR., received his B.S. degree in mathematics and economics from Carnegie-Mellon University in 1969 and the M.A. degree in statistics at Harvard University in 1970. He is currently a graduate student in the Department of Statistics, Harvard University.

Mr. Hauck's principal research interests include statistical decision theory and least-squares estimation.

NOTICE

This series of Special Reports was instituted under the supervision of Dr. F. L. Whipple, Director of the Astrophysical Observatory of the Smithsonian Institution, shortly after the launching of the first artificial earth satellite on October 4, 1957. Contributions come from the Staff of the Observatory.

First issued to ensure the immediate dissemination of data for satellite tracking, the reports have continued to provide a rapid distribution of catalogs of satellite observations, orbital information, and preliminary results of data analyses prior to formal publication in the appropriate journals. The Reports are also used extensively for the rapid publication of preliminary or special results in other fields of astrophysics.

The Reports are regularly distributed to all institutions participating in the U. S. space research program and to individual scientists who request them from the Publications Division, Distribution Section, Smithsonian Astrophysical Observatory, Cambridge, Massachusetts 02138.