ON BIASES IN ESTIMATION DUE TO THE USE OF PRELIMINARY TESTS OF SIGNIFICANCE

By T. A. Bancroft

Iowa State College

I. INTRODUCTION

In problems of statistical estimation we often express the joint frequency distribution of the sample observations $x_1, x_2, \cdots x_n$ in the form

$$(1) \qquad f(x_1, \cdots, x_n; \alpha, \beta, \gamma, \cdots)\Pi\, dx_i, \qquad (i = 1, \cdots, n)$$

where the functional form, f, is assumed known, and $\alpha, \beta, \gamma, \cdots$ are certain population parameters whose values may or may not be known. Given this specification, statistical theory provides routine mathematical processes for obtaining estimates of the parameters $\alpha, \beta, \gamma, \cdots$ from the observations $x_1, x_2, \cdots, x_n$.

In performing tests of significance we often assume that the data follow some distribution

$$(2) \qquad f_1(x_1, \cdots, x_n; \alpha, \beta, \gamma, \cdots)\Pi\, dx_i, \qquad (i = 1, \cdots, n)$$

where f_1 is a known function or family of functions. We may wish to test the hypothesis that the data follow the more specialized distribution

$$(3) \qquad f_2(x_1, \cdots, x_n; \alpha', \beta', \gamma', \cdots)\Pi\, dx_i, \qquad (i = 1, \cdots, n)$$

where f_2 is some member or sub-group of the family f_1. Given this specification, statistical theory provides routine mathematical processes for testing such hypotheses.

In the application of statistical theory to specific data, there is often some uncertainty about the appropriate specifications in equations (1), (2) and (3). In such cases preliminary tests of significance have been used, in practice, as an aid in choosing a specification. We shall give several examples from the literature of statistical methodology.

(1) In an analysis of variance, in order to obtain a best estimate of variance, we may be uncertain as to whether two mean squares in the lines of the analysis may be assumed homogeneous, [1]. Suppose that it is desired to estimate the variance σ_1^2, of which an unbiased estimate s_1^2 is available. In addition, there is an unbiased estimate s_2^2 of σ_2^2, where from the nature of the data it is known that either $\sigma_2^2 = \sigma_1^2$ or $\sigma_2^2 < \sigma_1^2$. As a criterion in making a decision the following rule of procedure is used frequently: test s_1^2/s_2^2 by the F-test, where s_1^2 and s_2^2 are the two mean squares. If F is not significant at some assigned significance level use $(n_1s_1^2 + n_2s_2^2)/(n_1 + n_2)$ as the estimate of σ_1^2. If F is significant at the assigned significance level, use s_1^2 as the estimate of σ_1^2.

(2) After working out the regression of y on a number of independent variates we may be uncertain as to the appropriateness of the retention of some one of

the independent variates, [2]. To illustrate let us consider the choice between the regression equations $y = b_1x_1 + b_2x_2$ and $y' = b_1'x_1$, after having fitted $y = b_1x_1 + b_2x_2$; the population regression equation being $y = \beta_1x_1 + \beta_2x_2$. In this case a procedure commonly used in deciding whether to retain x_2 is as follows: we test s_2^2/s_3^2 by the F-test, where s_2^2 is the reduction in sum of squares due to x_2 after fitting x_1, and s_3^2 is the residual mean square. If F is not significant at some assigned significance level we omit the term containing x_2 and use b_1' as the estimate of β_1. If F is significant we retain the term containing x_2 and use b_1 as the estimate of β_1. A similar example occurs in fitting a polynomial, when there is uncertainty as to the appropriate degree for the polynomial [3].

(3) In certain analyses we may be uncertain as to the appropriateness of the use of the χ^2 test. Bartlett gives an illustration in a discussion of binomial variation, [4]. He performs two supplementary χ^2 tests of significance as an aid in deciding to abandon the main use of the χ^2 test altogether, and proceeds to use an analysis of variance instead. It is of interest to note that the main use of the χ^2 test gives a significant difference at the 5% level while, in the analysis of variance, Fisher's z is not significant at the 5% level. Here again we might formulate a "rule of procedure" and follow through the analysis as in the preceding cases.

