MULTIVARIATE REGRESSION ANALYSIS

by

John H. Halton

Computer Sciences Technical Report #425

March 1981

MULTIVARIATE REGRESSION ANALYSIS

John H. Halton

ABSTRACT

This paper gives an exposition of the statistical method for finding an optimal estimate of the regression $E[\eta | \boldsymbol{x}] = f(\boldsymbol{x})$ in the form $f(\boldsymbol{x}) = \Sigma_{h=1}^{g} \alpha_h \phi_h(\boldsymbol{x})$, where η is a random variable whose distribution is normal about the regression surface (or line) $y = f(\boldsymbol{x})$. Emphasis is placed on the assumptions and choices underlying the results used, and these results are derived rigorously. The validity of these assumptions and choices is discussed.

MULTIVARIATE REGRESSION ANALYSIS

John H. HALTON

CORRECTIONS

P. 13, l. 6/7: For "n-1 times" read "related to"

l. 8: For "$N(\frac{n}{n-1}\mathrm{E}[\xi\eta] + x\sqrt{n(n-1)\mathrm{var}[\xi\eta]})$"

read "$N(\frac{n}{n-1}\{\mathrm{E}[\xi\eta] + x\sqrt{\mathrm{var}[\xi\eta]/n}\})$"

P. 14, l. 4 up: For "$-\frac{1}{2}\frac{(y - \mathrm{E}[\xi])}{\mathrm{var}[\xi]}(1 - 2it)$"

read "$-\frac{1}{2}\frac{(y-\mathrm{E}[\xi])^2}{\mathrm{var}[\xi]}(1 - 2it)$"

P. 21, l. 6 up: For "$\|W^{\frac{1}{2}}\Phi^{\mathrm{T}}\gamma\| \geqslant 0$" read "$\|W^{\frac{1}{2}}\Phi^{\mathrm{T}}\gamma\|^2 \geqslant 0$"

P. 28, eqs. (105) - (109), throughout: Replace h, i, j by t, u, v, respectively.

P. 31, l. 5, 4 up: For "by (117), that","$W^{\frac{1}{2}}\eta = \zeta + W^{\frac{1}{2}}\Phi\,\beta_0$ ="

read "by (84) and (117), that","$W^{\frac{1}{2}}\eta = \zeta + W^{\frac{1}{2}}\Phi^{\mathrm{T}}\beta_0$ ="

P. 32, bottom l.: Eq. (125) should be preceded by "$\begin{pmatrix}\omega_1\\ \omega_2\end{pmatrix} = \begin{pmatrix}\mathbf{I} & A\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}$ or"

P. 34, eq. (134): For "$S_n(\beta_1) = \zeta\ \zeta\ =$" read "$S_n(\beta_1) = \zeta^{\mathrm{T}}\zeta\ =$"

P. 35, l. 11, 12: Should read:

"*(e) The deviations from the underlying regression function are indeed distributed* ***independently*** *of one-another, about* ***zero***"

P. 36, l. 3 up: For "of the ξ values" read "of the ξ_j values"

MULTIVARIATE REGRESSION ANALYSIS

John H. Halton

RANDOM VARIABLES[1]

Let S be a set, the *global set*, and let $\boldsymbol{S}$ be a σ-algebra of subsets of S [containing the *empty set* $\emptyset$ and the global set S, and *closed* under complementation (if $E \in \boldsymbol{S}$, then $E^c \in \boldsymbol{S}$) and countable union (if $E_1, E_2, E_3, \ldots \in \boldsymbol{S}$, then $\cup_{i=1}^{\infty} E_i \in \boldsymbol{S}$)]; and let p be a **probability** on $\boldsymbol{S}$ [that is, a non-negative, totally finite *measure*: such that $p(\emptyset) = 0$, $p(S) = 1$, and if $E_1, E_2, E_3, \ldots \in \boldsymbol{S}$ and these sets are all *disjoint*, then

$$p(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} p(E_i); \tag{1}$$

whence, if the sets E_i are *not* required to be disjoint, then

$$p(\bigcup_{i=1}^{\infty} E_i) \leqslant \sum_{i=1}^{\infty} p(E_i), \tag{2}$$

a property referred to as *sub-additivity*] Then the global set S is also called the **sample space**, and the complete structure $(S, \boldsymbol{S}, p)$ is called a **probability space**.

EXAMPLE: If $S = \{1, 2, \ldots, n\}$, a finite set, we may take $\boldsymbol{S}$ to be the collection of all subsets of S, and then $p(E) = \Sigma_{i \in E} p_i$, where $p_i = p(\{i\})$ is the probability of the singleton set $\{i\}$, by (1).

If R denotes the set of real numbers and $\boldsymbol{R}$ the σ-algebra of *Borel sets* [which contains, in particular, all the open intervals $\{x \in R: a < x < b\}$ of R, and hence all the *open sets* (i.e., the *topology*) of R], then a function $\xi: S \to R$ is *measurable* if and only if [iff] $\xi^{-1}(B) \in \boldsymbol{S}$ whenever $B \in \boldsymbol{R}$, where

$$\xi^{-1}(B) = \{s \in S: \xi(s) \in B\} \tag{3}$$

is the *inverse image* of B under ξ: the set of all s in S, such that $\xi(s)$ is in B. A **random variable** [r.v.] is simply a measurable function, when the measure p is a probability. Since, for any r.v. ξ, the inverse image of any Borel set B is in the domain $\boldsymbol{S}$ of p, as the definition is intended to guarantee, it follows that the probability $p\xi^{-1}(B)$ is well defined [in the sense of $p(\xi^{-1}(B))$], and so $(R, \boldsymbol{R}, p\xi^{-1})$ is itself a probability space, called the **distribution** of ξ. Further, it is known that the distribution of a r.v. is determined uniquely if we are given the probability of all values less than any given value x:

$$F_\xi(x) = p\xi^{-1}(\{r \in R: r < x\}) = p(\{s \in S: \xi(s) < x\}) \tag{4}$$

and this function is referred to as the *cumulative distribution function* [c.d.f.] of ξ.

EXAMPLE: A r.v. which takes on only a finite number of values, say $\alpha_1 < \alpha_2 < \ldots < \alpha_m$, is called *simple*. If $p\xi^{-1}(\alpha_i) = p_i$ for each i, then $F_\xi(x) = \sum_{i=1}^{k(x)} p_i$, where $\alpha_{k(x)} < x \leqslant \alpha_{k(x)+1}$; and we see that $F_\xi(x)$ takes the form of a *step function*.

For any distribution, it is clear that the c.d.f. is *continuous to the left*. If the c.d.f. is *differentiable*, its derivative $\rho_\xi(x)$ is called

the **probability density** of the r.v. ξ, at x.

The *Lebesgue theory of integration*[2] defines the integral of a measurable function by extension from that of a simple function, given by

$$\int_S \xi dp = \sum_{i=1}^{m} \alpha_i p_i, \tag{5}$$

and we define the *mathematical* **expectation** (or *mean value*) of a r.v. ξ as

$$E[\xi] = \int_S \xi dp = \int_{-\infty}^{+\infty} x dF_\xi(x) = \int_{-\infty}^{+\infty} x \rho_\xi(x) dx, \tag{6}$$

the last form only if a probability density exists. Similarly, the **variance** of ξ is defined as

$$\text{var}[\xi] = E[(\xi - E[\xi])^2] = E[\xi^2] - (E[\xi])^2, \tag{7}$$

and if ξ and η are both r.v., then the **covariance** of ξ and η is

$$\begin{aligned} \text{cov}[\xi, \eta] &= E[(\xi - E[\xi])(\eta - E[\eta])] \\ &= E[\xi\eta] - E[\xi]E[\eta], \end{aligned} \tag{8}$$

so that

$$\text{var}[\xi] = \text{cov}[\xi, \xi]. \tag{9}$$

If ξ is an m-dimensional vector of r.v. $\xi_1, \xi_2, \ldots, \xi_m$, then we may define the expectation vector and the *variance-covariance* matrix of ξ as

$$E[\xi] = \begin{pmatrix} E[\xi_1] \\ E[\xi_2] \\ \cdots \\ E[\xi_m] \end{pmatrix}, \quad \mathbf{V}[\xi] = \begin{pmatrix} \text{var}[\xi_1] & \text{cov}[\xi_1, \xi_2] & \cdots & \text{cov}[\xi_1, \xi_m] \\ \text{cov}[\xi_2, \xi_1] & \text{var}[\xi_2] & \cdots & \text{cov}[\xi_2, \xi_m] \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}[\xi_m, \xi_1] & \text{cov}[\xi_m, \xi_2] & \cdots & \text{var}[\xi_m] \end{pmatrix}, \tag{10}$$

or, equivalently, by

$$(E[\xi])_i = E[\xi_i], \quad (\mathbf{V}[\xi])_{ij} = \text{cov}[\xi_i, \xi_j]. \tag{11}$$

If $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$, we may define a *joint c.d.f.* for the r.v. ξ by

$$F_{\xi}(\boldsymbol{x}) = p(\{s \in S: (\forall i \in \{1, 2, \ldots, m\})\ \xi_i(s) < x_i\}). \tag{12}$$

Similarly, if $F_{\xi}(\boldsymbol{x})$ is appropriately differentiable, we define a *joint probability density* by

$$\rho_{\xi}(\boldsymbol{x}) = \frac{\partial^m}{\partial x_1 \partial x_2 \ldots \partial x_m} F_{\xi}(\boldsymbol{x}); \tag{13}$$

whence

$$E[\xi_i] = \int_{-\infty}^{+\infty} dx_1 \ldots \int_{-\infty}^{+\infty} dx_i \ldots \int_{-\infty}^{+\infty} dx_m\ x_i\ \rho_{\xi}(\boldsymbol{x}). \tag{14}$$

If the set $E \in \boldsymbol{S}$, we often refer to it as an *event*; and we may define the **conditional expectation** of the r.v. ξ, given that the event E occurs, by

$$E[\xi \mid E] = \int_E \xi dp/p(E). \tag{15}$$

The events of greatest interest are those which are defined in terms of the values of random variables. For example, suppose that we define a r.v. η and a vector ξ of r.v. Then the conditional expectation of η, given that $\xi = \boldsymbol{x}$, is

$$E[\eta \mid \xi = \boldsymbol{x}] = \int_{-\infty}^{+\infty} y \rho_{\eta,\xi}(y, \boldsymbol{x}) dy \Big/ \int_{-\infty}^{+\infty} \rho_{\eta,\xi}(y, \boldsymbol{x}) dy, \tag{16}$$

where we denote the joint probability density for η and ξ by $\rho_{\eta,\xi}(y, \boldsymbol{x})$ and assume that it exists in the given case.

If it is the case that, for all choices of $\boldsymbol{x}$,

$$E[\eta \mid \xi = \boldsymbol{x}] = E[\eta], \tag{17}$$

so that the conditional expectation of η is the same for all values of $\boldsymbol{x}$ and equals the unconditional expectation of η, then we say that η is

independent of the r.v. ξ; and then we have that

$$F_{\eta,\xi}(y, x) = F_{\eta}(y)F_{\xi}(x) \text{ and } \rho_{\eta,\xi}(y, x) = \rho_{\eta}(y)\rho_{\xi}(x), \tag{18}$$

the latter if the densities exist. Thus *independence allows us to separate the variables in integration.* We note, in particular, that, if (17) holds, then $\text{cov}[\eta, \xi_i] = 0$ for all i. This is expressed by saying that independent r.v. are *uncorrelated.* [The converse, that uncorrelated r.v. are independent is not always true.]

INDEPENDENT TRIALS, UNBIASED ESTIMATORS

Given a probability space $(S, \boldsymbol{S}, p)$, it is thought of as representing a *statistical experiment* in which a member s of the sample space S is randomly sampled according to the probability distribution specified by p, and appropriate r.v. are then evaluated. It is usually the case that we expand the experiment by **repeated independent trials**, in which the r.v. are sampled without regard to previous results: it is then assumed that the r.v. evaluated in each trial are independent of all r.v. evaluated in other trials. This corresponds to a *product space*[3], denoted by $(S^n, \boldsymbol{S}^n, p^n)$ if there are n trials.

Let ξ be a r.v. on $(S, \boldsymbol{S}, p)$ and let n independent trials be made, with $s_1, s_2, \ldots, s_n$ as outcomes. The corresponding results of the evaluations of ξ are $\xi(s_1), \xi(s_2), \ldots, \xi(s_n)$, and we may define a new r.v. on the product space

$$\psi_n(s_1, s_2, \ldots, s_n) = \frac{1}{n}\sum_{i=1}^{n} \xi(s_i). \tag{19}$$

We now verify that

$$E[\psi_n] = E[\xi] \quad \text{and} \quad \text{var}[\psi_n] = \text{var}[\xi]/n. \tag{20}$$

Kolmogorov[4] has shown (in a result known as the *Strong Law of Large Numbers*) that, if $E[\xi]$ is finite (even if $\text{var}[\xi]$ is not) then

$$\psi_n \to E[\xi] \text{ (a.s.) as } n \to \infty; \tag{21}$$

that is, if the trials are repeated infinitely often, the sequence of corresponding values of $\psi_1, \psi_2, \psi_3, \ldots$ will converge to $E[\xi]$ *with probability one*[5] (or *almost surely*: a.s.) It is this result which justifies the use of repeated independent trials and averaged r.v. such as ψ_n to estimate the value of $E[\xi]$; or conversely, it is because such averages as ψ_n converge (a.s.) to the limit $E[\xi]$ that $E[\xi]$ is a significant parameter of the distribution of ξ. Related results are *Chebyshev's inequality*[6], which asserts that, for any $\varepsilon > 0$, however small,

$$\text{Prob}\left\{ |\psi_n - E[\xi]| \geqslant \left(\frac{\text{var}[\xi]}{n\varepsilon} \right)^{\frac{1}{2}} \right\} \leqslant \varepsilon; \tag{22}$$

and the *Central Limit Theorem*[7], which states that

$$\text{Prob}\left\{ |\psi_n - E[\xi]| \geqslant \left(\frac{\text{var}[\xi]}{n\varepsilon} \right)^{\frac{1}{2}} \right\} \to \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \int_{1/\sqrt{\varepsilon}}^{\infty} e^{-\frac{1}{2}x^2} dx, \tag{23}$$

as $n \to \infty$. These results indicate that $\text{var}[\xi]/n$ is a natural measure of the scale of the dispersion of the values of the r.v. ψ_n, and point to the major importance of the *normal distribution*[8], whose c.d.f. is

$$N(x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy. \tag{24}$$

Here, "Prob" denotes the probability in the product space (also denoted by "p^n") and both results hold for any r.v. which has finite mean and variance.

If ξ is a r.v. and $E[\xi] = \theta$, then ξ is often referred to as an **unbiased estimator** for (or of) the parameter θ of the distribution of ξ. We have seen that both ξ and ψ_n are unbiased estimators of $E[\xi]$. It is slightly less immediate that, if ξ and η are r.v. (possibly the same), then an unbiased estimator for $\mathrm{cov}[\xi, \eta]$ (possibly $\mathrm{var}[\xi]$) is

$$T_n(s_1, s_2, \ldots, s_n) = \frac{1}{n-1}\left\{\sum_{i=1}^{n} \xi(s_i)\eta(s_i) - \frac{1}{n}\sum_{i=1}^{n} \xi(s_i) \sum_{j=1}^{n} \eta(s_j)\right\}. \quad (25)$$

[[*Proof*: Write $X_i = \xi(s_i)$ and $Y_i = \eta(s_i)$. Then

$$E\left[\frac{1}{n-1}\left\{\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j\right\}\right] = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[X_iY_i] - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iY_j]\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} (E[\xi]E[\eta] + \mathrm{cov}[\xi, \eta]) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} (E[\xi]E[\eta] + \delta_{ij}\,\mathrm{cov}[\xi,$$

$$\eta])\right\} = \frac{n - n^2/n}{n-1} E[\xi]E[\eta] + \frac{n - n/n}{n-1}\mathrm{cov}[\xi, \eta] = \mathrm{cov}[\xi, \eta]. \,]]$$

In order to compute the estimator ψ_n efficiently, one must accumulate the sum $S_{\xi n} = \Sigma_{i=1}^{n}\xi(s_i)$, by the recurrence relation

$$S_{\xi 0} = 0, \quad S_{\xi n} = S_{\xi(n-1)} + \xi(s_n), \quad (26)$$

and, for the similar estimator of $E[\eta]$, we use $S_{\eta n} = \Sigma_{i=1}^{n}\eta(s_i)$, by

$$S_{\eta 0} = 0, \quad S_{\eta n} = S_{\eta(n-1)} + \eta(s_n). \quad (27)$$

It is tempting to assume that T_n is similarly computed by accumulating the sum $\Sigma_{i=1}^{n}\xi(s_i)\eta(s_i)$; but this is not efficient: we must obtain the relatively small difference of two large sums to compute T_n, which leads to large round-off errors. Instead, we accumulate the sum

$$T_{\xi\eta n} = \sum_{q=2}^{n} \frac{q}{q-1}\left\{\xi(s_q) - \frac{1}{q}\sum_{i=1}^{q}\xi(s_i)\right\}\left\{\eta(s_q) - \frac{1}{q}\sum_{j=1}^{q}\eta(s_j)\right\}, \qquad (28)$$

by the recurrence relation

$$T_{\xi\eta 1} = 0, \quad T_{\xi\eta n} = T_{\xi\eta(n-1)} + \frac{n}{n-1}\left\{\xi(s_n) - \frac{1}{n}S_{\xi n}\right\}\left\{\eta(s_n) - \frac{1}{n}S_{\eta n}\right\}; \qquad (29)$$

so that

$$E[\tfrac{1}{n}S_{\xi n}] = E[\xi],\ E[\tfrac{1}{n}S_{\eta n}] = E[\eta],\ \text{and}\ E[\tfrac{1}{n-1}T_{\xi\eta n}] = \operatorname{cov}[\xi, \eta]. \qquad (30)$$

[[*Proof*: The first two results are immediate from (20). For the last, by (29),

$$\begin{aligned} T_{\xi\eta n} - T_{\xi\eta(n-1)} &= \frac{n}{n-1}\left\{\frac{n-1}{n}\xi(s_n) - \frac{1}{n}\Sigma_{i=1}^{n-1}\xi(s_i)\right\}\left\{\frac{n-1}{n}\eta(s_n) - \frac{1}{n}\Sigma_{j=1}^{n-1}\eta(s_j)\right\} \\ &= \frac{n-1}{n}\xi(s_n)\eta(s_n) - \frac{1}{n}\xi(s_n)\Sigma_{i=1}^{n-1}\eta(s_i) - \frac{1}{n}\eta(s_n)\Sigma_{i=1}^{n-1}\xi(s_i) + \\ &\quad \frac{1}{n(n-1)}\Sigma_{i=1}^{n-1}\Sigma_{j=1}^{n-1}\xi(s_i)\eta(s_j), \end{aligned} \qquad (31)$$

while, by (25),

$$\begin{aligned} (n-1)T_n - (n-2)T_{n-1} &= \Sigma_{i=1}^{n}\xi(s_i)\eta(s_i) - \frac{1}{n}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\xi(s_i)\eta(s_j) \\ &\quad - \Sigma_{i=1}^{n-1}\xi(s_i)\eta(s_i) + \frac{1}{n-1}\Sigma_{i=1}^{n-1}\Sigma_{j=1}^{n-1}\xi(s_i)\eta(s_j) = \xi(s_n)\eta(s_n) \\ &\quad - \frac{1}{n}\xi(s_n)\eta(s_n) - \frac{1}{n}\xi(s_n)\Sigma_{i=1}^{n-1}\eta(s_i) - \frac{1}{n}\eta(s_n)\Sigma_{i=1}^{n-1}\xi(s_i) \\ &\quad - \frac{1}{n}\Sigma_{i=1}^{n-1}\Sigma_{j=1}^{n-1}\xi(s_i)\eta(s_j) + \frac{1}{n-1}\Sigma_{i=1}^{n-1}\Sigma_{j=1}^{n-1}\xi(s_i)\eta(s_j), \end{aligned} \qquad (32)$$

and clearly (31) and (32) are identical, and further, $T_1 = 0$; so that

$$T_n = \frac{1}{n-1}T_{\xi\eta n}, \qquad (33)$$

and (30) follows.]] *The gain in accuracy is well worth the additional arithmetic required by* (29).