This use of tests of significance as an aid in determining an appropriate specification, and hence the form that the completed analysis shall take, involves acting as if the null hypothesis is false in those cases in which it is refuted at some assigned significance level, and, on the other hand, acting as if the null hypothesis is true in those cases in which we fail to refute it at the assigned significance level. An investigation of the consequences of some of these uses is the purpose of this paper.

It is proposed to consider the first two cases mentioned above: (1) a test of the homogeneity of variances, and (2) a test of a regression coefficient. A complete investigation of the consequences of the rules of procedure would be very extensive, since these consequences depend on the form of the subsequent statistical analysis. As a beginning, it is proposed to limit the study to the efficiency of these "rules of procedure" in the control of bias.

The need for solutions of a whole family of problems of this kind has been pointed out recently by Berkson [5].

II. EXAMPLE ONE: TEST OF HOMOGENEITY OF VARIANCES

1. Statement of the problem. s_1^2 and s_2^2 are two independent estimates of variances σ_1^2 and σ_2^2 respectively, (such that $n_1s_1^2/\sigma_1^2$, $n_2s_2^2/\sigma_2^2$ are distributed independently according to χ_1^2 and χ_2^2, with n_1 and n_2 degrees of freedom). It is known that $\sigma_2^2 \leq \sigma_1^2$. To obtain from these an estimate of σ_1^2, to be used in the particular analysis in hand, we formulate a rule of procedure.

2. Rule of procedure. Test s_1^2/s_2^2 by the F-test. If F is non-significant at some assigned significance level, we use $(n_1s_1^2 + n_2s_2^2)/(n_1 + n_2)$ as the estimate of σ_1^2. If F is significant at some assigned significance level we use s_1^2 as the estimate of σ_1^2. The estimate of σ_1^2 obtained by this rule of procedure will be denoted by e^*.

3. Object of this investigation. If we follow such a rule of procedure, what will be the bias in our estimate e^* of σ_1^2 ?

4. Derivation of the expected value of e^*. First we wish to find

$$E\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\right), \quad \text{if} \quad \frac{s_1^2}{s_2^2} < \lambda,$$

where λ is the value on the F-distribution corresponding to some assigned significance level for n_1 and n_2 degrees of freedom.

Let $v_1 = s_1^2$, $v_2 = s_2^2$. Since s_1^2 and s_2^2 are independently distributed, the joint distribution of v_1 and v_2 is

$$c_1 v_1^{\frac{1}{2}n_1 - 1} v_2^{\frac{1}{2}n_2 - 1} \exp\left[-\frac{1}{2}\left(\frac{n_1 v_1}{\sigma_1^2} + \frac{n_2 v_2}{\sigma_2^2}\right)\right] dv_1\, dv_2 ,$$

where c_1 is a constant and n_1 and n_2 are the respective degrees of freedom.

Let us make the transformation of variables

$$u_1 = n_1 v_1 + n_2 v_2 , \qquad 0 < u_1 < \infty$$

$$u_2 = \frac{v_1}{v_2} \qquad 0 < u_2 < \lambda ,$$

then the expected value, E_1, of $\dfrac{u_1}{n_1 + n_2}$ for $u_2 < \lambda$ is given by

$$(n_1 + n_2)E_1 = \frac{c_1}{P(u_2 < \lambda)} \int_0^\lambda \int_0^\infty \frac{u_2^{\frac{1}{2}n_1 - 1}}{(n_1 u_2 + n_2)^{\frac{1}{2}(n_1+n_2)}}$$

$$\cdot u_1^{\frac{1}{2}(n_1+n_2)} \exp\left[-\frac{1}{2}\,\frac{u_1}{n_1 u_2 + n_2}\left(\frac{n_1 u_2}{\sigma_1^2} + \frac{n_2}{\sigma_2^2}\right)\right] du_1\, du_2$$

where $P(u_2 < \lambda)$ is the probability of u_2 being less than λ.

Integrating out u_1 and expressing the result in terms of the incomplete beta function we obtain

$$(4) \qquad (n_1 + n_2)E_1 = \frac{n_1 I_{x_0}(\frac{1}{2}n_1 + 1, \frac{1}{2}n_2)\sigma_1^2 + n_2 I_{x_0}(\frac{1}{2}n_1, \frac{1}{2}n_2 + 1)\sigma_2^2}{P(u_2 < \lambda)}$$

where $x_0 = (n_1\varphi\lambda)/(n_2 + n_1\varphi\lambda)$, $\varphi = \sigma_2^2/\sigma_1^2$.

We wish now to find the expected value of s_1^2 when $s_1^2/s_2^2 \geq \lambda$. Again we start with the joint distribution of v_1 and v_2, given above and this time let

$$\frac{v_2}{v_1} = Y, \qquad v_1 = v_1 ,$$

then the expected value, E_2, of v_1 when $Y \leqq \frac{1}{\lambda}$ is

$$E_2 = \frac{c_1}{P\left(Y \leqq \frac{1}{\lambda}\right)} \int_0^{1/\lambda} \int_0^{\infty} Y^{\frac{1}{2}n_2 - 1} v_1^{\frac{1}{2}(n_1+n_2)} \exp\left[-\frac{1}{2} v_1 \left(\frac{n_1}{\sigma_1^2} + \frac{n_2 Y}{\sigma_2^2}\right)\right] dv_1 \, dY.$$

Integrating out v_1 as a gamma function, and expressing the results as incomplete beta functions we obtain

$$(5) \qquad E_2 = \frac{[1 - I_{x_0}(\frac{1}{2}n_1 + 1, \frac{1}{2}n_2)]\sigma_1^2}{P(Y \leq 1/\lambda)},$$

where

$$x_0 = n_1\varphi\lambda/(n_2 + n_1\varphi\lambda) \text{ as before.}$$

5. Final Results. The probability that we use $(n_1s_1^2 + n_2s_2^2)/(n_1 + n_2)$ is $P(u_2 < \lambda)$. From equation (4) the contribution from this case to the mean value of e^* is

$$\frac{n_1 I_{x_0}(\frac{1}{2}n_1 + 1, \frac{1}{2}n_2)\sigma_1^2 + n_2 I_{x_0}(\frac{1}{2}n_1, \frac{1}{2}n_2 + 1)\sigma_2^2}{n_1 + n_2}.$$

The probability that we use s_1^2 is $P(Y \leq 1/\lambda)$. From equation (5) the contribution from this case is

$$[1 - I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2)]\sigma_1^2.$$

The expected value of e^* is obtained by combining the two cases, i.e.,

$$(6) \qquad E(e^*) = \left[1 + \frac{n_2}{n_1 + n_2}\left\{I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 1)\frac{\sigma_2^2}{\sigma_1^2} - I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2)\right\}\right]\sigma_1^2.$$

Hence the bias in e^*, expressed as a fraction of σ_1^2 is

$$(7) \qquad \frac{n_1}{n_1 + n_2}\left[I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 1)\frac{\sigma_2^2}{\sigma_1^2} - I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2)\right]$$

We note that in estimating σ_1^2 there will be a positive bias, no bias, or a negative bias according as

$$\frac{I_{x_0}(\frac{1}{2}n_1, \frac{1}{2}n_2 + 1)}{I_{x_0}(\frac{1}{2}n_1 + 1, \frac{1}{2}n_2)}$$

is greater than, equal to, or less than σ_1^2/σ_2^2.

6. Identity and checks. If $\sigma_1^2 = \sigma_2^2$, then in section 4, $E_1 = \sigma_1^2$ and

$$P(u_2 < \lambda) = I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2).$$

From (4) this gives the identity

$$(n_1 + n_2)I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2) = n_1 I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2) + n_2 I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 1)$$

where $x_0 = n_1\lambda/(n_2 + n_1\lambda)$. This identity may be established easily by elementary calculus.

The first result in equation (6) may be checked by noting that when $\lambda = \infty$, i.e. when the two mean squares are always pooled, x_0 is 1 and equation (6) reduces to $(n_1\sigma_1^2 + n_2\sigma_2^2)/(n_1 + n_2)$. Similarly when $\lambda = 0$, in which case there is no pooling, $x_0 = 0$, and equation (6) reduces to σ_1^2.

7. Discussion. In making a choice of an appropriate estimate of σ_1^2 we may consider three procedures:

(1) Use s_1^2 always. This has the merit of having no bias, but is likely to have a large sampling error.