The Central Limit Theorem[7] tells us that, if ξ is *any* r.v. and ψ_n is the average of a sample of n independent values of ξ, then the *standardized variable*

$$\psi_n^\circ = \frac{\psi_n - E[\psi_n]}{\sqrt{\text{var}[\psi_n]}} = \frac{\psi_n - E[\xi]}{\sqrt{\text{var}[\xi]/n}} \tag{34}$$

has a distribution approaching the normal distribution as $n \to \infty$ [see (23) and (24)]; or, in other words, the average ψ_n itself has a distribution whose c.d.f. is asymptotic to

$$N(E[\xi] + x\sqrt{\text{var}[\xi]/n}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{E[\xi]+x\sqrt{\text{var}[\xi]/n}} e^{-\frac{1}{2}y^2}\, dy$$

$$= \frac{1}{\sqrt{2\pi\text{var}[\xi]/n}} \int_{-\infty}^{x} e^{-\frac{1}{2}(z-E[\xi])^2 n/\text{var}[\xi]}\, dz, \tag{35}$$

where we have put $z = E[\xi] + y\sqrt{\text{var}[\xi]/n}$. The normal distribution is therefore often assumed to hold, where the r.v. observed can reasonably be assumed to be the result of many independent, identically distributed r.v., whose effects are summed. This applies to such phenomena as round-off errors and experimental perturbations of measurements.

We note also that the sums $S_{\xi n}$ and $S_{\eta n}$ are respectively distributed with c.d.f. asymptotic to $N(nE[\xi] + x\sqrt{n\ \text{var}[\xi]})$ and $N(nE[\eta] + x\sqrt{n\ \text{var}[\eta]})$. It is now reasonable to ask what the distributions of T_n and the sum $T_{\xi\eta n}$ look like, as $n \to \infty$.

First, we observe that there is an extension[9] of the Central Limit Theorem to a vector ξ of r.v. on a probability space $(S, \boldsymbol{S}, p)$. If, by repeated independent trials, we obtain outcomes $s_1, s_2, \ldots, s_n, \ldots$ and

compute the values of the r.v. $\xi(s_i) = (\xi_1(s_i), \xi_2(s_i), \ldots, \xi_m(s_i))$ for $i = 1, 2, \ldots, n, \ldots$, we may define new vectors $\psi_1, \psi_2, \ldots$, by

$$\psi_n(s_1, s_2, \ldots, s_n) = \frac{1}{n} \sum_{i=1}^{n} \xi(s_i) \tag{36}$$

[compare (19)]; and then, if the r.v. ξ have the expectation vector $E[\xi]$ and the variance-covariance matrix $\mathbf{V}[\xi]$ defined in (10) and (11), then the averages ψ_n have a distribution whose joint c.d.f. (12) is asymptotic as $n \to \infty$ to that of the *multivariate normal distribution*

$$N_m(\boldsymbol{x};\ E[\xi],\ \mathbf{V}[\xi]/n) = \left(\frac{1}{(2\pi/n)^m \det\mathbf{V}[\xi]}\right)^{\frac{1}{2}} \int_{-\infty}^{x_1} dy_1 \int_{-\infty}^{x_2} dy_2 \cdots \int_{-\infty}^{x_m} dy_m\ e^{-\frac{1}{2}n(\boldsymbol{y}-E[\xi])^{\mathrm{T}}\mathbf{V}[\xi]^{-1}(\boldsymbol{y}-E[\xi])}, \tag{37}$$

with the same expectation and variance-covariance parameters. [Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are m-dimensional column vectors, $^{\mathrm{T}}$ is the transposing operator, in this case converting a column to a row vector, and $\det\mathbf{V}[\xi]$ and $\mathbf{V}[\xi]^{-1}$ are respectively the determinant (assumed non-zero) and reciprocal of the variance-covariance matrix $\mathbf{V}[\xi]$.] In particular, two r.v. ξ and η with given means, variances, and covariance will yield averages $\psi_{\xi n}$ (hitherto written ψ_n) and $\psi_{\eta n}$ whose distribution approaches a bivariate normal distribution with the same means and with variances and covariance $1/n$ of those for ξ and η. Assuming this asymptotic normal distribution, we write

$$\left.\begin{aligned} &E[\xi] = a, \quad \mathbf{V}[\xi] = \begin{pmatrix} A & r_{\xi\eta}\sqrt{AB} \\ r_{\xi\eta}\sqrt{AB} & B \end{pmatrix}, \quad r = r_{\xi\eta}, \text{ for brevity,} \\ &E[\eta] = b, \qquad\qquad\qquad\qquad \Delta = \det\mathbf{V}[\xi] = AB(1 - r^2), \\ &\text{and } \mathbf{V}[\xi]^{-1} = \begin{pmatrix} B/\Delta & -r\sqrt{AB}/\Delta \\ -r\sqrt{AB}/\Delta & A/\Delta \end{pmatrix}, \text{ with } X = x - a, \quad Y = y - b. \end{aligned}\right\} \tag{38}$$

Then, by (37), the asymptotic joint c.d.f. of $\psi_{\xi n}$ and $\psi_{\eta n}$ is

$$N_2(x, y; a, b, \mathbf{V}[\xi]/n) = \frac{1}{(2\pi/n)\sqrt{\Delta}} \int_{-\infty}^{x} dx \int_{-\infty}^{y} dy \; e^{\frac{-n}{2(1-r^2)}[\frac{X^2}{A} - 2r\frac{XY}{\sqrt{AB}} + \frac{Y^2}{B}]} \tag{39}$$

and the parameter $r = r_{\xi\eta}$ is called the **correlation** of ξ and η: we thus have, by (38),

$$r_{\xi\eta} = \frac{\mathrm{cov}[\xi, \eta]}{\sqrt{\mathrm{var}[\xi]\mathrm{var}[\eta]}}. \tag{40}$$

A further statistical concept is useful here: it is that of the **characteristic function** of a r.v. If a r.v. ξ has a c.d.f. F_ξ, its characteristic function is defined to be

$$\Psi_\xi(t) = \int_{-\infty}^{+\infty} e^{ixt} dF_\xi(x). \tag{41}$$

[This is connected with the *moments* of the distribution of ξ: by formal expansion, we see that

$$\Psi_\xi(t) = \int_{-\infty}^{+\infty} dF_\xi + i\int_{-\infty}^{+\infty} xt\, dF_\xi - \frac{1}{2}\int_{-\infty}^{+\infty} x^2 t^2 dF_\xi - \frac{i}{6}\int_{-\infty}^{+\infty} x^3 t^3 dF_\xi + \ldots$$

$$= 1 + itE[\xi] - \frac{1}{2}t^2 E[\xi^2] - \frac{i}{6}t^3 E[\xi^3] + \ldots, \tag{42}$$

and the parameters $E[\xi^k]$ are the k-th moments of ξ. (We have already encountered the first moment, the *expectation*, and the second moment, which is $\mathrm{var}[\xi] + E[\xi]^2$.)] Note that the characteristic function *necessarily exists* for any r.v., and *uniquely determines* the distribution of the r.v.[10] Indeed, we get that

$$F_\xi(x) = F_\xi(a) + \lim_{K\to\infty} \frac{1}{2\pi} \int_{-K}^{+K} \frac{e^{-iat} - e^{-ixt}}{it} \Psi_\xi(t) dt, \tag{43}$$

if F_ξ is continuous in the closed interval $[a, x]$; and if ξ has a probability density ρ_ξ, then

$$\rho_\xi(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ixt} \Psi_\xi(t)\, dt. \tag{44}$$

We may now use (39) to compute the characteristic functions of

$$\alpha = n(\psi^{\circ}_{\xi n})^2, \quad \beta = n(\psi^{\circ}_{\eta n})^2, \quad \gamma = n\psi^{\circ}_{\xi}\,\psi^{\circ}_{\eta}\,. \tag{45}$$

Thus:

$$\begin{aligned}\Psi_{\alpha}(t) &= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \frac{1}{(2\pi/n)\sqrt{\Delta}}\, e^{\frac{-n}{2(1-r^2)}\left[\frac{X^2}{A} - 2r\frac{XY}{\sqrt{AB}} + \frac{Y^2}{B}\right] + i n\frac{X^2}{A}t} \\ &= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \frac{1}{(2\pi/n)\sqrt{\Delta}}\, e^{\frac{-n}{2(1-r^2)}\left[\left(\frac{Y}{\sqrt{B}} - \frac{rX}{\sqrt{A}}\right)^2 + \frac{X^2}{A}(1-r^2)(1-2it)\right]} \\ &= \int_{-\infty}^{+\infty} dx \frac{1}{\sqrt{2\pi A/n}}\, e^{-(n/2)\frac{X^2}{A}(1-2it)} = \frac{1}{\sqrt{1 - 2it}},\end{aligned} \tag{46}$$

where we have used the facts that (by (38)) $\Delta = AB(1 - r^2)$ and

$$\int_{-\infty}^{+\infty} du\; e^{-\frac{1}{2}(u-c)^2/K} = \sqrt{2\pi K}. \tag{47}$$

By the symmetry of (46), we obtain similarly that

$$\Psi_{\beta}(t) = \frac{1}{\sqrt{1 - 2it}}\,, \tag{48}$$

and this common characteristic function inverts to yield the well-known density

$$\rho_{\alpha}(x) = \rho_{\beta}(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x} x^{-\frac{1}{2}}, \tag{49}$$

called the χ^2 (or **chi-squared**) distribution. [Note that we could have obtained (49) by simply considering the distribution of ξ^2 when ξ is distributed with the standard normal distribution (24); i.e., by replacing y^2 by x and so dy by $d(\sqrt{x}) = \frac{1}{2}x^{-\frac{1}{2}}dx$, in (24), and then noting that the same element dx corresponds to *two* elements dy (with $\pm y$), so that the density must be doubled.] Similarly, we get that

$$\Psi_\gamma(t) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \; \frac{1}{(2\pi/n)\sqrt{AB(1-r^2)}} \; e^{\frac{-n}{2(1-r^2)}\left[\frac{Y}{\sqrt{B}} - \frac{X}{\sqrt{A}}(it+r-itr^2)\right]^2}$$
$$\times \; e^{\frac{-n}{2(1-r^2)}\frac{X^2}{A}\left[1-(it+r-itr^2)^2\right]}$$

$$= \frac{\sqrt{(2\pi/n)B(1-r^2)}\;\sqrt{(2\pi/n)A/[1-2irt+(1-r^2)t^2]}}{(2\pi/n)\sqrt{AB(1-r^2)}}$$

$$= \frac{1}{\sqrt{1-2irt+(1-r^2)t^2}} \; . \qquad (50)$$

In this case, there also exists a probability density function, ρ_γ, but it is not well known or easily expressed in closed form. [It is related to certain Bessel functions.] The r.v. α, β, and γ are seen to be related to the second term in (25); but the first term is, again by the Central Limit Theorem, asymptotically distributed like $N(\frac{n}{n-1}\{E[\xi\eta] + x\sqrt{\text{var}[\xi\eta]/n}\})$ and the distribution of T_n and $T_{\xi\eta n}$ is not at all easily obtained, in general. *This is why statisticians resort to assumptions of normality of distribution in circumstances in which justification is only intuitive or tenuous, at best.* Even in these restricted circumstances, the derivation of the sampling distributions are complicated and the results are sometimes incomplete.

It must be emphasized that one possible source of error is indeed an unwarranted assumption of normality in the underlying distribution.

THE NORMALITY ASSUMPTION

Henceforth, we shall assume that the underlying distribution is indeed normal, and that repeated independent trials are made. The joint c.d.f. of

n independent trials of a r.v. ξ is then [see (18) and (37)]

$$\prod_{i=1}^{n} N(x_i;\ \mathrm{E}[\xi],\ \mathrm{var}[\xi]) = (2\pi\mathrm{var}[\xi])^{-\frac{1}{2}n} \int_{-\infty}^{x_1} dy_1 \int_{-\infty}^{x_2} dy_2 \cdots \int_{-\infty}^{x_n} dy_n\ e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mathrm{E}[\xi])^2/\mathrm{var}[\xi]}; \quad (51)$$

whence we see that [by (41)] the characteristic function of $\psi_{\xi n}$ is

$$\begin{aligned}\Psi_{\psi_{\xi n}}(t) &= \left\{(2\pi\mathrm{var}[\xi])^{-\frac{1}{2}} \int_{-\infty}^{+\infty} dy\ e^{-\frac{1}{2}\left(\frac{(y-\mathrm{E}[\xi])^2}{\mathrm{var}[\xi]} - 2it\frac{y}{n}\right)}\right\}^n \\ &= \left\{(2\pi\mathrm{var}[\xi])^{-\frac{1}{2}} \int_{-\infty}^{+\infty} dy\ e^{-\frac{1}{2}\left(\left(\frac{y-\mathrm{E}[\xi]}{\sqrt{\mathrm{var}[\xi]}} - it\frac{\sqrt{\mathrm{var}[\xi]}}{n}\right)^2 + \frac{t^2}{n^2}\mathrm{var}[\xi] - 2it\frac{\mathrm{E}[\xi]}{n}\right)}\right\}^n \\ &= e^{-\frac{n}{2}\left(\frac{t^2}{n^2}\mathrm{var}[\xi] - 2it\frac{\mathrm{E}[\xi]}{n}\right)}, \end{aligned} \quad (52)$$

by (47); and so the probability density of $\psi_{\xi n}$ is [by (44)]

$$\begin{aligned}\rho_{\psi_{\xi n}}(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} dt\ e^{-\frac{n}{2}\left(\left(\frac{t}{n}\sqrt{\mathrm{var}[\xi]} + i\frac{(x-\mathrm{E}[\xi])}{\sqrt{\mathrm{var}[\xi]}}\right)^2 + \frac{(x-\mathrm{E}[\xi])^2}{\mathrm{var}[\xi]}\right)} \\ &= (2\pi\mathrm{var}[\xi]/n)^{-\frac{1}{2}}\ e^{-\frac{1}{2}\frac{(x-\mathrm{E}[\xi])^2}{\mathrm{var}[\xi]/n}}, \end{aligned} \quad (53)$$

the density of $N(x;\ \mathrm{E}[\xi],\ \mathrm{var}[\xi]/n)$, as we would expect [see (35).] The characteristic function of $\sum_{i=1}^{n}(\xi(s_i) - \mathrm{E}[\xi])^2/\mathrm{var}[\xi] = \chi^2_{\xi n}$ is similarly

$$\begin{aligned}\Psi_{\chi^2_{\xi n}}(t) &= \left\{(2\pi\mathrm{var}[\xi])^{-\frac{1}{2}} \int_{-\infty}^{+\infty} dy\ e^{-\frac{1}{2}\frac{(y-\mathrm{E}[\xi])^2}{\mathrm{var}[\xi]}(1-2it)}\right\}^n \\ &= (1-2it)^{-\frac{1}{2}n}, \end{aligned} \quad (54)$$

which yields the density

$$\rho_{\chi^2_{\xi n}}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-\frac{1}{2}x}\ x^{(n/2)-1} \quad (55)$$

of the **chi-squared distribution** with n **degrees of freedom** [d.f.] (compare (49), the density with *one* d.f.)[11] [The *gamma function* $\Gamma(z)$ is defined by

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt; \tag{56}$$

so that (by integration by parts)

$$\Gamma(z) = (z - 1)\,\Gamma(z - 1), \tag{57}$$

whence, for integer n, $\Gamma(n) = (n - 1)!$

and $$\Gamma(n + \tfrac{1}{2}) = (n-\tfrac{1}{2})(n-\tfrac{3}{2})\ldots(\tfrac{3}{2})(\tfrac{1}{2})\sqrt{\pi}, \tag{58}$$

since $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Because of the obvious relation between (55) and (56), we see that the chi-squared distribution is related to the so-called *gamma distribution*.]

A more complicated derivation[12] shows that the sum $T_{\xi\xi n}/\mathrm{var}[\xi]$ is itself distributed as $\chi^2_{\xi(n-1)}$, the chi-squared r.v. with n-1 d.f. Of course, corresponding r.v. derived from another r.v. η are distributed exactly analogously; and if ξ and η are distributed according to a bivariate normal distribution of the form (39) with $n = 1$, it can be shown that the averages $\psi_{\xi n}$ and $\psi_{\eta n}$ have the exact distribution (39), with the same means as ξ and η and variances and covariance less by a factor $1/n$, so that the correlation is also the same; while the quadratic estimators have a joint density (*independent of the distribution of the averages!*) whose exact form is known[13], but need not concern us here.

A final distribution will be of interest in what follows. We shall be concerned with testing whether the variance estimators V_1 with n_1 d.f. and V_2 with n_2 d.f. are obtained from samples of distributions with the *same variance*. The test is applied by computing V_1/V_2, the **variance-ratio** , and

looking up the tabulated values, under n_1 and n_2, of the c.d.f. of this ratio when both estimates are obtained from samples of normal distributions with the same (unspecified) variance. The corresponding probability density is[14]

$$\rho_{V_1/V_2}(x) = \frac{\Gamma(\frac{n_1+n_2}{2})\ n_1^{\frac{1}{2}n_1}\ n_2^{\frac{1}{2}n_2}\ x^{\frac{1}{2}n_1-1}}{\Gamma(\frac{n_1}{2})\ \Gamma(\frac{n_2}{2})\ \ (n_1x + n_2)^{\frac{1}{2}(n_1+n_2)}}. \tag{59}$$

Tests of significance are to be interpreted as follows: Suppose that a certain *statistic* is observed (as the value of a r.v.) in a given sample. We wish to decide between two alternative *hypotheses*, H_0 and H_1, the first being a situation commonly occurring (such as: both variance estimates are drawn from normal distributions with the same variance) and the second indicating an event we wish to take note of (such as: the variances of the two populations differ.) On the basis of H_0, we compute the probability that the statistic V_1/V_2 should be as extreme as it is (i.e., as far from the expected value, in our example, 1.) This is the probability of an error of *Type I* (the hypothesis is true and we reject it.) An error of *Type II* (the hypothesis H_0 is false and we accept it) is also to be avoided, of course, but is usually harder to compute.

REGRESSION ANALYSIS

Suppose that we observe r.v. η, ξ_1, ξ_2, ..., ξ_m (the last forming an m-vector ξ) and that the conditional expectation of η, given that $\xi = \boldsymbol{x}$, is [compare (15) and (16)]

$$E[\eta \mid \xi = \boldsymbol{x}] = f(\boldsymbol{x}). \tag{60}$$

Since we cannot hope to determine the completely unspecified function f, we seek instead to approximate f by a suitable choice from a family of functions $\varphi(\boldsymbol{x}; \boldsymbol{a})$ with parameter-vector $\boldsymbol{a} = (\alpha_1, \alpha_2, \ldots, \alpha_g)$, selected *a priori* on the basis of physical intuition and convenience. ***Note that the choice of the family*** φ ***is crucial to the efficacy of the method and to the accuracy of the estimates obtained.***

The theoretical criterion for optimizing the estimate of f is the minimization of the variance of η from the *regression surface* [or regression *line*, when $m = 1$ and the vector $\boldsymbol{x}$ reduces to a single variable x.] Since, again, this variance is unknown, we seek instead to minimize the **sum of squares**

$$S_n(\boldsymbol{a}) = \sum_{j=1}^{n} \left\{\eta(s_j) - \varphi\big(\xi(s_j); \boldsymbol{a}\big)\right\}^2 W\big(\xi(s_j)\big), \tag{61}$$

arising from n repeated independent trials, where $W(\boldsymbol{x})$ is a *weight function* indicating the importance we attach to the value of η when $\xi = \boldsymbol{x}$. This may reflect both the *accuracy* of η at $\boldsymbol{x}$ (in terms of an estimate, at $\boldsymbol{x}$, of the *variance* of η) and the *cost* to us of an error at $\boldsymbol{x}$.

If we assume that φ is differentiable with respect to the α_h, we can use a Taylor expansion to show that, in the vicinity of $\boldsymbol{a}$ (at $\boldsymbol{a} + \delta$, with $|\delta| \leqslant \varepsilon$), as $\varepsilon \to 0$,

$$\begin{aligned} S_n(\boldsymbol{a} + \delta) = \sum_{j=1}^{n} \Bigg\{ & \Big(\eta_j - \varphi(\xi_j; \boldsymbol{a})\Big)^2 - 2\Big(\eta_j - \varphi(\xi_j; \boldsymbol{a})\Big) \sum_{h=1}^{g} \varphi_h(\xi_j; \boldsymbol{a})\, \delta_h \\ & - \Big(\eta_j - \varphi(\xi_j; \boldsymbol{a})\Big) \sum_{h=1}^{g} \sum_{k=1}^{g} \varphi_{hk}(\xi_j; \boldsymbol{a})\, \delta_h\, \delta_k \\ & + \sum_{h=1}^{g} \sum_{k=1}^{g} \varphi_h(\xi_j; \boldsymbol{a}) \varphi_k(\xi_j; \boldsymbol{a})\, \delta_h\, \delta_k + O(\varepsilon^3) \Bigg\} W(\xi_j); \end{aligned} \tag{62}$$

where we have written η_j for $\eta(s_j)$ and ξ_j for $\xi(s_j)$, for brevity, $\delta = (\delta_1, \delta_2, \ldots, \delta_g)$, $\varphi_h(\xi_j; a)$ is $\partial\varphi/\partial\alpha_h$ at ξ_j and a, and $\varphi_{hk}(\xi_j; a)$ is $\partial^2\varphi/\partial\alpha_h\partial\alpha_k$ at ξ_j and a; and $O(\varepsilon^3)$ is, as usual, a function of $s_1, s_2, \ldots, s_n, a, \delta$, which is bounded by a multiple of ε^3 in the neighborhood $|\delta| \leqslant \varepsilon$. To minimize $S_n(a)$, we must have that, for some $\varepsilon > 0$ and all δ such that $|\delta| \leqslant \varepsilon$,

$$S_n(a + \delta) \geqslant S_n(a); \tag{63}$$

and this necessitates that the terms in δ_h in (62) should vanish:

$$\sum_{j=1}^{n} \left(\eta_j - \varphi(\xi_j; a)\right) \varphi_h(\xi_j; a)\ W(\xi_j) = 0; \tag{64}$$

and further, if we define the $(g \times g)$ matrix $M(y, \boldsymbol{x}; \boldsymbol{a})$ as having components

$$\left(M(y, \boldsymbol{x}; \boldsymbol{a})\right)_{hk} = \left\{\varphi_h(\boldsymbol{x}; \boldsymbol{a})\varphi_k(\boldsymbol{x}; \boldsymbol{a}) - \left(y - \varphi(\boldsymbol{x}; \boldsymbol{a})\right)\varphi_{hk}(\boldsymbol{x}; \boldsymbol{a})\right\}W(\boldsymbol{x}), \tag{65}$$

then the matrix $\Sigma_{j=1}^{n} M(\eta_j, \xi_j; a)$ must be *non-negative definite* [i.e., the sum of all terms of order ε^2 in (62) should be non-negative: guaranteed iff no eigenvalue of the matrix is negative.] Indeed, if we are to be able to avoid consideration of terms in $\delta_h\delta_k\delta_l$, and so on, we must ask that the matrix be *positive-definite* [i.e., all eigenvalues of the matrix should be strictly positive.] Further, to avoid dependence of this condition on the specific observed values η_j and ξ_j, we are effectively compelled to ask that, for all y and $\boldsymbol{x}$ in some broad range encompassing all likely values of η_j and ξ_j, the matrix $M(y, \boldsymbol{x}; \boldsymbol{a})$ itself should be positive-definite. *Even this condition is by no means easy to verify, in general.*

Because of the mathematical difficulties outlined above, regression analysis (and indeed, theoretical physics in general) is limited in practice to rather simple models of possible relationships. Usually, we are restricted to selecting a set of g functions of $\boldsymbol{x}$, say $\phi_1(\boldsymbol{x})$, $\phi_2(\boldsymbol{x})$, ..., $\phi_g(\boldsymbol{x})$, which are believed, *on the basis of an understanding of the physical realities underlying the observations*, together with a great deal of professional intuition (i.e., guesswork!), to affect the values of η *linearly*; and defining the simple linear family of functions

$$\varphi(\boldsymbol{x};\ \boldsymbol{\alpha}) = \sum_{h=1}^{g} \alpha_h \phi_h(\boldsymbol{x}) = \alpha_1\phi_1(\boldsymbol{x}) + \alpha_2\phi_2(\boldsymbol{x}) + \ldots + \alpha_g\phi_g(\boldsymbol{x}). \quad (66)$$

For example, in the simplest case, when $m = 1$ and $\boldsymbol{x}$ is a single variable x, we may wish to set $\phi_h(x) = x^{h-1}$ and seek a polynomial relationship,

$$E[\eta|\ x] \approx \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \ldots + \alpha_g x^{g-1}. \quad (67)$$

Of course, this can be extended to polynomials in several variables; e.g.,

$$E[\eta|\ x_1, x_2, x_3] \approx \alpha_1 + \alpha_2 x_1 + \alpha_3 x_2 + \alpha_4 x_3 + \alpha_5 x_1^2 + \alpha_6 x_2^2 + \alpha_7 x_3^2 + \alpha_8 x_2 x_3 + \alpha_9 x_3 x_1 + \alpha_{10} x_1 x_2, \quad (68)$$

when only quadratic terms are considered: the number of terms grows rapidly with the degree and m.

Nevertheless, the model (66) can be very natural, as when η is the time required to complete a process and the ϕ_h are approximate formulae for the time required to complete distinct and consecutive parts of this process. However, in this case, if the parts may proceed concurrently, in parallel, then the model (66) is clearly a poor one; as it would be if the formulae ϕ_h were to be inappropriately chosen.

We shall limit ourselves henceforth to families of the linear form (66); so that

$$\varphi_h(\boldsymbol{x};\ \boldsymbol{a}) = \phi_h(\boldsymbol{x}) \quad \text{and} \quad \varphi_{hk}(\boldsymbol{x};\ \boldsymbol{a}) = 0. \tag{69}$$

Thus, for the matrix M of (65) and the derived matrix $\Sigma_{j=1}^{n}\mathrm{M}(\eta_j,\ \xi_j;\ \boldsymbol{a})$, we get

$$\big(\mathrm{M}(y,\ \boldsymbol{x};\ \boldsymbol{a})\big)_{hk} = \phi_h(\boldsymbol{x})\phi_k(\boldsymbol{x})W(\boldsymbol{x}) \tag{70}$$

and

$$\big(\Lambda(\eta,\ \Xi;\ \boldsymbol{a})\big)_{hk} = \sum_{j=1}^{n} \big(\mathrm{M}(\eta_j,\ \xi_j;\ \boldsymbol{a})\big)_{hk} = \Lambda_{hk}$$

$$= \sum_{j=1}^{n} \phi_h(\xi_j)\phi_k(\xi_j)W(\xi_j) = (\Phi \boldsymbol{W} \Phi^{\mathrm{T}})_{hk}, \tag{71}$$

where we write η for the column vector $(\eta_1,\ \eta_2,\ \ldots,\ \eta_n)$, Ξ for the $(n \times m)$ matrix with rows $\xi_1,\ \xi_2,\ \ldots,\ \xi_n$, $\boldsymbol{W}$ for the $(n \times n)$ diagonal matrix with components $(\boldsymbol{W})_{ij} = \delta_{ij}\ W(\xi_j)$, and Φ for the $(g \times n)$ matrix with components $(\Phi)_{hj} = \phi_h(\xi_j)$; so that (71) may be written

$$\Lambda = \Phi \boldsymbol{W} \Phi^{\mathrm{T}}\ . \tag{72}$$

Now, the conditions (64) become the *normal equations*,

$$\sum_{k=1}^{g} \left(\sum_{j=1}^{n} \phi_h(\xi_j)\phi_k(\xi_j)W(\xi_j) \right) \alpha_k = \sum_{j=1}^{n} \eta_j \phi_h(\xi_j)W(\xi_j), \tag{73}$$

which, with the aid of our matrix notation, we may write as

$$\Lambda \boldsymbol{a} = \lambda, \tag{74}$$

where

$$\lambda = \Phi \boldsymbol{W} \eta. \tag{75}$$

It is notable that the normal equations are obtained by multiplying the idealized equation $\Phi^{\mathrm{T}}\boldsymbol{a} = \eta$ on the left by $\Phi \boldsymbol{W}$. This idealized equation asserts that $\boldsymbol{a}$ combines the columns of Φ^{T} to yield η; i.e., that each $\eta_j = \Sigma_{h=1}^{g}\alpha_h\phi_h(\xi_j)$; so that the fit of η by (66) is exact at every ξ_j.

If the matrix Λ is *regular* (i.e. invertible), then it has a reciprocal Λ^{-1} and we may solve (74) for the unknown vector a in terms of the computable matrix Λ^{-1} and vector λ, in the form

$$a = \Lambda^{-1}\lambda. \tag{76}$$

Of course, in practice, we would not compute the reciprocal, but rather solve the equations by Gaussian elimination, successive over-relaxation, or some other efficient computational technique. [The reader is referred to a treatise on numerical methods for further details on this point.]

If the functions $\phi_h(\boldsymbol{x})$ are *linearly independent*, in the sense that the only solution of the equation

$$\alpha_1\phi_1(\boldsymbol{x}) + \alpha_2\phi_2(\boldsymbol{x}) + \ldots + \alpha_g\phi_g(\boldsymbol{x}) = 0 \tag{77}$$

for *almost all* $\boldsymbol{x}$ is $\quad \alpha_1 = \alpha_2 = \ldots = \alpha_g = 0 \tag{78}$

[in the sense that, if A is the set of values of $\boldsymbol{x}$ for which (77) has a solution $a \neq \mathbf{0}$, then $p(\{s \in S\colon \xi(s) \in A\}) = 0$], and if the diagonal elements of $\boldsymbol{W}$ are all strictly positive [so that the matrix $\boldsymbol{W}^{\frac{1}{2}}$ with components $\delta_{ij}\sqrt{W(\xi_j)}$ is real and has all its diagonal elements positive, and $\boldsymbol{W} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}}$], and if the number n of data is not less than the number g of parameters in a; then, for any g-dimensional column vector $\gamma \neq \mathbf{0}$, we have $\gamma^T\Lambda\gamma = \gamma^T\Phi\boldsymbol{W}\Phi^T\gamma = (\boldsymbol{W}^{\frac{1}{2}}\Phi^T\gamma)^T\boldsymbol{W}^{\frac{1}{2}}\Phi^T\gamma = \|\boldsymbol{W}^{\frac{1}{2}}\Phi^T\gamma\|^2 \geqslant 0$. Now, the vector $\boldsymbol{W}^{\frac{1}{2}}\Phi^T\gamma$ has, for its j-th component, $\sqrt{W(\xi_j)}$ times the j-th component of the vector $\Phi^T\gamma$ (and $\sqrt{W(\xi_j)} > 0$); so the former vector is non-null if the latter vector is non-null; and since γ is non-null, it follows that $\Phi^T\gamma$ is also non-null with probability one. Therefore Λ is positive definite with probability one.

We note in the argument above that it is necessary for it to be possible for the vectors $(\phi_h(\xi_1), \phi_h(\xi_2), \ldots, \phi_h(\xi_n))$ —— the rows of the matrix Φ —— to be linearly independent, and the condition for this is that $n \geqslant g$. Also, if ζ is an eigenvector of the matrix Λ, so that $\Lambda\zeta = \lambda\zeta$; then clearly $\zeta^T\Lambda\zeta = \zeta^T\lambda\zeta = \lambda\|\zeta\|^2 > 0$; so that $\lambda > 0$ (i.e., all eigenvalues of Λ are strictly positive) and Λ is invertible. We conclude that *the equation* (74) *is almost surely uniquely soluble for* a *if the above conditions hold.*

We return to (61): the sum of squares for the family (66) takes the form

$$S_n(a) = \sum_{j=1}^{n} \left\{\eta_j - \sum_{h=1}^{g} \alpha_h \phi_h(\xi_j)\right\}^2 W(\xi_j) = (\eta - \Phi^T a)^T W (\eta - \Phi^T a); \quad (79)$$

and the minimal value obtained from the normal equations is (by (75), (76))

$$\begin{aligned} S_n(a_{\min}) &= (\eta - \Phi^T\Lambda^{-1}\Phi W\eta)^T W (\eta - \Phi^T\Lambda^{-1}\Phi W\eta) \\ &= \eta^T W\eta - \eta^T W\Phi^T\Lambda^{-1}\Phi W\eta, \end{aligned} \quad (80)$$

since $\eta^T W\Phi^T\Lambda^{-1}\Phi W\Phi^T\Lambda^{-1}\Phi W\eta = \eta^T W\Phi^T\Lambda^{-1}\Phi W\eta$, by (72). By subtraction, we obtain that

$$S_n(a) = [\eta^T W\eta - \eta^T W\Phi^T\Lambda^{-1}\Phi W\eta] + (a - \Lambda^{-1}\Phi W\eta)^T \Phi W\Phi^T (a - \Lambda^{-1}\Phi W\eta), \quad (81)$$

which exhibits the minimality of (80) explicitly, since $\Lambda = \Phi W\Phi^T$ is positive definite.

What we have obtained above is the set of parameters of the family (66) for which the *sample sum of squares of deviations from the regression function* (61) is minimized. This is sometimes referred to as **least-squares approximation**, and we note that the *statistical properties* of η and ξ do not play a role in the formulation: it is simply a method of optimal approximation.

Now suppose that η *is distributed with a normal distribution about a regression function* $\Sigma_{h=1}^{g}\beta_h\phi_h(\boldsymbol{x})$ *as mean, with variance* $1/W(\boldsymbol{x})$. ***It is extremely important to realize that this is a very strong assumption which may simply not be valid in a given case.*** We observe that this hypothesis fixes the weight function W, which has hitherto been unspecified, and *which may be governed by considerations of cost, as well as variance, if it is to be realistic.* It follows from our assumption that

$$\zeta(\boldsymbol{x}) = \left\{\eta - \sum_{h=1}^{g}\beta_h\phi_h(\boldsymbol{x})\right\}\sqrt{W(\boldsymbol{x})} \tag{82}$$

is distributed (*without regard to the value of* $\boldsymbol{x}$) with a standard normal distribution (24) [with mean 0 and variance 1]; so that the sum of squares $S_n(\beta)$ will be distributed with a chi-square distribution with n d.f. [see (55).] With our previous notation, we may put

$$\zeta = W^{\frac{1}{2}}(\eta - \Phi^{T}\beta) \tag{83}$$

for the column vector $(\zeta_1, \zeta_2, \ldots, \zeta_n) = (\zeta(s_1), \zeta(s_2), \ldots, \zeta(s_n))$, where it is understood that the random sampling of s entails the corresponding value of $\boldsymbol{x} = \xi(s)$, as well as of $\eta(s)$. More particularly, since the distribution of ζ is mathematically independent of $\boldsymbol{x}$ [i.e., the conditional distribution of ζ, given $\boldsymbol{x}$, does not depend on the value of $\boldsymbol{x}$], it follows that the absolute distribution of ζ is the same as the conditional, and that ζ is independent of ξ, if we replace the variable $\boldsymbol{x}$ by the r.v. ξ in (82).