(2) Always pool, i.e., use $\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$. When $\sigma_1^2 \neq \sigma_2^2$ this is biased, but in compensation will have less sampling error than (1) since it will be based on $(n_1 + n_2)$ degrees of freedom.

(3) Use the test of significance of $\dfrac{s_1^2}{s_2^2}$ as a criterion in making the decision as to whether to pool the two mean squares or not. If the test discriminates properly between cases where pooling should or should not be made, the preliminary test of significance criterion will utilize the extra n_2 degrees of freedom whenever permissible and also avoid the bias in method (2).

In Table I the expected value $E(e^*)$ divided by σ_1^2, is given for two sets of values of n_1, n_2 somewhat typical of those frequently encountered in applied work, and for a series of values of σ_2^2/σ_1^2. In addition to the case of always pooling ($\lambda = \infty$) and that of never pooling ($\lambda = 0$), the results for λ at the 1 percent, 5 percent, 20 percent levels and for $\lambda = 1$ have been tabulated. By subtracting unity from the results the bias is obtained as a fraction of σ_1^2. The Table was computed from the incomplete beta function Tables [6].

When the two mean squares are always pooled, the fractional bias is negative and increases numerically as σ_2^2 becomes small relative to σ_1^2. By examination of the values in Table I for $\sigma_2^2/\sigma_1^2 = .1$, it will be seen that the preliminary test of significance controls the bias well when σ_2^2 is much smaller than σ_1^2, that is when a large bias from pooling is most to be feared. This result happens because in such cases the preliminary test allows pooling only in a small proportion of samples.

If λ is taken at the 1 or 5 percent levels, the maximum bias appears to occur when σ_2^2/σ_1^2 is in the region 0.4–0.5, there being little bias when σ_2^2 is near σ_1^2. The lower values of λ (20 percent or λ equals 1) control the bias satisfactorily in the region $\sigma_2^2 < .6\sigma_1^2$, but have a fairly substantial positive bias when $\sigma_2^2 = \sigma_1^2$, that is when pooling would actually be justified. By use of the relation between the incomplete beta function and the sum of the terms of a binomial series it can be shown that there is always a positive bias when $\sigma_1^2 = \sigma_2^2$ and that for given numbers of degrees of freedom this bias is greatest when $\lambda = 1$.

To summarize from the example in Table I, it seems that for small values of n_1 and n_2 none of the values of λ which have been investigated controls the bias throughout the whole range $0 \leq \sigma_2^2/\sigma_1^2 \leq 1$.

TABLE I

Expected Value of e^*/σ_1^2: $E(e^*)/\sigma_1^2$

Case 1: $n_1 = 4, n_2, = 20$

	σ_2^2/σ_1^2									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$\lambda \quad = \infty$	.250	.333	.417	.500	.583	.667	.750	.833	.917	1.00
$\lambda_{.01} = 4.43$	.965	.870	.791	.750	.748	.775	.821	.880	.948	1.02
$\lambda_{.05} = 2.87$	.991	.960	.924	.901	.892	.903	.930	.970	1.02	1.08
$\lambda_{.20} = 0.00$	1.00	.999	1.00	1.01	1.02	1.04	1.07	1.11	1.15	1.20
$\lambda \quad = 1$	1.00	1.00	1.01	1.03	1.05	1.08	1.11	1.15	1.20	1.25
$\lambda \quad = 0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Case 2: $n_1 = 12, n_2 = 10$

	σ_2^2/σ_1^2									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$\lambda \quad = \infty$	.591	.636	.682	.727	.773	.818	.864	.909	.955	1.00
$\lambda_{.01} = 4.71$	.981	.896	.833	.814	.824	.850	.884	.922	.963	1.00
$\lambda_{.05} = 2.91$	.998	.973	.935	.909	.901	.908	.928	.955	.989	1.03
$\lambda_{.20} = 0.00$	1.00	.998	.993	.987	.986	.991	1.00	1.02	1.04	1.07
$\lambda \quad = 1$	1.00	1.00	1.00	1.00	1.01	1.02	1.04	1.06	1.08	1.11
$\lambda \quad = 0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