From (83), we get $$\eta = W^{-\frac{1}{2}}\zeta + \Phi^{T}\beta; \tag{84}$$

whence $$a_{\min} = \Lambda^{-1}\Phi W\eta = \Lambda^{-1}\Phi W^{\frac{1}{2}}\zeta + \beta. \tag{85}$$

Noting that, by (10) and (11), we can write in general that

$$\mathbf{V}[\xi] = E[(\xi - E[\xi])(\xi - E[\xi])^T], \tag{86}$$

so that, for ζ,

$$E[\zeta] = \mathbf{0} \quad \text{and} \quad \mathbf{V}[\zeta] = E[\zeta\zeta^T] = \mathbf{I}; \tag{87}$$

we see that, since ζ is independent of Ξ,

$$E[a_{\min}] = \Lambda^{-1}\Phi W^{\frac{1}{2}}E[\zeta] + \beta = \beta \tag{88}$$

and

$$\mathbf{V}[a_{\min}] = E[(a_{\min} - \beta)(a_{\min} - \beta)^T] = E[\Lambda^{-1}\Phi W^{\frac{1}{2}}E[\zeta\zeta^T]W^{\frac{1}{2}}\Phi^T\Lambda^{-1}]$$
$$= E[\Lambda^{-1}\Phi W^{\frac{1}{2}}W^{\frac{1}{2}}\Phi^T\Lambda^{-1}] = E[\Lambda^{-1}]. \tag{89}$$

when we assume the Ξ to be given fixed values; i.e., we are looking at the distribution of $a_{\min}$ for given Ξ.

Returning once again to the sums of squares, we see from (81) that

$$S_n(\beta) = S_n(a_{\min}) + (\beta - a_{\min})^T\Lambda(\beta - a_{\min}), \tag{90}$$

and since, by (85), $a_{\min}$ is a linear combination of normal r.v. and therefore itself normal; we conclude that the second term on the right of (90) is distributed as chi-squared with g d.f. We note, further, that

$$E[S_n(a_{\min})] = E[(\eta - \Phi^T a_{\min})^T W(\eta - \Phi^T a_{\min})]$$
$$= E[(W^{-\frac{1}{2}}\zeta - \Phi^T\Lambda^{-1}\Phi W^{\frac{1}{2}}\zeta)^T W(W^{-\frac{1}{2}}\zeta - \Phi^T\Lambda^{-1}\Phi W^{\frac{1}{2}}\zeta)$$
$$= E[\zeta^T W^{-\frac{1}{2}}WW^{-\frac{1}{2}}\zeta] - E[\zeta^T W^{-\frac{1}{2}}W\Phi^T\Lambda^{-1}\Phi W^{\frac{1}{2}}\zeta]$$
$$= E[\zeta^T\zeta] - E[\zeta^T W^{\frac{1}{2}}\Phi^T\Lambda^{-1}\Phi W^{\frac{1}{2}}\zeta]$$
$$= \Sigma_{j=1}^{n}E[\zeta_j^2]$$
$$- \Sigma_{i=1}^{n}\Sigma_{h=1}^{g}\Sigma_{k=1}^{g}\Sigma_{j=1}^{n}(\Lambda^{-1})_{hk}(\Phi)_{hi}(\Phi)_{kj}\sqrt{W(\xi_i)W(\xi_j)}E[\zeta_i\zeta_j]$$
$$= n - \Sigma_{h=1}^{g}\Sigma_{k=1}^{g}\Sigma_{j=1}^{n}(\Lambda^{-1})_{hk}(\Phi)_{hj}(\Phi)_{kj}(W)_{jj} = n - g, \tag{91}$$

where we use (87) and (72). Consider further the matrix

$$\Upsilon = \mathbf{I} - W^{\frac{1}{2}}\Phi^T\Lambda^{-1}\Phi W^{\frac{1}{2}} \tag{92}$$

for which

$$S_n(a_{\min}) = \zeta^T\Upsilon\zeta. \tag{93}$$

Since $S_n(a_{\min}) \geqslant 0$, the matrix Υ is non-negative definite [with eigenvalues all non-negative real numbers]; and we can easily verify that (by (72))

$$\Upsilon^2 = \Upsilon. \tag{94}$$

Since the eigenvalues of Υ^2 are the squares of those of Υ, it follows that the eigenvalues of Υ can only be 0 or 1. It now follows from the theorem on diagonalizing symmetric matrices and quadratic forms by orthogonal transformations[15] that (since Υ is clearly symmetric: $\Upsilon^{\mathrm{T}} = \Upsilon$) there is an orthogonal matrix $\boldsymbol{J}$ [such that $\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J} = \mathbf{I} = \boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}$] such that $\boldsymbol{J}^{\mathrm{T}}\Upsilon\boldsymbol{J}$ is diagonal with diagonal entries 0 and 1 only. We note that the corresponding transformation applied to the unit matrix $\mathbf{I}$ leaves it invariant [$\boldsymbol{J}^{\mathrm{T}}\mathbf{I}\boldsymbol{J} = \mathbf{I}$.] We now see that the characteristic function of $S_n(a_{\min})$ will be

$$\begin{aligned}\Psi_{S_n(a_{\min})}(t) &= (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{+\infty} dy_1 \cdots \int_{-\infty}^{+\infty} dy_n e^{-\frac{1}{2}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2it\boldsymbol{y}^{\mathrm{T}}\Upsilon\boldsymbol{y})} \\ &= (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{+\infty} dz_1 \cdots \int_{-\infty}^{+\infty} dz_n e^{-\frac{1}{2}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} - 2it\boldsymbol{z}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}}\Upsilon\boldsymbol{J}\boldsymbol{z})},\end{aligned}$$

where we have substituted $\boldsymbol{z} = \boldsymbol{J}^{\mathrm{T}}\boldsymbol{y}$ and note that the Jacobian of the transformation is 1. Now, $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z} = \sum_{j=1}^{n} z_j^2$, while $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}}\Upsilon\boldsymbol{J}\boldsymbol{z} = \sum_{j=1}^{n} \upsilon_j z_j^2$, where the υ_j are the 0 or 1 eigenvalues of Υ. For $\upsilon_j = 0$, the integral reduces to $(2\pi)^{\frac{1}{2}}$, while for $\upsilon_j = 1$, the corresponding integral gives a factor of $[2\pi/(1 - 2it)]^{\frac{1}{2}}$. Thus, if there are r unit eigenvalues, the characteristic function is

$$\Psi_{S_n(a_{\min})}(t) = (1 - 2it)^{-\frac{1}{2}r} \tag{95}$$

[compare (41), (51), (54), (55), and (87)]; so that $S_n(a_{\min})$ is distributed as chi-squared with r d.f. It remains only to determine r. The *rank* of a matrix is unchanged by an orthogonal transformation; so the rank of Υ will equal that of $\boldsymbol{J}^{\mathrm{T}}\Upsilon\boldsymbol{J}$, which clearly equals r; and this will be n minus the

rank of the matrix $\boldsymbol{K} = \boldsymbol{J}^{\mathrm{T}}\boldsymbol{W}^{\frac{1}{2}}\Phi^{\mathrm{T}}\Lambda^{-1}\Phi\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{J}$ [Since $\boldsymbol{J}^{\mathrm{T}}\Upsilon\boldsymbol{J} = \mathbf{I} - \boldsymbol{K}$, $\boldsymbol{K}$ has $n - r$ unit diagonal elements.] Finally, we note that, since rank is not affected by multiplication by a non-singular matrix, the rank of $\boldsymbol{K}$ equals that of $\Phi^{\mathrm{T}}\Phi$, which equals that of Φ, which our assumptions make equal to g. Thus, finally,

$$r = n - g. \tag{96}$$

[For the theoretical basis of the rank argument, see, e.g., D3, ch. 5 (in particular, pp. 137 - 139.)] We have thus demonstrated that $S_n(a_{\min})$ *is distributed with the chi-squared distribution with n-g d.f.*

Again, we see that [by (79) and (80)]

$$S_n(\mathbf{0}) = \eta^{\mathrm{T}}\boldsymbol{W}\eta = S_n(a_{\min}) + \eta^{\mathrm{T}}\boldsymbol{W}\Phi^{\mathrm{T}}\Lambda^{-1}\Phi\boldsymbol{W}\eta = S_n(a_{\min}) + a_{\min}{}^{\mathrm{T}}\Lambda a_{\min}, \tag{97}$$

and we may interpret $S_n(\mathbf{0})$ as the "total sum of squares", $S_n(a_{\min})$ as the "sum of squares of residuals [after allowing for the regression]", and $a_{\min}{}^{\mathrm{T}}\Lambda a_{\min}$ as the "sum of squares due to the regression." We know that the first term is distributed as chi-squared with n d.f., the second as chi-squared with n-g d.f. (as we have just proved); and *if we postulate that* $\beta = \mathbf{0}$, the third term becomes $(a_{\min}-\beta)^{\mathrm{T}}\Lambda(a_{\min}-\beta)$, which we have already shown to be distributed as chi-squared with g d.f.

At this point, we require that the two terms on the right of (97) be *statistically independent*, whereupon we may form the variance-ratio

$$\frac{S_n(a_{\min})/(n - g)}{a_{\min}{}^{\mathrm{T}}\Lambda a_{\min}/g} \tag{98}$$

[see (59)], and test the hypothesis that two such different variance estimates derive from independent samples of the same population; i.e., that indeed $\beta = \mathbf{0}$, so that η is *not* dependent on the $\phi_h(\xi(s_j))$.

The independence of the components of a decomposition of a chi-squared variable into a sum of chi-squared variables as in (90) or (97) is the result of a thorem due to W. G. Cochran [see *Proceedings of the Cambridge Philosophical Society*, vol. 30 (1934) p. 178; also **C3**, §§(1)15.16-19, and **C6**, ch.4], which states that, *if the unit matrix* $\mathbf{I}$ *is decomposed into* k *symmetric matrices,*

$$\mathbf{I} = \sum_{i=1}^{k} \Omega_i , \tag{99}$$

such that the rank of Ω_i *is* $r(\Omega_i) = r_i$, *and* ζ *denotes* (as previously) *a vector of independent standard normal variates; and if we write*

$$Q_i = \zeta^T \Omega_i \zeta ; \tag{100}$$

then each of the following conditions implies the other two:

(a) the sum of the ranks r_i *of the* Ω_i *equals the order* n *of the vector* ζ *and matrices* Ω_i *and* $\mathbf{I}$*;*

(b) each of the r.v. Q_i *is distributed as chi-squared with* r_i *d.f.; and*

(c) all the r.v. Q_i *are independent.*

Proof: For any i, $\mathbf{I} = \Omega_i + (\mathbf{I} - \Omega_i)$. There is an orthogonal transformation which diagonalizes Ω_i, and this leaves $\mathbf{I}$ invariant. Since $\mathbf{I}$ and $J^T \Omega_i J$ are both diagonal, so is $J^T(\mathbf{I} - \Omega_i) J$; and since the rank of Ω_i is r_i, just $(n-r_i)$ diagonal elements of $J^T \Omega_i J$ vanish, so that the corresponding elements of $J^T(\mathbf{I} - \Omega_i) J$ are 1. On the other hand,

$$r(J^T(\mathbf{I} - \Omega_i) J) = r(\mathbf{I} - \Omega_i) = r\Big(\sum_{j \neq i} \Omega_j\Big) \leqslant \sum_{j \neq i} r(\Omega_j) ; \tag{101}$$

so that, if condition *(a)* holds, $J^T(\mathbf{I} - \Omega_i) J$ will have at most $(n-r_i)$

non-zero diagonal elements; whence this matrix will have exactly r_i zero diagonal elements (the others are 1), and therefore $\boldsymbol{J}^{\mathrm{T}}\Omega_i\boldsymbol{J}$ will have 1 in the corresponding positions. Consequently, Q_i will be a chi-squared variable with just r_i d.f. [see the argument leading to (95).] Thus, *(a)* implies *(b)* [*(a)* ⇒ *(b)*.] Further, we note that we have shown that

$$\Omega_i^2 = \boldsymbol{J}(\boldsymbol{J}^{\mathrm{T}}\Omega_i\boldsymbol{J})^2\boldsymbol{J}^{\mathrm{T}} = \boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}\Omega_i\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}} = \Omega_i, \tag{102}$$

since the diagonalized matrix has all 0 and 1 elements. Thus, by (99),

$$\mathbf{I} = \Big(\sum_{i=1}^{k}\Omega_i\Big)^2 = \sum_{i=1}^{k}\Omega_i + 2\sum_{i=2}^{k}\sum_{j=1}^{k-1}\Omega_i\Omega_j. \tag{103}$$

Therefore

$$\sum_{i=2}^{k}\sum_{j=1}^{k-1}\Omega_i\Omega_j = \mathbf{0}. \tag{104}$$

Now, the *trace* of a matrix $\boldsymbol{A}$ is defined as

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{u=1}^{n}A_{uu}; \tag{105}$$

so that

$$\mathrm{tr}(\boldsymbol{J}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{J}) = \sum_{t=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{n}J_{ut}A_{uv}J_{vt} = \sum_{u=1}^{n}\sum_{v=1}^{n}\delta_{uv}A_{uv} = \mathrm{tr}(\boldsymbol{A}), \tag{106}$$

since

$$(\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}})_{uv} = \Sigma_{t=1}^{n}J_{ut}J_{vt} = (\mathbf{I})_{uv} = \delta_{uv}; \tag{107}$$

and also

$$\mathrm{tr}(\boldsymbol{A}\,\boldsymbol{B}) = \sum_{u=1}^{n}\sum_{v=1}^{n}A_{uv}B_{vu} = \mathrm{tr}(\boldsymbol{B}\,\boldsymbol{A}); \tag{108}$$

whence, in particular,

$$\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}) = \sum_{u=1}^{n}\sum_{v=1}^{n}A_{uv}^2 \geqslant 0. \tag{109}$$

Therefore, since the matrices Ω_i are *symmetric* (i.e. $\Omega_i^{\mathrm{T}} = \Omega_i$) and *idempotent* (i.e. $\Omega_i^2 = \Omega_i$), it follows that

$$\mathrm{tr}(\Omega_i\Omega_j) = \mathrm{tr}(\Omega_i^2\Omega_j^2) = \mathrm{tr}(\Omega_j\Omega_i^2\Omega_j) = \mathrm{tr}((\Omega_i\Omega_j)^{\mathrm{T}}(\Omega_i\Omega_j)) \geqslant 0, \tag{110}$$

since $(\boldsymbol{A}\,\boldsymbol{B})^{\mathrm{T}} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}$. Consequently, since (104) implies that the trace of

the sum, which equals the sum of the corresponding traces, is zero; we see from (110) that *every* $\mathrm{tr}(\Omega_i \Omega_j) = 0$, and again that, by (109),

$$\Omega_i \Omega_j = \mathbf{0}. \tag{111}$$

When two r.v. are *independent*, their joint c.d.f., probability density, and characteristic function are respectively the products of their individual c.d.f., probability densities, and characteristic functions; and the condition is both necessary and sufficient. The joint characteristic function of two quadratic forms $\zeta^{\mathrm{T}} \boldsymbol{A} \zeta$ and $\zeta^{\mathrm{T}} \boldsymbol{B} \zeta$ is clearly

$$\begin{aligned} \Psi_{\zeta^{\mathrm{T}} \boldsymbol{A} \zeta, \zeta^{\mathrm{T}} \boldsymbol{B} \zeta}(u, v) &= (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{+\infty} dy_1 \cdots \int_{-\infty}^{+\infty} dy_n e^{-\frac{1}{2}\left(\boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} - 2i \boldsymbol{y}^{\mathrm{T}} (u\boldsymbol{A} + v\boldsymbol{B}) \boldsymbol{y}\right)} \\ &= \left\{ \det\left(\mathbf{I} - 2i(u\boldsymbol{A} + v\boldsymbol{B})\right) \right\}^{-\frac{1}{2}}, \end{aligned} \tag{112}$$

where we have used an argument analogous to that leading to (95), and noted that the determinant of any matrix equals the product of its eigenvalues. The individual characteristic functions of the two quadratic forms are clearly obtained from (112) by putting $v = 0$ and $u = 0$, respectively; so that the condition for independence becomes (for all u and v)

$$\det\left(\mathbf{I} - 2i(u\boldsymbol{A} + v\boldsymbol{B})\right) = \det\left((\mathbf{I} - 2iu\boldsymbol{A})(\mathbf{I} - 2iv\boldsymbol{B})\right), \tag{113}$$

since the product of two determinants is the determinant of the product of the corresponding matrices. We conclude that, *if* $\boldsymbol{AB} = 0$, *then* (113) *holds and therefore the two quadratic forms are independent.* Conversely, if the forms are independent, we can show that $\boldsymbol{AB} = \mathbf{0}$ [e.g., see **C6**, Thm. C7, p. 182.] This result is due to A. T. Craig [see *Annals of Mathematical Statistics*, vol. 14 (1943) p. 195; also **C3**, §15.13.] Thus, (111) implies that all the Q_i are independent. Since (as we showed in proving (95)) *(b)* is equivalent to asserting the *idempotence* of the matrices Ω_i, and this is the

property used in proving the independence *(c)* of the Q_i, we see that *(b)* $\Rightarrow$ *(c)*.

Conversely, assuming *(c)*, we have, by Craig's theorem, that (111) holds for every unequal pair i, j; and hence, from (99), for all successive positive powers,

$$\mathbf{I} = \sum_{i=1}^{k} \Omega_i^{s}, \tag{114}$$

whence

$$n = \mathrm{tr}(\mathbf{I}) = \sum_{i=1}^{k} \mathrm{tr}(\Omega_i^{s}). \tag{115}$$

and this can only be [since the trace of a matrix is the sum of its eigenvalues, and the eigenvalues of the s-th power of a matrix are the s-th powers of the eigenvalues of the matrix] if the eigenvalues of each Ω_i are only 0 or 1. Since, for each i, there will be an orthogonal transformation which diagonalizes Ω_i with zeros and ones in the diagonal, it follows that the Ω_i are idempotent, which is equivalent to *(b)*. Thus, *(c)* $\Rightarrow$ *(b)*. Finally, if we assume *(b)*, all the Ω_i are idempotent, with eigenvalues 0 and 1 only. Let $\mathrm{r}(\Omega_i) = r_i$; then $\mathrm{tr}(\Omega_i) = r_i$. Therefore, from the trace of (99),

$$n = \sum_{i=1}^{k} r_i; \tag{116}$$

which establishes that *(b)* $\Rightarrow$ *(a)*, and completes the proof of Cochran's theorem. /// [It is pointed out by Kendall & Stuart in **C3**, (1) p. 361, that, if two chi-squared r.v. are such that their sum is also a chi-squared r.v., with the d.f. also adding-up correctly, they may nevertheless *not be independent*: it is necessary that the r.v. be quadratic forms in standard

normal variates. A simple counter-example is provided by G. S. James (see *Proceedings of the Cambridge Philosophical Society*, vol. 48 (1952) p. 443; or C3, Ex. 7.6, (1) p. 190.)]

A further special case should be mentioned here. It is very common to add a "zeroth" function to the set of g functions $\phi_h(\boldsymbol{x})$ in (66). This is the *constant* function

$$\phi_0(\boldsymbol{x}) = 1. \tag{117}$$

Of course, all the results obtained above remain essentially the same, with the adjunction of a zero index and therefore a zeroth component in the direction indexed 1, 2, ..., g, which becomes indexed by 0, 1, 2, ..., g. We may now consider the *modified null-hypothesis* in which we suppose that $\beta = (\beta_0, 0, 0, \ldots, 0) = \beta_0$, say:

$$S_n(\beta_0) = (\eta - \beta_0\mathbf{1})^T W (\eta - \beta_0\mathbf{1}) = S_n(a_{\min}) + (a_{\min} - \beta_0)^T \Lambda (a_{\min} - \beta_0), \tag{118}$$

where $\mathbf{1}$ denotes a column of ones. Of course, by (88), $E[(a_{\min})_0] = \beta_0$; and, just as before, all three terms in (118) are chi-squared variates with n, $(n-g)$, and g d.f., and (by Cochran's theorem) the last two are independent. The modified null-hypothesis gives us, by (82), that

$$E[\eta] = \beta_0; \tag{119}$$

so that the hypothesis of normal distribution about the regression yields, by (84) and (117), that

$$W^{\frac{1}{2}}\eta = \zeta + W^{\frac{1}{2}}\Phi^T\beta_0 = \zeta + \beta_0 W^{\frac{1}{2}}\mathbf{1}. \tag{120}$$

Thus,
$$\overline{S}_n = (\eta - \frac{\mathbf{1}^T W\eta}{\mathbf{1}^T W\mathbf{1}}\mathbf{1})^T W (\eta - \frac{\mathbf{1}^T W\eta}{\mathbf{1}^T W\mathbf{1}}\mathbf{1}) = \eta^T W\eta - \frac{(\mathbf{1}^T W\eta)^2}{\mathbf{1}^T W\mathbf{1}}$$

$$= (\zeta + \beta_0 W^{\frac{1}{2}}\mathbf{1})^T (\zeta + \beta_0 W^{\frac{1}{2}}\mathbf{1}) - \frac{(\mathbf{1}^T W^{\frac{1}{2}}\zeta + \beta_0 \mathbf{1}^T W\mathbf{1})^2}{\mathbf{1}^T W\mathbf{1}}$$

$$= \zeta^T\zeta - \frac{(\mathbf{1}^T W^{\frac{1}{2}}\zeta)^2}{\mathbf{1}^T W\mathbf{1}}, \tag{121}$$

independently of β_0 [compare (25), noting that $\mathbf{1}^T W \mathbf{1} = \Sigma_{j=1}^{n} W(\xi_j)$.] This sum of squares measures the deviations, not from the actual mean β_0, which is unknown, but from the sample mean, $(\mathbf{1}^T W \eta)/(\mathbf{1}^T W \mathbf{1})$, which is computable. Since (120) holds,

$$W^{\frac{1}{2}}(\eta - \frac{\mathbf{1}^T W \eta}{\mathbf{1}^T W \mathbf{1}}\mathbf{1}) = \zeta + \beta_0 W^{\frac{1}{2}}\mathbf{1} - \frac{\mathbf{1}^T W^{\frac{1}{2}}\zeta + \beta_0 \mathbf{1}^T W \mathbf{1}}{\mathbf{1}^T W \mathbf{1}} W^{\frac{1}{2}}\mathbf{1}$$

$$= \zeta - \frac{W^{\frac{1}{2}}\mathbf{1}\mathbf{1}^T W^{\frac{1}{2}}}{\mathbf{1}^T W \mathbf{1}}\zeta = (\mathbf{I} - \frac{W^{\frac{1}{2}}\mathbf{1}\mathbf{1}^T W^{\frac{1}{2}}}{\mathbf{1}^T W \mathbf{1}})\zeta ; \qquad (122)$$

so that the matrix $\mathbf{I} - (\mathbf{1}^T W \mathbf{1})^{-1} W^{\frac{1}{2}}\mathbf{1}\mathbf{1}^T W^{\frac{1}{2}}$ is idempotent, whence $\overline{S}_n$ is a chi-squared variable, and since the term on the right is an idempotent matrix of rank 1, it follows as before that $\overline{S}_n$ has $(n\text{-}1)$ d.f. [The argument is again essentially that leading to (95) and (96); based on the result that, *if* ζ *is a vector of independent standard normal variates and* Υ *is an idempotent symmetric matrix of rank* r, *then* $\zeta^T \Upsilon \zeta$ *is a chi-squared r.v. with* r *d.f.*; and also that $\mathbf{I} - \Upsilon$ *is idempotent and symmetric, and has rank* $n\text{-}r$.]

In the same spirit as the modified null hypothesis above, we may consider the question, whether *some* of the parameters β_h are zero: let us say, that

$$\beta_h = 0 \quad \text{for} \quad h > e. \qquad (123)$$

Write Φ_1 and Φ_2 for the first e and the next $(g\text{-}e)$ rows of Φ, and β_1 and β_2 for the corresponding parameters; so that our hypothesis becomes

$$\beta_2 = \mathbf{0}. \qquad (124)$$

Then the transformation of parameters from (a_1, a_2) to (ω_1, ω_2) given by

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & -A \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & A \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad (125)$$

yielding the equation [compare (72), (74), and (75)]

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -A^{\mathrm{T}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} W \begin{pmatrix} \Phi_1^{\mathrm{T}} & \Phi_2^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -A \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -A^{\mathrm{T}} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} W\eta, \quad (126)$$

which reduces immediately to

$$\begin{pmatrix} \Phi_1 W \Phi_1^{\mathrm{T}} & \Phi_1 W (\Phi_2^{\mathrm{T}} - \Phi_1^{\mathrm{T}} A) \\ (\Phi_2 - A^{\mathrm{T}} \Phi_1) W \Phi_1^{\mathrm{T}} & (\Phi_2 - A^{\mathrm{T}} \Phi_1) W (\Phi_2^{\mathrm{T}} - \Phi_1^{\mathrm{T}} A) \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} \Phi_1 \\ \Phi_2 - A^{\mathrm{T}} \Phi_1 \end{pmatrix} W\eta. \quad (127)$$

In order to remove the off-diagonal blocks in the matrix on the left of (127), we must choose

$$A = (\Phi_1 W \Phi_1^{\mathrm{T}})^{-1} \Phi_1 W \Phi_2^{\mathrm{T}}. \quad (128)$$

The equations then become

$$\Phi_1 W \Phi_1^{\mathrm{T}} \omega_1 = \Phi_1 W\eta \qquad \text{and} \qquad \Phi_2 W^{\frac{1}{2}} G W^{\frac{1}{2}} \Phi_2^{\mathrm{T}} \omega_2 = \Phi_2 W^{\frac{1}{2}} G W^{\frac{1}{2}} \eta, \quad (129)$$

where

$$G = \mathbf{I} - W^{\frac{1}{2}} \Phi_1^{\mathrm{T}} (\Phi_1 W \Phi_1^{\mathrm{T}})^{-1} \Phi_1 W^{\frac{1}{2}}, \quad (130)$$

as is easily verified.

The question is to compare the sums of squares due to the two regression models, under the hypothesis that the more elaborate model is null; i.e., (123) or (124). By (90), we have that

$$\begin{aligned} S_n(\beta_1) &= S_n(\omega_{\min}^{(e)}) + (\beta_1 - \omega_{\min}^{(e)})^{\mathrm{T}} \Phi_1 W \Phi_1^{\mathrm{T}} (\beta_1 - \omega_{\min}^{(e)}) \\ &= S_n(\omega_{\min}^{(g)}) + (\beta - \omega_{\min}^{(g)})^{\mathrm{T}} \Lambda (\beta - \omega_{\min}^{(g)}); \end{aligned} \quad (131)$$

where $S_n(\beta_1)$ refers to $S_n(\beta)$ when the hypothesis applies: this is the total residual sum of squares from the true regression, and therefore is the same for both sampling models. In the simpler model, in which only e functions appear in the regression formula (66), we write $\omega_{\min}^{(e)}$ for the minimizing set of parameters: by (129), this is simply ω_1. In the more complex model with g functions, we write $\omega_{\min}^{(g)}$ for the full set of parameters, ω_1, ω_2; and Λ denotes the full matrix (72).

The sums of squares due to the regressions are the second terms on the right in each line of (131); so that their difference, the "extra sum of squares" (as Draper & Smith [E1, §2.7] call it) equals

$$\begin{aligned}\Delta_{12} &= (\beta - \omega_{\min}^{(g)})^T \Lambda (\beta - \omega_{\min}^{(g)}) - (\beta_1 - \omega_{\min}^{(e)})^T \Phi_1 W \Phi_1^T (\beta_1 - \omega_{\min}^{(e)}) \\ &= S_n(\omega_{\min}^{(e)}) - S_n(\omega_{\min}^{(g)}) = \zeta^T G \zeta - \zeta^T \Upsilon \zeta \\ &= \zeta^T W^{\frac{1}{2}} (\Phi_2^T - \Phi_1^T A) (\Phi_2 W^{\frac{1}{2}} G W^{\frac{1}{2}} \Phi_2^T)^{-1} (\Phi_2 - A^T \Phi_1) W^{\frac{1}{2}} \zeta , \end{aligned} \tag{132}$$

where Υ denotes the matrix (92) derived from the equations (129), for $\omega_{\min}^{(g)}$ with the full set of parameters ω_1 and ω_2, and the full matrix Λ; and we note that Λ^{-1} may be partitioned into the reciprocals of its two diagonal blocks [the coefficient matrices in the two equations of (129).] It follows from (127) - (130) that the matrix $G - \Upsilon$ in (132) is both symmetric and idempotent. [We have that

$$(\Phi_2 - A^T \Phi_1) W (\Phi_2^T - \Phi_1^T A) = \Phi_2 W^{\frac{1}{2}} G W^{\frac{1}{2}} \Phi_2^T , \tag{133}$$

which establishes the idempotence.] As we have seen, *the rank of an idempotent matrix equals its trace*; so $r(G - \Upsilon) = tr(G - \Upsilon)$, which equals the number of rows in the matrix in (133), which is g-e. It follows that Δ_{12} is distributed as chi-squared with g-e d.f. Since, by (90), (93), (131),

$$\begin{aligned} S_n(\beta_1) = \zeta^T \zeta &= (\beta_1 - \omega_{\min}^{(e)}) \Phi_1 W \Phi_1^T (\beta_1 - \omega_{\min}^{(e)}) + S_n(\omega_{\min}^{(g)}) + \Delta_{12} \\ &= \zeta^T (I - G) \zeta + \zeta^T \Upsilon \zeta + \zeta^T (G - \Upsilon) \zeta , \end{aligned} \tag{134}$$

we may apply Cochran's theorem to prove that Δ_{12} is independent of the other quadratic forms on the right of (134). We may thus apply a variance-ratio test to determine the significance of the added sum of squares due to the use of the more complex model. *Indeed, we may introduce the functions* $\phi_h(x)$ *one by one, testing for significance of improvement at each step.*

Summary of crucial assumptions and choices:

(a) The actual ***regression function*** (60) *is well-approximated by a member of a* ***linear*** *family of the form* (66).

(b) In particular, the functions $\phi_h(\boldsymbol{x})$ *are of suitable* ***form*** *for the regression analysis, in terms of the physical realities of the problem.*

(c) The functions are ***independent,*** in the sense of (77) and (78).

(d) The ***weight function*** $W(\boldsymbol{x})$ *properly attaches importance to different points of the space of independent variables, both in terms of* ***accuracy*** *and* ***cost*** *of errors, and satisfies (e) below, also.*

(e) The ***deviations*** *from the underlying regression function are indeed distributed* ***independently*** *of one-another, about* ***zero mean,*** *with* ***variance*** $1/W(\boldsymbol{x})$, *with a* ***normal distribution.***

None of these assumptions is obviously true, nor are the choices easy to make. In particular, it is predominantly the case that the weight function is chosen to be *constant*. Some justification is to be found in that the residuals presumably represent effects of variables other than the ξ, which may be independent of the ξ. [This may work for assumption *(e)*, but is less obviously valid for assumption *(d)*.] It is the intention of the mathematical derivations given above, to show that, without these assumptions, the conclusions and tests of regression analysis become incorrect. Another point where the assumptions may not be reasonable is in the normality of the residuals: this assumption is justified, when the observations are averages of many readings, or are individually the result of many additive

effects [the Central Limit Theorem ensures it]; but when there are only a few observations, themselves otherwise distributed, the resulting distribution may be far from normal.

One final matter needs to be considered. So far, all that we have said assumes, in effect, that the values $\xi_j = \xi(s_j)$ of the independent variables $\boldsymbol{x}$ are constant: the matrices Φ and $\boldsymbol{W}$ occurring in the analysis are assumed to be constant in our calculations of expected values. This may be interpreted in two ways:

(I) The values ξ_j are not randomly selected, but are either ***all*** *of the values occurring in reality, or a carefully preselected* ***experimental design****, set up to optimize the collection of information.*

(II) Our results are all ***conditional*** *on the observed random values ξ_j and will only give us predictive powers when we take expectations over the distribution of the r.v. ξ: this will necessitate some further assumptions about the distribution of the ξ, and leads to considerable additional complications.* Some analysis is carried out by Kendall & Stuart [C3] and by Draper & Smith [E1]: the net verdict is that the conditional estimators are biased and that the direct application of the standard (type I) analysis may be very misleading... ***Caveat emptor!*** When, disregarding these warnings of bias, we may wish to perform an experiment in which the ξ_j values are known to be some kind of random selection; we must still realize that, in practical situations in which we take either all available data or some sort of "random sample", the distribution of the ξ_j values may be far from representative, and the theory may be far from the type I theory presented in this paper.

ANALYSIS OF VARIANCE[16]

We have already considered, under regression analysis, the partition of a sum of squares into separate sums of squares attributed to distinct statistical and structural effects [see (81), (90), (97), (118), (121), (131), (134), and the general form (99) of Cochran's theorem.] *Analysis of Variance* [AV or "ANOVA"] and its generalization, *Analysis of Covariance*, are based on ideas of R. A. Fisher, and essentially embody the application of Cochran's theorem to analyze the significance of models of the regressional type for statistical situations. The principal kind of problem to which the analysis was applied was that in which the "predictor" variables took on discrete values, most usually, just two, say 0 and 1, denoting the absence or presence of a given factor or treatment. In this most simple case, we call the variables "dummy" or "label" variables. The analysis of such situations under a variety of circumstances is far from easy, as is indicated by Kendall & Stuart's devotion of ten chapters (entitled "Analysis of Variance in the Linear Model: Classified Data", "Other Models for the Analysis of Variance", "The Assumptions of the Analysis of Variance", "The Design of Experiments", "Sample Survey Theory: Designs", "Sample Survey Theory: Supplementary Information", "Multivariate Distribution Theory", "Tests of Hypotheses in Multivariate Analysis", "Canonical Variables", and "Discrimination and Classification", and comprising some 341 tightly-packed pages) to related matters. *It should be noted that many of the results and techniques of AV refer to specialized situations arising in specific experimental designs, and may not without peril be blythely applied to more general regression problems.*

Returning to the more general considerations of AV applied to the regression model discussed here, we find once again that things are by no means simple. In addition to the general considerations *(a)* - *(e)* described on page 35, we must ask such questions as:

(i) *Are there any* ***hidden variables***, *correlated in unknown ways with the known variables* $\boldsymbol{x}$, *and affecting the observed values* η *in a regressive way?* This can seriously distort the results of our analysis.

(ii) Generally, even if it is reasonable to assume that the r.v. η is *linearly* related to the variables $\boldsymbol{x}$, in the sense that the regression function takes the form (66), the dependence will be ***non-linear***, through the functions $\phi_h(\boldsymbol{x})$; and part of our problem will be *to determine a sufficiently large family of functions* $\phi_h(\boldsymbol{x})$ *to include the functional dependence of* η *on* $\boldsymbol{x}$. For even a modest number m of components of $\boldsymbol{x}$, and a relatively limited family of functions (say all products of non-negative integer powers of the components totalling less than q), we find ourselves laden with a large number g of functions and parameters (e.g., $g = \binom{m+q-1}{m}$; so, if $m = 10$ and $q = 4$, $g = \binom{13}{10} = 286$.) *Since the complete set of functions is not practical to use, what functions are to be chosen?* ***In practice, this will depend strongly on physical intuition.***

Several techniques for determining *which of a very large family of functions should be used* are discussed in the literature [see, e.g., the extensive discussion by Draper & Smith, in **E1**, chs. 2, 3, and 6.] There is no single "best" method, and certainly no deductive argument leading to a unique technique.

(iii) Even though we guarantee that the functions chosen are linearly

independent [in the sense of (77) and (78)], they may form a very ill-conditioned matrix Φ, which will lead to great instability in the values of the parameters $a_{\min}$, when these are determined numerically. *Can we choose the functions so that the rows of* $\Phi W^{\frac{1}{2}}$ *are mutually* ***orthogonal*** (the best situation)? Approximate orthogonality between all rows of $\Phi W^{\frac{1}{2}}$ leads to good computational behavior; exact orthogonality means that $\Lambda = (\Phi W^{\frac{1}{2}})(\Phi W^{\frac{1}{2}})^{T}$ is a diagonal matrix, whose reciprocal is trivial to compute.

(iv) *What can we infer from the* ***residuals****, given a partially completed regression model?* Of course, this is really the same question as the initial one: How do we get the "best" regression? But it is to be hoped that, after a first pass at obtaining a model, we are looking at a "small correction" type of situation. [See chapter 3 of **E1**.] It is also possible to get information from residual-plots when what tests-out to be a good model on paper does not stand up to predictive testing.

(v) *In deciding what predictor functions to incorporate in a regression model, and when to* ***quit****, what constitutes* ***significance****?* The usual answer is to look at the "significance level" of a table of test-criteria, such as variance-ratios, and accept anything below 5% (or 1%, or whatever the user considers a risk he or she is willing accept.) Of course, this only refers to Type I error (the null hypothesis is true and we reject it: see page 16): the analysis of Type II errors (related to the *power* of the test) is much more complicated and full elucidation is not available.