8. The variance of e^***.** Using the same method we may obtain the variance of e^*. The final result is

$$V = \frac{\begin{array}{r} n_1(n_1 + 2)I_{x_0}(\tfrac{1}{2}n_1 + 2, \tfrac{1}{2}n_2)\sigma_1^4 + 2n_1 n_2 I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2 + 1)\sigma_1^2\sigma_2^2 \\ + n_2(n_2 + 2)I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 2)\sigma_2^4 \end{array}}{(n_1 + n_2)^2}$$

$$\begin{aligned} (8) \qquad &+ \frac{n_1 + 2}{n_1}[1 - I_{x_0}(\tfrac{1}{2}n_1 + 2, \tfrac{1}{2}n_2)]\sigma_1^4 - \Bigg[1 + \frac{n_2}{n_1 + n_2} \\ &\left\{I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 1)\frac{\sigma_2^2}{\sigma_1^2} - I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2)\right\}\Bigg]\sigma_1^4. \end{aligned}$$

From the relations in deriving this result the following identity was obtained:

$$(n_1 + n_2 + 2)(n_1 + n_2)I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2) = n_1(n_1 + 2)I_{x_0}(\tfrac{1}{2}n_1 + 2, \tfrac{1}{2}n_2) + 2n_1n_2I_{x_0}(\tfrac{1}{2}n_1 + 1, \tfrac{1}{2}n_2 + 1) + n_2(n_2 + 2)I_{x_0}(\tfrac{1}{2}n_1, \tfrac{1}{2}n_2 + 2).$$

This identity can be readily established by elementary calculus.

As a check on the result in equation (8), we note that if $\lambda = \infty$, then $x_0 = 1$, and $\frac{s_1^2}{s_2^2} < \lambda$ always. The variance of the estimate of variance becomes $2(n_1\sigma_1^4 + n_2\sigma_2^4)/(n_1 + n_2)^2$, which checks with the variance of $(n_1s_1^2 + n_2s_2^2)/(n_1 + n_2)$ for the case of always pooling. If in addition $\sigma_1^2 = \sigma_2^2$, then $V = 2\sigma_1^4/(n_1 + n_2)$. If $\lambda = 0$, then $x_0 = 0$, and $s_1^2/s_2^2 \geqq \lambda$ always. The variance of the estimate of variance becomes $2\sigma_1^4/n_1$ which checks with the variance of s_1^2 for the case of never pooling.

The expression for the variance of e^* enables us to investigate how much has been gained in terms of reduction in variance by the use of the preliminary test. The quantity $\{V + (\text{Bias})^2\}$ is the appropriate value for the whole sampling error, where V is given by (8) and the bias by (7). For the two numerical examples these quantities are shown as fractions of σ_1^4 in Table II.

As a standard of comparison the variances of the estimate s_1^2 (no pooling) will be used. In these examples the preliminary test with $\lambda = 1$ produces a variance smaller than that of s_1^2 for all values of σ_2^2/σ_1^2 except the lowest (0.1) where the two variances are equal. As λ is taken successively higher there is a substantial reduction in variance when σ_2^2 is near σ_1^2 but an increase in variance over that of s_1^2 when σ_2^2/σ_1^2 is small. Throughout nearly all the range of values of σ_2^2/σ_1^2, the smallest variance is obtained by always pooling ($\lambda = \infty$), despite the relatively large bias given by that method. This result is a reflection of the instability of estimates of variance which are based on only a few degrees of freedom.

III. EXAMPLE TWO: TEST OF A REGRESSION COEFFICIENT

1. Regression and some properties of orthogonal functions. Let

$$y = \beta_1x_1 + \beta_2x_2 + e$$

be a linear regression of y on the two variates x_1 and x_2 in which β_1 and β_2 are the respective population regression coefficients and e is the error. We assume that x_1, x_2 and y are measured from their respective sample means and that the values of x_1 and x_2 are fixed from sample to sample. In order to make comparisons among samples of different sizes we assume that x_1 and x_2 have unit variances and correlation coefficient[1] ρ so that

$$S(x_1^2) = n - 1, \qquad S(x_2^2) = n - 1, \qquad S(x_1x_2) = \rho(n - 1),$$

[1] Although ρ is commonly used to denote a population correlation coefficient, we are using it here for the sample correlation coefficient between the *fixed* variates x_1 and x_2.