However, other considerations also cloud this issue. Draper & Smith refer to a thesis by J. Wetz, which suggests that the variance-ratio must

be at least *four times* the tabulated minimum for the chosen level of significance (of course, this multiple has a degree of arbitrariness) [see **E1**, Appendix 2C, pp. 129-133.] They also point out that, if we choose among, say, m predictor functions to be introduced into a regression formula on the basis of a variance-ratio test applied to each of their "extra sums of squares" as explained on pages 32 - 34 above, and if the tabulated significance-level is, say, γ; then the probability of getting a ratio no smaller than that in the table (with the null-hypothesis in force) in one test is γ. Consequently, the probability of getting a ratio no smaller than that in the table, among m test values, is $[1 - (1 - \gamma)^m]$, which may be quite large: e.g., if $\gamma = 0.01$ and $m = 40$, then the significance-level becomes 33% (if we begin with $\gamma = 0.05$, we arrive at 87%.) Thus, the selection of significant predictor-functions may well be done too hastily. [This formula applies if the possible functions are totally uncorrelated: if not, the significance-level improves, but the difference between the functions decreases!]

(vi) *How are we to test the stability of the parameters over the sample space?* In other words, how do we test whether the parameters obtained by the regression analysis are representative and unbiased for further observations: *what is the* ***predictive ability*** *of the model?* Again, various techniques have been proposed, for dividing the given observations between the set to be used to obtain a model and that to be used to test it. No "best" technique emerges; but it is clear that *the selection* (if not exhaustive) *must be random*, if grave errors are to be avoided.

(vii) *Is there a* ***time-dependence*** *in the observations?* This is a common source of inexplicable non-predictiveness, as well as of general variability, and may be detected by a plot of residuals against time. A similar approach may be used for seeking the effect of other collateral or hidden variables [see also (i) above.]

FOOD FOR THOUGHT

In conclusion, we reproduce a number of reflections, explanations, and warnings, culled from the pages of two of the best references in this important field: Kendall & Stuart [C3] and Draper & Smith [E1].

QUOTES FROM KENDALL & STUART

CHAPTER 17

ESTIMATION

The problem

17.1 On several occasions in previous chapters we have encountered the problem of estimating from a sample the values of the parameters of the parent population. We have hitherto dealt on somewhat intuitive lines with such questions as arose—for example, in the theory of large samples we have taken the means and moments of the sample to be satisfactory estimates of the corresponding means and moments in the parent.

We now proceed to study this branch of the subject in more detail. In the present chapter, we shall examine the sort of criteria which we require a " good " estimate to satisfy, and discuss the question whether there exist " best " estimates in an acceptable sense of the term. In the next few chapters, we shall consider methods of obtaining estimates with the required properties.

17.2 It will be evident that if a sample is not random and nothing precise is known about the nature of the bias operating when it was chosen, very little can be inferred from it about the parent population. Certain conclusions of a trivial kind are sometimes possible—for instance, if we take ten turnips from a pile of 100 and find that they weigh ten pounds altogether, the mean weight of turnips in the pile must be greater than one-tenth of a pound ; but such information is rarely of value, and estimation based on biassed samples remains very much a matter of individual opinion and cannot be reduced to exact and objective terms. We shall therefore confine our attention to random samples only. Our general problem, in its simplest terms, is then to estimate the value of a parameter in the parent from the information given by the sample. In the first instance we consider the case when only one parameter is to be estimated. The case of several parameters will be discussed later.

QUOTES FROM KENDALL & STUART

17.3 Let us in the first place consider what we mean by " estimation. " We know, or assume as a working hypothesis, that the parent population is distributed in a form which is completely determinate but for the value of some parameter θ. We are given a sample of observations $x_1, \ldots, x_n$. We require to determine, with the aid of observations, a number which can be taken to be the value of θ, or a range of numbers which can be taken to include that value.

Now the observations are random variables, and any function of the observations will also be a random variable. A function of the observations alone is called a *statistic*. If we use a statistic to estimate θ, it may on occasion differ considerably from the true value of θ. It appears, therefore, that we cannot expect to find any method of estimation which can be guaranteed to give us a close estimate of θ on every occasion and for every sample. We must content ourselves with formulating a rule which will give good results " in the long run " or " on the average, " or which has " a high probability of success "—phrases which express the fundamental fact that we have to regard our method of estimation as generating a distribution of estimates and to assess its merits according to the properties of this distribution.

17.4 It will clarify our ideas if we draw a distinction between the method or rule of estimation, which we shall call an estimator, and the value to which it gives rise in particular cases, the estimate. The distinction is the same as that between a function $f(x)$, regarded as defined for a range of the variable x, and the particular value which the function assumes, say $f(a)$, for a specified value of x equal to a. Our problem is not to find estimates, but to find estimators. We do not reject an estimator because it gives a bad result in a particular case (in the sense that the estimate differs materially from the true value). We should only reject it if it gave bad results in the long run, that is to say, if the distribution of possible values of the estimator were seriously discrepant with the true value of θ. The merit of the estimator is judged by the distribution of estimates to which it gives rise, i.e. by the properties of its sampling distribution.

CHAPTER 22

TESTS OF HYPOTHESES: SIMPLE HYPOTHESES

22.1 We now pass from the problems of estimating parameters to those of testing hypotheses concerning parameters. Instead of seeking the best (unique or interval) estimator of an unknown parameter, we shall now be concerned with deciding whether some pre-designated value is acceptable in the light of the observations.

In a sense, the testing problem is logically prior to that of estimation. If, for example, we are examining the difference between the means of two normal popula-

tions, our first question is whether the observations indicate that there is *any* true difference between the means. In other words, we have to compare the observed differences between the two samples with what might be expected on the hypothesis that there is no true difference at all, but only random sampling variation. If this hypothesis is not sustained, we proceed to the second step of estimating the *magnitude* of the difference between the population means.

Quite obviously, the problems of testing hypotheses and of estimation are closely related, but it is nevertheless useful to preserve a distinction between them, if only for expository purposes. Many of the ideas expounded in this and the following chapters are due to Neyman and E. S. Pearson, whose remarkable series of papers (1928, 1933a, b, 1936a, b, 1938) is fundamental.

22.2 The kind of hypothesis which we test in statistics is more restricted than the general scientific hypothesis. It is a scientific hypothesis that every particle of matter in the universe attracts every other particle, or that life exists on Mars ; but these are not hypotheses such as arise for testing from the statistical viewpoint. Statistical hypotheses concern the behaviour of observable random variables. More precisely, suppose that we have a set of random variables $x_1, \ldots, x_n$. As before, we may represent them as the co-ordinates of a point ($\mathbf{x}$, say) in the n-dimensional sample space, one of whose axes corresponds to each variable. Since $\mathbf{x}$ is a random variable, it has a probability distribution, and if we select any region, say w, in the sample space W, we may (at least in principle) calculate the probability that the sample point $\mathbf{x}$ falls in w, say $P(\mathbf{x} \in w)$. We shall say that any hypothesis concerning $P(\mathbf{x} \in w)$ is a statistical hypothesis. In other words, any hypothesis concerning the behaviour of observable random variables is a statistical hypothesis.

For example, the hypothesis (a) that a normal distribution has a specified mean and variance is statistical ; so is the hypothesis (b) that it has a given mean but unspecified variance ; so is the hypothesis (c) that a distribution is of normal form, both mean and variance remaining unspecified ; and so, finally, is the hypothesis (d) that two unspecified continuous distributions are identical. Each of these four examples implies certain properties of the sample space. Each of them is therefore translatable into statements concerning the sample space, which may be tested by comparison with observation.

Critical regions and alternative hypotheses

22.5 To test any hypothesis on the basis of a (random) sample of observations, we must divide the sample space (i.e. all possible sets of observations) into two regions. If the observed sample point $\mathbf{x}$ falls into one of these regions, say w, we shall reject the hypothesis ; if $\mathbf{x}$ falls into the complementary region, $W-w$, we shall accept the hypothesis. w is known as the *critical region* of the test, and $W-w$ is called the *acceptance region.*

It is necessary to make it clear at the outset that the rather peremptory terms " reject " and " accept," used of a hypothesis under test in the last paragraph, are now conventional usage, to which we shall adhere, and are not intended to imply that

any hypothesis is ever finally accepted or rejected in science. If the reader cannot overcome his philosophical dislike of these admittedly inapposite expressions, he will perhaps agree to regard them as code words, " reject " standing for " decide that the observations are unfavourable to " and " accept " for the opposite. We are concerned to investigate procedures which make such decisions with calculable probabilities of error, in a sense to be explained.

22.6 Now if we know the probability distribution of the observations under the hypothesis being tested, which we shall call H_0, we can determine w so that, given H_0, the probability of rejecting H_0 (i.e. the probability that $\mathbf{x}$ falls in w) is equal to a pre-assigned value α, i.e.

$$\text{Prob}\{\mathbf{x} \in w \mid H_0\} = \alpha. \tag{22.1}$$

If we are dealing with a discontinuous distribution, it may not be possible to satisfy (22.1) for every α in the interval (0, 1). The value α is called the *size* of the test.(*) For the moment, we shall regard α as determined in some way. We shall discuss the choice of α later.

Evidently, we can in general find many, and often even an infinity, of sub-regions w of the sample space, all obeying (22.1). Which of them should we prefer to the others ? This is the problem of the theory of testing hypotheses. To put it in every-day terms, which sets of observations are we to regard as favouring, and which as disfavouring, a given hypothesis ?

22.7 Once the question is put in this way, we are directed to the heart of the problem. For it is of no use whatever to know merely what properties a critical region will have when H_0 holds. What happens when some other hypothesis holds ? In other words, we cannot say whether a given body of observations favours a given hypothesis unless we know to what alternative(s) this hypothesis is being compared. It is perfectly possible for a sample of observations to be a rather " unlikely " one if the original hypothesis were true ; but it may be much more " unlikely " on another hypothesis. If the situation is such that we are forced to choose one hypothesis or the other, we shall obviously choose the first, notwithstanding the " unlikeliness " of the observations. The problem of testing a hypothesis is essentially one of choice between it and some other or others. It follows immediately that whether or not we accept the original hypothesis depends crucially upon the alternatives against which it is being tested.

The power of a test

22.8 The discussion of **22.7** leads us to the recognition that a critical region (or, synonymously, a test) must be judged by its properties both when the hypothesis tested is true and when it is false. Thus we may say that the errors made in testing a statistical hypothesis are of two types :

(I) We may wrongly reject it, when it is true ;
(II) We may wrongly accept it, when it is false.

These are known as Type I and Type II errors respectively. The probability of a Type I error is equal to the size of the critical region used, α. The probability of a Type II error is, of course, a function of the alternative hypothesis (say, H_1) considered, and is usually denoted by β. Thus

$$\text{Prob}\{\mathbf{x} \in W - w \mid H_1\} = \beta$$

or

$$\text{Prob}\{\mathbf{x} \in w \mid H_1\} = 1 - \beta. \tag{22.2}$$

This complementary probability, $1-\beta$, is called the *power* of the test of the hypothesis H_0 against the alternative hypothesis H_1. The specification of H_1 in the last sentence is essential, since power is a function of H_1.

CHAPTER 26

STATISTICAL RELATIONSHIP: LINEAR REGRESSION AND CORRELATION

26.1 For this and the next three chapters we shall be concerned with one or another aspect of the relationships between two or more variables. We have already, at various points in our exposition, discussed bivariate and multivariate distributions, their moments and cumulants ; in particular, we have discussed the properties of bivariate and multivariate normal distributions. However, a systematic discussion of the relationships between variables was deferred until the theory of estimation and testing hypotheses had been explored. Even in this group of four chapters, we shall not be able to address ourselves to the whole problem, the more complicated distributional problems of three or more variables being deferred until we discuss Multivariate Analysis in Volume 3.

26.2 Even so, the area which we are about to study is a very large one, and it will be helpful if we begin by reviewing it in a general way.

Most of our work stems from an interest in the joint distribution of a pair of random variables : we may describe this as the problem of *statistical relationship*. There is a quite distinct field of interest concerning relationships of a strictly functional kind between variables, such as those of classical physics ; this subject is of statistical interest because the functionally related variables are subject to observational or instrumental errors. We call this the problem of *functional relationship*, and discuss it in Chapter 29 below. Before we reach that chapter, we shall be concerned with the problem of statistical relationship alone, where the variables are not (except in degenerate cases) functionally related, although they may also be subject to observational or instrumental errors ; we regard them simply as members of a distributional complex.

26.3 Within the field of statistical relationship there is a further useful distinction to be made. We may be interested either in the *interdependence* between a number (not necessarily all) of our variables or in the *dependence* of one or more variables upon others. For example, we may be interested in whether there is a relationship between length of arm and length of leg in men ; put this way, it is a problem of interdependence. But if we are interested in using leg-length measurements to convey information about arm-length, we are considering the dependence of the latter upon the former. This is a case in which either interdependence or dependence may be of interest. On the other hand, there are situations when only dependence if of interest. The relationship of crop-yields and rainfall is an example in which non-statistical considerations make it clear that there is an essential asymmetry in the situation : we say, loosely, that rainfall " causes " crop-yield to vary, and we are quite certain that crops do not affect the rainfall, so we measure the dependence of yield upon rainfall.

There is no clear-cut distinction in statistical terminology for the techniques appropriate to these essentially different types of problem. For example, we shall see in Chapter 27 that if we are interested in the interdependence of two variables with the effects of other variables eliminated, we use the method called " partial correlation," while if we are interested in the dependence of a single variable upon a group of others, we use " multiple correlation." Nevertheless, it is true in the main that the study of *interdependence* leads to the theory of correlation dealt with in Chapters 26–27, while the study of *dependence* leads to the theory of regression discussed in these chapters and in Chapter 28.

26.4 Before proceeding to the exposition of the theory of correlation (largely developed around the beginning of this century by Karl Pearson and by Yule), which will occupy most of this chapter, we make one final general point. A statistical relationship, however strong and however suggestive, can never *establish* a causal connexion : our ideas on causation must come from outside statistics, ultimately from some theory or other. Even in the simple example of crop-yield and rainfall discussed in **26.3**, we had no *statistical* reason for dismissing the idea of dependence of rainfall upon crop-yield : the dismissal is based on quite different considerations. Even if rainfall and crop-yields were in perfect functional correspondence, we should not dream of reversing the " obvious " causal connexion. We need not enter into the philosophical implications of this ; for our purposes, we need only reiterate that statistical relationship, of whatever kind, cannot logically imply causation.

G. B. Shaw made this point brilliantly in his Preface to *The Doctor's Dilemma* (1906) : " Even trained statisticians often fail to appreciate the extent to which statistics are vitiated by the unrecorded assumptions of their interpreters . . . It is easy to prove that the wearing of tall hats and the carrying of umbrellas enlarges the chest, prolongs life, and confers comparative immunity from disease. . . . A university degree, a daily bath, the owning of thirty pairs of trousers, a knowledge of Wagner's music, a pew in church, anything, in short, that implies more means and better nurture . . . can be statistically palmed off as a magic-spell conferring all sorts of privileges. . . . The mathematician whose correlations would fill a Newton with

admiration, may, in collecting and accepting data and drawing conclusions from them, fall into quite crude errors by just such popular oversights as I have been describing."

Although Shaw was on this occasion supporting a characteristically doubtful cause, his logic was valid. In the first flush of enthusiasm for correlation techniques, it was easy for early followers of Karl Pearson and Yule to be incautious. It was not until twenty years after Shaw wrote that Yule (1926) frightened statisticians by adducing cases of very high correlations which were obviously not causal : e.g. the annual suicide rate was highly correlated with the membership of the Church of England. Most of these " nonsense " correlations operate through concomitant variation in time, and they had the salutary effect of bringing home to the statistician that causation cannot be deduced from any observed co-variation, however close. Now, more than thirty years later, the reaction has perhaps gone too far : correlation analysis is very unfashionable among statisticians. Yet there are large fields of application (the social sciences and psychology, for example) where patterns of causation are not yet sufficiently well understood for correlation analysis to be replaced by more specifically " structural " statistical methods, and also large areas of multivariate analysis where the computation of what is in effect a matrix of correlation coefficients is a necessary prelude to the detailed statistical analysis ; on both these accounts, some study of the subject is necessary.

The screening of variables in investigatory work

27.27 In new fields of research, a preliminary investigation of the relations between variables often begins with the calculation of the zero-order correlations between all possible pairs of variables, giving the correlation matrix **C.** If we are only interested in " predicting " the value of one variable, x_1, from the others, it is tempting first to calculate only the correlations of x_1 with the others, and to discard those variables with which it has zero or very small correlations : this would perhaps be done as a means of reducing the number of variables to a manageable figure. The next stage would be to calculate the correlation matrix of the retained variables and the multiple correlations of x_1 on combinations of the remaining variables.

Unfortunately, this procedure may be seriously misleading. Whilst it is perfectly true that the whole set of zero-order correlations completely determine the whole complex of partial correlations, it is not true that small zero-order coefficients of x_1 with other variables guarantee small higher-order coefficients, and this is so even if we ignore sampling considerations. Since by (27.62) the multiple correlation must be as great as the largest correlation of any order, we may be throwing away valuable information by the " screening " procedure described above.

Consider (27.5) again :

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\{(1-\rho_{13}^2)(1-\rho_{23}^2)\}^{\frac{1}{2}}}. \qquad (27.67)$$

If ρ_{12} and ρ_{13} or ρ_{23} are zero, so is $\rho_{12.3}$: if $\rho_{13} = \rho_{23} = 0$, $\rho_{12.3} = \rho_{12}$. *But ρ_{12} and ρ_{13} can both be very small while $\rho_{12.3}$ is very large.* In fact, suppose that $\rho_{13} = 0$. Then (27.67) becomes

$$\rho_{12.3} = \rho_{12}/(1-\rho_{23}^2)^{\frac{1}{2}}, \quad \rho_{13} = 0. \tag{27.68}$$

If ρ_{12} is very small and ρ_{23}^2 is very large, (27.68) can be large, too. To consider a specific example, let

$$\left.\begin{aligned} \rho_{13} &= 0, \\ \rho_{12} &= \cos\theta, \\ \rho_{23} &= \cos(\tfrac{1}{2}\pi - \theta) = \sin\theta. \end{aligned}\right\}$$

Then (27.68) becomes

$$\rho_{12.3} = 1.$$

A similar result occurs if we put

$$\rho_{23} = \cos(\tfrac{1}{2}\pi + \theta),$$

for then ρ_{23}^2 is unchanged.

Now we may make $\cos\theta$ (or $\cos(\frac{1}{2}\pi+\theta)$) as small as we like, say ε. Thus we have

$$\left.\begin{aligned} \rho_{13} &= 0, \\ \rho_{12} &= \varepsilon, \\ \rho_{12.3} &= 1. \end{aligned}\right\} \tag{27.69}$$

Since the multiple correlation $\boldsymbol{R}_{1(23)} \geqslant |\rho_{12.3}|$ by (27.62), we have in this case

$$\boldsymbol{R}_{1(23)} = 1. \tag{27.70}$$

By (27.70), x_1 is a perfect linear function of x_2 and x_3, despite the values of the zero-order coefficients in (27.69). We should clearly have been unwise to discard x_2 and x_3 as predictors of x_1 on the evidence of the zero-order correlations alone.

It is easy to see what has happened here in geometrical terms. The vector PQ_1 is orthogonal to PQ_3 and almost orthogonal (making an angle θ near $\frac{1}{2}\pi$) to PQ_2, but all three vectors lie in the same plane, in which PQ_2 and PQ_3 are either at an angle $(\frac{1}{2}\pi - \theta)$ to each other (when $\cos(\frac{1}{2}\pi-\theta)$ is very near 1) or at an angle $(\frac{1}{2}\pi+\theta)$ to each other (when $\cos(\frac{1}{2}\pi+\theta)$ is very near -1).

We have been considering a simple example, but the same argument applies *a fortiori* with more variables, where there is more room for relationships of this kind to appear. The value of $\boldsymbol{R}$ depends on *all* the partial correlations.

Fortunately for human impatience, life has a habit of being less complicated than it need be, and we usually escape the worse possible consequences of simplifying procedures for the selection of " predictor " variables; we usually have enough background knowledge, even in new fields, to help us to avoid the more egregious oversights, but the logical difficulty remains.

CHAPTER 29

FUNCTIONAL AND STRUCTURAL RELATIONSHIP

Functional relations between mathematical variables

29.1 It is common in the natural sciences, and to some extent in the social sciences, to set up a model of a system in which certain mathematical (not random) variables are functionally related. A well-known example is Boyle's law, which states that, at constant temperature, the pressure (P) and the volume (V) of a given quantity of gas are related by the equation

$$PV = \text{constant}. \tag{29.1}$$

(29.1) may not hold near the liquefaction point of the gas, or possibly in other parts of the range of P and V. If we wish to discuss the pressure–volume relationship in the so-called adiabatic expansion, when internal heat does not have time to adjust itself to surrounding conditions, we may have to modify (29.1) to

$$PV^{\gamma} = \text{constant}, \tag{29.2}$$

where γ is an additional constant which may have to be estimated. Moreover, at some stage we may wish to take temperature (T) into account and extend (29.1) to the form

$$PVT^{-1} = \text{constant}.$$

In general, we have a set of variables $X_1, \ldots, X_k$ related in p functional forms

$$f_j(X_1, \ldots, X_k;\ \alpha_1, \ldots, \alpha_l) = 0, \qquad j = 1, 2, \ldots, p, \tag{29.3}$$

depending on l parameters α_r, $r = 1, 2, \ldots, l$. Our object is usually to estimate the α_r from a set of observations, and possibly also to determine the actual functional forms f_j, especially in cases where neither theoretical considerations nor previous experience provide a complete specification of these forms. If we were able to observe values of X without error, there would be no statistical problem here at all: we should simply have a set of values satisfying (29.3) and the problem would be merely the mathematical one of solving the set of equations. However, experimental or observational error usually affects our measurements. What we then observe is not a " true " value X, but X together with some random element. We thus have to estimate the parameters α_r (and possibly the forms f_j) from data which are, to some extent at least, composed of samples from frequency distributions of error. Our problem then immediately becomes statistical.

29.2 In our view, it is particularly important in this subject, which has suffered from confusion in the past, to use a clear terminology and notation. In this chapter, we shall denote mathematical variables by capital Roman letters (actually italic). As usual, we denote parameters by small Greek letters (here we shall particularly use α and β) and random variables generally by a small Roman letter or, in the case of Maximum Likelihood estimators, by the parameter covered by a circumflex, e.g. $\hat{\alpha}$. Error random variables will be symbolized by other small Greek letters, particularly

δ and ε, and the observed random variables corresponding to unobservable variables will be denoted by a " corresponding " (*) Greek letter, e.g., ξ for X. The only possible source of confusion in this system of notation is that Greek letters are performing three roles (parameters, error variables, observable variables) but distinct groups of letters are used throughout, and there is a simple way of expressing our notation which may serve as a rescuer : any Greek letter " corresponding " to a capital Roman letter is the observable random variable emanating from that mathematical variable ; all other Greek letters are unobservables, being either parameters or error variables.

29.3 We begin with the simplest case. Two mathematical variables X and Y are known to be linearly related, so that we have

$$Y = \alpha_0 + \alpha_1 X, \tag{29.4}$$

and we wish to estimate the parameters α_0, α_1. We are not able to observe X and Y ; we observe only the values of two random variables ξ, η defined by

$$\left.\begin{aligned} \xi_i &= X_i + \delta_i, \\ \eta_i &= Y_i + \varepsilon_i, \end{aligned}\right\} \qquad i = 1, 2, \ldots, n. \tag{29.5}$$

The suffixes in (29.5) are important. Observations about any " true " value are distributed in a frequency distribution of an " error " random variable, and the form of this distribution may depend on i. For example, errors may tend to be larger for large values of X than for small X, and this might be expressed by an increase in the variance of the error variable δ.

In this simplest case, however, we suppose the δ_i to be identically distributed, so that δ_i has the same mean (taken to be zero without loss of generality) and variance for all X_i ; and thus also for ε and Y. We also suppose the errors δ, ε to be uncorrelated amongst themselves and with each other. For the present, we do not assume that δ and ε are normally distributed. Our model is thus (29.4) and (29.5) with

$$\left.\begin{aligned} &E(\delta_i) = E(\varepsilon_i) = 0, \quad \text{var } \delta_i = \sigma_\delta^2, \quad \text{var}\,\varepsilon_i = \sigma_\varepsilon^2, \qquad \text{all } i, \\ &\text{cov}\,(\delta_i, \delta_j) = \text{cov}\,(\varepsilon_i, \varepsilon_j) = 0, \qquad i \neq j, \\ &\text{cov}\,(\delta_i, \varepsilon_j) = 0, \qquad \text{all } i, j. \end{aligned}\right\} \tag{29.6}$$

The restrictive assumption on the means of the δ_i is only that they are all equal, and similarly for the ε_i—we may reduce their means μ_δ and μ_ε to zero by absorbing them into α_0, since we clearly could not distinguish α_0 from these biases in any case.

In view of (29.6) we may on occasion unambiguously write the model as

$$\left.\begin{aligned} \xi &= X + \delta, \\ \eta &= Y + \varepsilon. \end{aligned}\right\} \tag{29.7}$$

29.4 At first sight, the estimation of the parameters in (29.4) looks like a problem in regression analysis ; and indeed, this resemblance has given rise to much confusion. In a regression situation, however, we are concerned with the dependence of the mean

(*) It will be seen that the Roman–Greek " correspondence " is not so much strictly alphabetical as aural and visual. In any case, it would be more logical to use the ordinary lower-case Roman letter, i.e. the observed x corresponding to the mathematical variable X, but there is danger of confusion in suffixes, and besides, we need x for another purpose—cf. **29.6.**

value of η (which is Y) upon X, which is not subject to error ; the error variable δ is identically zero in value, so that $\sigma_\delta^2 = 0$. Thus the regression situation is essentially a special case of our present model. In addition (though this is a difference of background, not of formal analysis), the variation of the dependent variable in a regression analysis is not necessarily, or even usually, due to error alone. It may be wholly or partly due to the inherent structure of the relationship between the variables. For example, body weight varies with height in an intrinsic way, quite unconnected with any errors of measurement.

We may easily convince ourselves that the existence of errors in both X and Y poses a problem quite distinct from that of regression. If we substitute for X and Y from (29.7) into (29.4), we obtain

$$\eta = \alpha_0 + \alpha_1 \xi + (\varepsilon - \alpha_1 \delta). \tag{29.8}$$

This is not a simple regression situation : ξ is a random variable, and it is correlated with the error term $(\varepsilon - \alpha_1 \delta)$. For, from (29.6) and (29.7),

$$\begin{aligned} \operatorname{cov}(\xi, \varepsilon - \alpha_1 \delta) &= E\{\xi(\varepsilon - \alpha_1 \delta)\} = E\{(X+\delta)(\varepsilon - \alpha_1 \delta)\} \\ &= -\alpha_1 \sigma_\delta^2, \end{aligned} \tag{29.9}$$

which is only zero if $\sigma_\delta^2 = 0$, which is the regression situation, or in the trivial case $\alpha_1 = 0$.

The equation (29.8) is called a *structural relation* between the observable random variables ξ, η. This structural relation is a result of the *functional relation* between the mathematical variables X, Y.

29.5 In regression analysis, the values of the regressor variable X may be selected arbitrarily, e.g. at equal intervals along its effective range. But they may also emerge as the result of some random selection, i.e. n pairs of observations may be randomly chosen from a bivariate distribution and the regression of one variable upon the other examined. (We have already discussed these alternative regression models in **26.24**, **27.29.**) In our present model also, the values of X might appear as a result of some random process or as a result of deliberate measurement at particular points, but in either case X remains unobserved due to the errors of observation. We now discuss the situation where X, and hence Y, becomes a random variable, so that the functional relation (29.4) itself becomes a *structural relation* between the unobservables.

Structural relations between random variables

29.6 Suppose that X, Y are themselves random variables (in accordance with our conventions we shall therefore now write them as x, y) and that (29.4), (29.5) and (29.6) hold as before. (29.8) will once more follow, but (29.9) will no longer hold without further assumptions, for in it X was treated as a constant. The correct version of (29.9) is now

$$\operatorname{cov}(\xi, \varepsilon - \alpha_1 \delta) = E\{(x+\delta)(\varepsilon - \alpha_1 \delta)\} = E(x\varepsilon) - \alpha_1 E(x\delta) - \alpha_1 \sigma_\delta^2, \tag{29.10}$$

and we now make the further assumptions (two for x and two for y)

$$\operatorname{cov}(x, \delta) = \operatorname{cov}(x, \varepsilon) = \operatorname{cov}(y, \delta) = \operatorname{cov}(y, \varepsilon) = 0. \tag{29.11}$$

(29.11) reduces (29.10) to (29.9) as before.

The present model is therefore

$$\left.\begin{aligned}\xi_i &= x_i+\delta_i,\\ \eta_i &= y_i+\varepsilon_i,\end{aligned}\right\} \tag{29.12}$$

$$y_i = \alpha_0+\alpha_1 x_i, \tag{29.13}$$

subject to (29.6) and (29.11), leading to (29.8) as before. We have replaced the *functional relation* (29.4) between mathematical variables by the *structural relation* (29.13) expressing an exact linear relationship between two unobservable random variables x, y. The present model is a generalization of our previous one, which is simply the case where x_i degenerates to a constant, X_i. The relation (29.8) between the observables ξ, η is a structural one, as before, but we also have a structural relation at the heart of the situation, so to speak.

The applications of structural relation models are principally to the social sciences, especially econometrics. We shall revert to this subject in connexion with multivariate analysis in Volume 3. Here, we may briefly mention by way of illustration that if the quantity sold (y) of a commodity and its price (x) are each regarded as random variables, the hypothesis that they are linearly related is expressed by (29.13). If both price and quantity can only be observed with error, we have (29.12) and are therefore in the structural relation situation. The essential point is that there is both inherent variability in each fundamental quantity with which we are concerned *and* observational error in determining each.

29.7 One consequence of the distinctions we have been making has frequently puzzled scientists. The investigator who is looking for a unique linear relationship between variables cannot accept two different lines, but he was liable in the early days of the subject (and perhaps sometimes even today) to be presented with a pair of regression lines. Our discussion should have made it clear that a regression line does not purport to represent a functional relation between mathematical variables or a structural relation between random variables : it either exhibits a property of a bivariate distribution or, when the regressor variable is not subject to error, gives the relation between the mean of the dependent variable and the value of the regressor variable. The methods of this chapter, which our references will show to have been developed largely within the last twenty years, permit the mathematical model to be more precisely fitted to the needs of the scientific situation.

38.5 The second inadequacy of the classical discussions is even more radical, and is again illustrated by the quotation from J. S. Mill in **38.4**. It arises from the danger of attributing to one or more of the experimental factors, effects upon the dependent variable which are in reality due to variations in some causal factors not included in the experiment. An unrecognized causal factor may (unknown to the experimenter) vary during the course of the experiment in such a way as to favour a particular combination of experimental factors; this combination will then appear to be highly effective, when it is really the unrecognized factor which is producing the good results.

QUOTES FROM KENDALL & STUART

The classical discussions had no solution to this problem, and it is essential to realize how deep-seated and ever-present the problem is. We can *never* be quite sure that all the important, or even the most important, causal factors have been incorporated in the structure of the experiment. Some may be quite unknown; others, although known, may wrongly be considered to be of minor importance and deliberately neglected. We always need to guard against the perversion of the inferences within an experiment by adventitious outside effects.

QUOTES FROM DRAPER & SMITH

CHAPTER 1

FITTING A STRAIGHT LINE BY LEAST SQUARES

1.0. Introduction: The Need for Statistical Analysis

In today's industry, there is no shortage of "information." No matter how small or how straightforward a process may be, measuring instruments abound. They tell us such things as input temperature, concentration of reactant, per cent catalyst, steam temperature, consumption rate, pressure, and so on, depending on the characteristics of the process being studied. Some of these readings are available at regular intervals, every five minutes perhaps or every half hour; others are observed continuously. Still other readings are available with a little extra time and effort. Samples of the end product may be taken at intervals and, after analysis, may provide measurements of such things as purity, per cent yield, glossiness, breaking strength, color, or whatever other properties of the end product are important to the manufacturer or user. In many plants we find huge accumulations of data of these types, and many times the figures are simply collected without any real purpose or reason in mind. Or else there may have been a purpose years before, and although the purpose no longer exists, the figures are still religiously compiled hour by hour, day by day, week by week.

The purpose of this book is not, however, to explain what type of information should or should not be collected for any given process. The purpose is to explain in some detail something of the technique of extracting, from masses of data of the type just mentioned, the main features of the relationships hidden or implied in the tabulated figures. Nevertheless, the study of regression analysis techniques will also provide certain insights into how to plan the collection of data, when the opportunity arises. See, for example, Section 1.8.

In any system in which variable quantities change, it is of interest to examine the effects that some variables exert (or appear to exert) on others. There may in fact be a simple functional relationship between variables; in most physical processes this is the exception rather than the rule. Often there exists a functional relationship which is too complicated to grasp or to describe in simple terms. In this case we may wish to approximate to this functional relationship by some simple mathematical function, such as a polynomial, which contains the appropriate variables and which graduates or approximates to the true function over some limited ranges of the variables involved. By examining such a graduating function we may be able to learn more about the underlying true relationship and to appreciate the separate and joint effects produced by changes in certain important variables.

Even where no sensible physical relationship exists between variables, we may wish to relate them by some sort of mathematical equation. While the equation might be physically meaningless, it may nevertheless be extremely valuable for predicting the values of some variables from knowledge of other variables, perhaps under certain stated restrictions.

In this book we shall use one particular method of obtaining a mathematical relationship. This involves the initial assumption that a certain type of relationship, linear in unknown parameters (except in Chapter 10, where nonlinear models are considered), holds. The unknown parameters are estimated under certain other assumptions with the help of available data, and a fitted equation is obtained. The value of the fitted equation can be gauged, and checks can be made on the underlying assumptions to see if any of these assumptions appears to be erroneous.

CHAPTER 6

SELECTING THE "BEST" REGRESSION EQUATION

6.0. Introduction

We shall defer discussion of the general model building process to Chapter 8, and in this chapter deal only with the use of specific statistical procedures for selecting variables in regression. Suppose we wish to establish a linear regression equation for a particular response Y in terms of the basic "independent" or predictor variables $X_1, X_2, \ldots, X_k$. Suppose further that $Z_1, Z_2, \ldots, Z_r$, all functions of one or more of the X's, represent the complete set of variables from which the equation is to be chosen and that this set includes any functions, such as squares, cross products, logarithms, inverses, and powers thought to be desirable and necessary. Two opposed criteria of selecting a resultant equation are usually involved:

1. To make the equation useful for predictive purposes we should want our model to include as many Z's as possible so that reliable fitted values can be determined.
2. Because of the costs involved in obtaining information on a large number of Z's and subsequently monitoring them, we should like the equation to include as few Z's as possible.

The compromise between these extremes is what is usually called *selecting the best regression equation.* There is no unique statistical procedure for doing this. If we knew the magnitude of σ^2 (the true random variance of the observations) for any single well-defined problem, our choice of a best regression equation would be much easier. Unfortunately, we are never in this position, so a great deal of personal judgment will be a necessary part of any of the methods discussed. In this chapter we shall describe several procedures which have been proposed; all of these appear to be in current use. To add to the confusion they do not all necessarily lead to the same solution when applied to the same problem, although for many problems they will achieve the same answer. We shall discuss: (1) all possible regressions using three criteria; R^2, s^2, and Mallows' C_p, (2) best subset regressions

using R^2, R^2 (adjusted), and C_p, (3) backward elimination, (4) stepwise regression, (5) some variations on previous methods, (6) ridge regression, (7) PRESS, (8) principal components regression, (9) latent root regression, and (10) stagewise regression. After each discussion, we state our personal opinion.

Some Cautionary Remarks on the Use of Unplanned Data

When we do regression calculations on unplanned data (that is, data arising from continuing operations and not from a designed experiment) some potentially dangerous possibilities can arise, as discussed by G. E. P. Box in "Use and abuse of regression," *Technometrics*, **8**, 1966, 625–629. The error in the model may well not be random but may result from the joint effect of several variables not incorporated in the regression equation nor, perhaps, even measured. (He calls these *latent* or *lurking* variables.) Due to the possibilities of bias in the estimates, discussed in Section 2.12, an observed false effect of a visible variable may, in fact, be caused by an unmeasured latent variable. Provided the system continues to run in the same way as when the data were recorded, this will not mislead. However, because the latent variable is not measured, its changes will not be seen or recorded, and such changes may well cause the predicted equation to become unreliable. Another defect in unplanned data is that, often, the most effective predictor variables are kept within quite a small range to keep the response(s) within specification limits. These small ranges will then frequently cause the corresponding regression coefficients to be found "nonsignificant," a conclusion which practical workers will interpret as ridiculous because they "know" the variable is effective. Both viewpoints are, of course, compatible; if an effective predictor variable is not varied much, it will show little or no effect. A third problem with unplanned data is that the operating policy (for example "if X_1 goes high, reduce X_2 to compensate") often causes large correlations between the predictors. This makes it impossible to see if changes in Y are associated with X_1, or X_2, or both. A carefully designed experiment can eliminate all the ambiguities described above. The effects of latent variables can be "randomized out," effective ranges of the predictor variables can be chosen, and correlations between predictors can be avoided. Where designed experiments are not feasible, happenstance data may still be analyzed via regression methods. However, the additional possibilities of jumping to erroneous conclusions must be kept in mind.