TABLE II

The Variance of e^* About its True Mean: $\dfrac{V + (Bias)^2}{\sigma_1^4}$

Case 1: $n_1 = 4$, $n_2 = 20$

	σ_2^2/σ_1^2									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$\lambda \;\;= \infty$	.577	.462	.360	.275	.205	.149	.111	.087	.076	.084
$\lambda_{.01} = 4.43$	.560	.620	.603	.523	.350	.323	.243	.184	.150	.137
$\lambda_{.05} = 2.87$	.514	.545	.554	.528	.479	.414	.353	.299	.260	.237
$\lambda_{.20} = 0.00$	.501	.500	.493	.480	.458	.435	.408	.374	.367	.360
$\lambda \;\;= 1$	.499	.493	.480	.462	.441	.423	.401	.389	.381	.387
$\lambda \;\;= 0$	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500

Case 2: $n_1 = 12$, $n_2 = 10$

	σ_2^2/σ_1^2									
	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$\lambda \;\;= \infty$	.217	.183	.154	.130	.112	.097	.088	.084	.085	.091
$\lambda_{.01} = 4.71$	.185	.218	.203	.171	.141	.118	.103	.094	.092	.096
$\lambda_{.05} = 2.91$	.170	.187	.194	.183	.163	.142	.125	.114	.109	.109
$\lambda_{.20} = 0.00$	.167	.169	.171	.170	.164	.156	.146	.139	.135	.135
$\lambda \;\;= 1$	.167	.166	.165	.162	.158	.152	.148	.144	.144	.147
$\lambda \;\;= 0$	.167	.167	.167	.167	.167	.167	.167	.167	.167	.167

where $S(x_1^2)$ denotes summation of x_1^2 over the sample, with similar meanings for $S(x_2^2)$ and $S(x_1x_2)$, where n is the sample size.

We make the orthogonal transformations

$$\xi_1 = x_1, \qquad \xi_2 = x_2 - \rho x_1,$$

then

$$y = \beta_1\xi_1 + \beta_2(\xi_2 + \rho\xi_1) + e.$$

But

$$S\xi_1^2 = n - 1, \qquad S\xi_2^2 = (n-1)(1-\rho^2), \qquad S(\xi_1\xi_2) = 0,$$

therefore

$$S(y\xi_1) = \beta_1(n-1) + \beta_2\rho(n-1) + S(x_1e),$$

and

$$S(y\xi_2) = \beta_2(n-1)(1-\rho^2) + S(x_2 - \rho x_1)e.$$

Now if we represent the regression coefficients of y on the ξ's as B's we have

$$B_1 S(\xi_1^2) = S(\xi_1 y), \qquad B_2 S(\xi_2^2) = S(\xi_2 y).$$

The reduction in the total sum of squares due to x_1 ignoring x_2 is

$$B_1 S(y\xi_1) = \frac{[S(\xi_1 y)]^2}{S(\xi_1^2)} = \frac{[(n-1)(\beta_1 + \beta_2 \rho) + S(x_1 e)]^2}{n-1}.$$

The reduction in the total sum of squares due to x_2 after fitting x_1 is

$$B_2 S(y\xi_2) = \frac{[S(\xi_2 y)]^2}{S(\xi_2^2)} = \frac{[(n-1)\beta_2(1-\rho^2) + S(x_2 - \rho x_1)e]^2}{(n-1)(1-\rho^2)}.$$

The reduction in the total sum of squares due to regression is

$$\frac{[(n-1)(\beta_1 + \beta_2 \rho) + S(x_1 e)]^2}{n-1} + \frac{[Se(x_2 - \rho x_1) + (n-1)\beta_2(1-\rho^2)]^2}{(n-1)(1-\rho^2)},$$

in which the two parts are independently distributed.