CHAPTER 8

MULTIPLE REGRESSION AND MATHEMATICAL MODEL BUILDING

8.0. Introduction

The multiple linear regression techniques we have discussed can be very useful but also very dangerous if improperly used and interpreted. Before tackling a large problem by multiple regression methods it makes sense to preplan the project as far as possible, to specify the objectives of the work, and to provide checkpoints as the work progresses. This planning will be the subject of this chapter. First, however, we shall discuss three main types of mathematical models often used by scientists:

1. The functional model.
2. The control model.
3. The predictive model.

The Functional Model

If the true functional relationship between a response and the predictor variables in a problem is known, then the experimenter is in an excellent position to be able to understand, control, and predict the response. However, there are very few situations in practice in which such models can be determined. Even in those situations, the functional equations are usually very complicated, difficult to interpret and to use, and are usually of nonlinear form. For example, many chemical processes are represented by systems of differential equations which lead to nonlinear models. In complicated cases numerical integration of the equations may be necessary. Examples of nonlinear models were mentioned in Chapter 5 and the fitting of nonlinear models will be discussed in Chapter 10. In such situations the linear regression procedures do not apply or else linear models can be used only as approximations to the correct models in iterative estimation procedures.

The Control Model

Even if it is known completely, the functional model is not always suitable for controlling a response variable. For example, in the problem of the amount of steam used in a plant, one of the most important variables is the ambient temperature, and this is not controllable in the sense that process temperature, process pressure, and other process variables are controllable. An advertising man who wishes to understand the effect of a television commercial on sales is quite aware that his competitor's activities are very important and are a necessary element in any functional model for sales. However, these activities constitute uncontrollable variables no matter how clearly they are specified in the functional model. A model which contains variables under the control of the experimenter is essential for control of a response.

A useful control model can sometimes be constructed by multiple regression techniques, if they are used carefully. If a designed experiment using the controllable variables is feasible, then the effect of these variables on the response can be obtained from a simple application of multiple regression such as those discussed in Chapter 9. However, there are many situations where designed experiments are not feasible: for example, an experiment conducted in a manufacturing plant usually disrupts day-to-day operations, and unless the potential return from a change in the response indicated by this experiment is great enough, the experiment will not be performed; as another example, an experiment conducted in the market place could be well designed and handled, but the uncontrollable factors (each identifiable) would make any calculated mathematical effect of the controlled variable so confusing as to be useless. These situations lead the practitioner to the use of predictive models.

Predictive Models

When the functional model is very complex and when the ability to obtain independent estimates of the effects of the control variables is limited, one can often obtain a linear predictive model which, though it may be in some senses unrealistic, at least reproduces the main features of the behavior of the response under study. These predictive models are very useful and under certain conditions can lead to real insight into the process or problem. It is in the construction of this type of predictive model that multiple regression techniques have their greatest contribution to make. These problems are usually referred to as "problems with messy data"—that is, data in which much intercorrelation exists. The predictive model is not necessarily functional and need not be useful for control purposes. This, of course, does not make it useless, contrary to the opinion of some scientists. If nothing else,

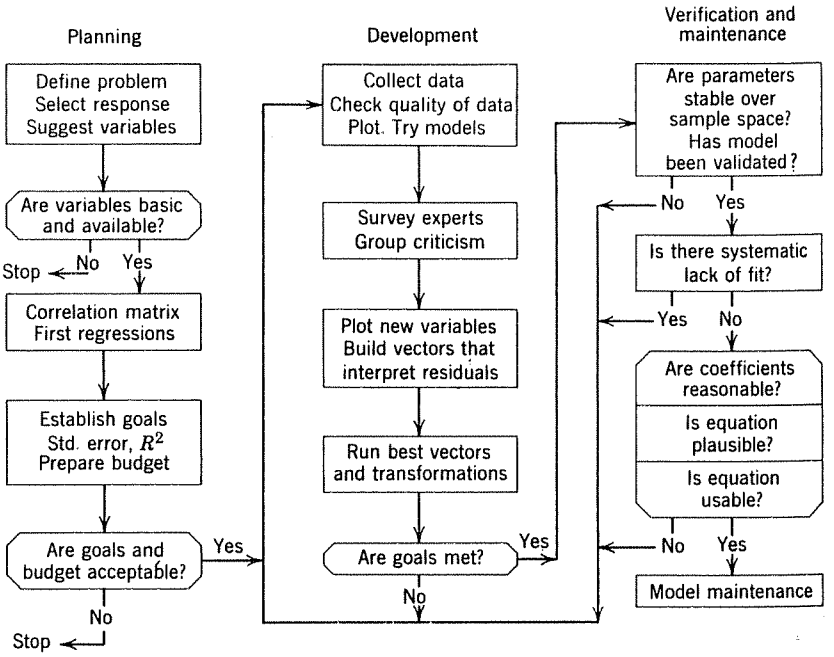

Figure 8.1 Summary of the model building procedure.

it can and does provide guidelines for further experimentation, it pinpoints important variables, and it is a very useful variable screening device.

It is necessary, however, to be very careful in using multiple regression, for it is easily misused and misunderstood. An organized plan for solving problems amenable to a multiple regression approach is both appropriate and necessary. This chapter is intended to be a proposal only, and anyone using this proposed scheme will find it necessary to adjust it to suit his or her particular situation.

While the plan below covers the development of a mathematical model for prediction purposes, it is sufficiently general in scope for use in building both functional and/or control models. The emphasis here will be on the "messy data" type problem. The plan is divided into three stages—planning, development, and maintenance. A schematic diagram of the plan is shown in Figure 8.1 and will be discussed in detail.

The problem definition must be to the point and both the response and predictor variables must be clearly identified. At the beginning of this planning phase, there should be no restraint on the scientist, who should write down every conceivable variable and response that he or she considers to have any possible effect on the problem. This list will be large, but pursuant discussions will gradually

reduce these to a reasonable number. The important point to remember is that *the screening of variables should never be left to the sole discretion of any statistical procedure*, including the multiple regression procedures covered in Chapter 6. Finally one arrives at a specific problem statement with a specified response or responses to be investigated in relation to a specific set of potential predictor variables.

Next, the list of variables obtained from the problem specification discussion must be examined carefully. Many of these predictor variables may turn out to be unmeasurable; for example, the temperature drop in the process could be considered a fundamental variable but it is not measured at the present time. Either a substitute variable which *is* measured and is related to temperature drop would have to be used, or else new instrumentation is needed. This latter alternative will cost money, and the scientist will have to determine which of these two alternatives is better. This scientific, practical assessment of all variables must be done at this point in the planning, before the main body of data is collected.

The next question that needs to be answered is "Can we get a complete set of real observations on all the specified X's and Y's at the same time?" Will the data set be complete? There are many situations in which this is not feasible, and more compromises must be made. A typical situation is one in which samples can be taken all at the same time but the control measurements need to be calculated later or read from a laboratory recording instrument. Based on the current work load, there will be several weeks delay before the data will come back from the laboratory. Should the analysis wait? Should we abandon the idea of collecting those pieces of data? These sorts of questions must be carefully considered before continuing and the time schedule must be *carefully* preplanned. After a complete check on all the variables has been made, a reassessment of the feasibility of the problem solution is essential.

Summary

If the scientist desires to use multiple regression as a tool to help him solve problems, it is imperative that he follow an outline similar to the one illustrated above. Much time and effort can be wasted by trying to make sense of highly intercorrelated data; a series of planned, cost-oriented checkpoints for using multiple regression techniques is a necessity. Finally, no scientist should be persuaded to abandon his scientific insight and principles in favor of some computerized statistical screening procedure. The use of multiple regression techniques is a powerful tool only if it is applied with intelligence and caution.

NOTES

1 The modern theory of probability was founded by A. N. Kolmogorov [see reference **B8**] and is deeply embedded in the theory of measure and integration developed by H. Lebesgue, E. Borel, and others. For a representative selection of texts on measure and Lebesgue integration, see **A1 - A20** ; and for texts presenting probability theory, see **B1 - B13** .

2 Again, for a selection of texts, see references **A1 - A20**.

3 *Product spaces* and *statistical independence* are closely linked. These are discussed in **A1** , ch. 6; **A2** , §5.8; **A6** , ch. 7; **A7** , ch. 6; **A8** , §35; **A15** , ch. 7; **A19** , §14; **B1** , §§2.6, 2.7; **B5** , §§(1)V.4, (2)IV.6; **B7** , ch. 6; **B9** , §(1)1.7; **B11** , ch. 3; and **B12** , §3.4.

4 *Kolmogorov's Strong Law of Large Numbers for independent identically distributed* [i.i.d.] r.v. is discussed, e.g., in **B1** , ch.7 (at pp. 275, 310; **B2** , p. 52; **B5** , chs. (1)8, 10, (2)7; **B6** , ch.6 (at p. 245); **B7** , p. 344; **B12** , §4.4 (at p. 199); and **B13** , §5.3 (at p. 124.)

5 *"With probability one"* (or *"almost surely"* -- abbreviated "a.s.") is the probabilistic equivalent of the measure-theoretic expression "almost everywhere" -- abbreviated "a.e." The definition of one or the other is given in most of the texts listed.

6 *Chebyshev's inequality*: see, e.g., **B5** , (1) p. 233; **B6** , p. 225; **B7** , p. 288; **B12** , p. 85; or **B13** , p. 40.

7 *The Central Limit Theorem* is a fundamental result of probability theory. It is discussed in various forms in **B1** , ch.8; **B2** , chs. 8-11; **B5** , ch. (1) 10 (at p. 244), §(2)VIII.4; **B6** , ch. 8; **B7** , §13.4; **B12**, §4.7; and **B13** , ch. 6 (at p. 205.)

8 *The normal distribution* (in part because of the Central Limit Theorem) is central to probability and statistics. Discussions will be found, e.g., in **B4** , chs. VI, X; **B5** , §(2)III.6; **B7** , §§11.7, 14.6; **C1** , ch. 24; **C3** , (1) [numerous scattered references: see *Index* under "bivariate normal distribution" (p. 420), "multivariate normal distribution" and "normal distribution" (pp. 427, 428); and especially ch. 15]; and **C4** , §§2.2-2.5.

9 *The Multivariate Central Limit Theorem* is discussed in **B4** , ch. 10 (at p. 112); **B5** , (2) p. 260; **C1** , p. 316; and **C3** , (1) p. 194.

10 *Characteristic functions* are discussed, e.g., in **B4**, ch. 4; **B5**, (2) ch. XV; **B7**, ch. 12; **C1**, ch. 10 and §§15.9, 21.3, 22.4; **C3**, (1) ch.4; and in **C5** and **C6**.

11 *The chi-squared distribution* is another major topic, found, e.g., in **B5**, (2) p. 48; **C1**, §18.1; **C2**, pp. 166, 216; **C3**, §(1)16.2-9; **C4**, p. 67 -72; and **C8**, §10.3.

12 The chi-squared distribution of the *sample variance-covariance* estimator is exhibited in **C3**, ch. (1)11 (Ex. 11.3, 11.7); and **C6**, §3.2. See also the discussions on pp. 25, 32 of the present paper.

13 The distribution of the sample variance-covariance estimators of a bivariate normal distribution is derived in **C3**, §(1)16.24 & seq., following the treatment by R. A. Fisher [*Biometrika*, vol. 10 (1915) p. 507.]

14 *The variance-ratio* (or *F*) *distribution* may be found, e.g., in **B5**, (2) p. 48; **C2**, §11.6; **C3**, §(1)16.5-22; **C6**, §3.3; and **C8**, §10.5.

15 *Diagonalization of quadratic forms*: a quadratic form corresponds to a symmetric matrix, and an orthogonal transformation is a change between orthonormal bases. See, e.g., **D1**, ch. 8 (at p. 169; **D2**, ch. 10 (at p. 302), ch. 12 (at p. 362); and **D3**, ch. 10 (at p. 321), ch. 12 (at p. 388.)

16 *Analysis of variance*: see **C1**, chs. 23, 37; **C3**, (1)15.11, chs. (2) 26-30 [note especially the remarks in §§26.2-4, 29.7], chs. (2)19, (3) 35-37 [Kendall & Stuart give by far the most extensive and thorough treatment of all aspects of the subject]; and **C8**, ch. 14. For a specific monograph on this subject, see, e.g., Draper & Smith, **E1**; also **E2-E4**.

REFERENCES

A. Measure Theory

1. S. K. BERBERIAN. *Measure and Integration.* Macmillan, New York, 1965.

2. N. BOURBAKI. *Éléments de Mathématique. Livre VI. Intégration.* Hermann, Paris, published piecemeal.

3. C. CARATHÉODORY. *Algebraic Theory of Measure and Integration.* (Translated by F. E. J. Linton.) Chelsea, New York, 1963.

4. H. FEDERER. *Geometric Measure Theory.* Springer, New York, 1969.

5. A. FRIEDMAN. *Foundations of Modern Analysis.* Holt, Rinehart & Winston, New York, 1970.

6. P. R. HALMOS. *Measure Theory.* Springer, New York, 1974.

7. E. HEWITT, K. STROMBERG. *Real and Abstract Analysis.* Springer, New York, 1965.

8. A. N. KOLMOGOROV, S. V. FOMIN. *Introductory Real Analysis.* (Translated by R. A. Silverman.) Prentice-Hall, Englewood Cliffs, N.J., 1970.

9. J. KOREVAAR. *Mathematical Methods. Volume 1.* Academic Press, New York, 1968.

10. H. LEBESGUE. *Measure and the Integral.* (Translated.) Holden-Day, San Francisco, 1966.

11. M. E. MUNROE. *Measure and Integration.* Addison-Wesley, Reading, Mass., Second Edition, 1971.

12. I. P. NATANSON. *Theory of Functions of a Real Variable.* (Translated by L. F. Boron.) F. Ungar, New York, *Volume 1*, Revised, 1961, *Volume 2*, 1960.

13. J. F. RANDOLPH. *Basic Real and Abstract Analysis.* Academic Press, New York, 1968.

14. H. L. ROYDEN. *Real Analysis.* Macmillan, New York, Second Edition, 1968.

15. W. RUDIN. *Real and Complex Analysis.* McGraw-Hill, New York, 1966.

16. G. E. SHILOV, B. L. GUREVICH. *Integral, Measure, and Derivative: A Unified Approach.* (Translated by R. A. Silverman.) Dover, New York, 1977.

17. E. C. TITCHMARSH. *The Theory of Functions.* Clarendon Press, Oxford, Second Edition, Corrected, 1952.

18. A. J. WEIR. *Lebesgue Integration and Measure.* University Press, Cambridge, 1973.

19. A. J. WEIR. *General Integration and Measure.* University Press, Cambridge, 1974.

20. M. ZAMANSKY. *Linear Algebra and Analysis.* D. Van Nostrand, London, 1969.

B. Probability Theory

1. B. ASH. *Real Analysis and Probability.* Academic Press, New York, 1972.

2. L. BREIMAN. *Probability.* Addison-Wesley, Reading, Mass., 1968.

3. K. L. CHUNG. *A Course in Probability Theory.* Harcourt, Brace & World, New York, 1968.

4. H. CRAMÉR. *Random Variables and Probability Distributions.* University Press, Cambridge, Third Edition, 1970.

5. W. FELLER. *An Introduction to Probability Theory and Its Applications.* J. Wiley, New York, *Volume 1*, Third Edition, 1968, *Volume 2*, Second Edition, 1971.

6. B. V. GNEDENKO. *The Theory of Probability.* (Translated by B. D. Seckler.) Chelsea, New York, Second Edition, 1963.

7. J. F. C. KINGMAN, S. J. TAYLOR. *Measure and Probability.* University Press, Cambridge, 1966.

8. A. N. KOLMOGOROV. *Foundations of the Theory of Probability.* (Translated by N. Morrison.) Chelsea, New York, Second Edition, 1956.

9. M. LOÈVE. *Probability Theory.* Springer, New York, *Volume 1*, Fourth Edition, 1977, *Volume 2*, Fourth Edition, 1978.

10. P. A. P. MORAN. *An Introduction to Probability Theory.* Clarendon Press, Oxford, 1968.

11. J. NEVEU. *Bases Mathématiques du Calcul des Probabilités.* Masson, Paris, 1964.

12. A. RÉNYI. *Foundations of Probability.* Holden-Day, San Francisco, 1970.

13. H. G. TUCKER. *A Graduate Course in Probability.* Academic Press, New York, 1967.

C. Statistics

1. H. CRAMÉR. *Mathematical Methods of Statistics.* University Press, Princeton, 1946.

2. P. G. HOEL. *Introduction to Mathematical Statistics.* J. Wiley, New York, Second Edition, 1954.

3. M. G. KENDALL, A. STUART. *The Advanced Theory of Statistics.* C. Griffin, London, *Volume 1*, Second Edition, 1963, *Volume 2*, 1961, *Volume 3*, 1966.

4. Y. V. LINNIK. *Méthode des Moindres Carrés.* (Translated into French by O. Arkhipoff.) Dunod, Paris, 1963.

5. E. LUKACS. *Characteristic Functions.* C. Griffin, London, 1960.

6. E. LUKACS, R. G. LAHA. *Applications of Characteristic Functions.* C. Griffin, London, 1964.

7. I. MILLER, J. E. FREUND. *Probability and Statistics for Engineers.* Prentice-Hall, Englewood Cliffs, N.J., 1965.

8. A. M. MOOD. *Introduction to the Theory of Statistics.* McGraw-Hill, New York, 1950.

D. Linear Algebra

1. C. C. MacDUFFEE. *Vectors and Matrices.* Mathematical Association of America/Open Court , LaSalle, Illinois, 1943.

2. L. MIRSKY. *An Introduction to Linear Algebra.* Clarendon Press, Oxford, Corrected, 1961.

3. B. NOBLE. *Applied Linear Algebra.* Prentice-Hall, Englewood Cliffs, N.J., 1969.

E. Regression Analysis

1. N. DRAPER, H. SMITH. *Applied Regression Analysis.* J. Wiley, New York, Second Edition, 1981.

2. R. L. PLACKETT. *Regression Analysis.* Clarendon Press, Oxford, 1960.

3. C. R. RAO. *Linear Statistical Inference and Its Applications.* J. Wiley, New York, 1973.

4. E. J. WILLIAMS. *Regression Analysis.* J. Wiley, New York, 1959.