Let b_1' be the regression coefficient of y on x_1 when the term containing x_2 is omitted from the regression equation. Now,

$$b_1' = B_1 = \frac{S(\xi_1 y)}{S(\xi_1^2)} = \frac{(n-1)(\beta_1 + \beta_2 \rho) + S(x_1 e)}{n-1}.$$

Hence

$$E(b_1') = \beta_1 + \beta_2 \rho. \tag{9}$$

Let b_2 be the regression coefficient of y on x_2 if both x_1 and x_2 are used. Then

$$b_2 = B_2 = \frac{S(\xi_2 y)}{S(\xi_2^2)} = \frac{(n-1)\beta_2(1-\rho^2) + S(x_2 - \rho x_1)e}{(n-1)(1-\rho^2)}.$$

And

$$V(b_2) = \frac{S(\xi_2^2)}{[S(\xi_2^2)]^2} = \frac{1}{S(\xi_2^2)} = \frac{1}{(n-1)(1-\rho^2)}.$$

The normal equations for $Y = b_1 x_1 + b_2 x_1$ are

$$b_1 S(x_1^2) + b_2 S(x_1 x_2) = S(x_1 y),$$

$$b_1 S(x_1 x_2) + b_2 S(x_2^2) = S(x_2 y).$$

Now

$$b_1' = \frac{S(x_1 y)}{S(x_1^2)} = b_1 + b_2 \frac{S(x_1 x_2)}{S(x_1^2)}.$$

Therefore

$$b_1' = b_1 + b_2 \rho,$$

or

$$b_1 = b_1' - b_2 \rho.$$

Therefore

$$E(b_1) = \beta_1 + \beta_2\rho - \rho E(b_2). \tag{10}$$

We notice that if $\rho = 0$, b_1 is unbiased in any selected portion of the population.

2. Statement of the problem. To obtain an estimate of b_1, in a particular analysis in hand, in which it is desirable to choose by means of a test of significance between using the regression equation $Y = b_1x_1 + b_2x_2$ and $Y' = b_1'x_1$, we formulate a rule of procedure.

3. Rule of procedure. Calculate the following analysis of variance:

	Degrees of freedom	Sum of squares	Mean square
Reduction due to x_1	1	$\frac{[(n-1)(\beta_1 + \beta_2\rho) + S(x_1e)]^2}{n-1}$	s_1^2
Reduction due to x_2 after fitting x_1	1	$\frac{[(n-1)\beta_2(1-\rho^2) + S(x_2 - px_1)e]^2}{(n-1)(1-\rho^2)}$	s_2^2
Residual	$n - 3$	$S(y - Y)^2$	s_3^2

Test $\frac{s_2^2}{s_3^2}$ by the F-test. If F is non-significant at some assigned significance level we omit the term containing x_2 and use

$$b_1' = \frac{(n-1)(\beta_1 + \beta_2\rho) + S(x_1e)}{n-1}$$

as the estimate of β_1. If F is significant, we retain the term containing x_2, and use b_1 as the estimate of β_1. The estimate obtained by this rule will be called b^*.

4. Object of this investigation. If we follow such a rule of procedure, what will be the bias in b^* as an estimate of β_1?

5. Mathematical derivation of the bias. First, we wish $E(b_1')$ when

$$\frac{s_2^2}{s_3^2} < \lambda \quad \text{or} \quad \frac{b_2^2}{s_3^2} < \frac{\lambda}{(1-\rho^2)(n-1)},$$

where λ is the value on Snedecor's F-distribution corresponding to some assigned significance level for 1 and $(n - 3)$ degrees of freedom. From (9) we have

$$E(b_1') = \beta_1 + \beta_2\rho, \tag{11}$$

no matter what the value of $\frac{s_2^2}{s_3^2}$; since from section 1, s_1^2 and s_2^2 are independently distributed.

Next we wish $E(b_1)$ when $\frac{s_2^2}{s_3^2} \geqq \lambda$ or $\frac{b_2^2}{s_3^2} \geqq \frac{\lambda}{(1-\rho^2)(n-1)}$. To obtain this we find it more convenient to find first $E(b_2)$ when

$$\frac{b_2^2}{s_3^2} \geq \frac{\lambda}{(1-\rho^2)(n-1)}, \quad \text{or} \quad \frac{v_3}{b_2^2} \leq \frac{1}{\lambda c_{22}},$$

where

$$v_3 = s_3^2 \quad \text{and} \quad c_{22} = \frac{1}{(n-1)(1-\rho^2)}.$$

The joint distribution of b_2, v_3 is

$$Ke^{-\frac{1}{2}(b_2-\beta_2)^2/c_{22}} v_3^{\frac{1}{2}(n-5)} e^{-\frac{1}{2}(n-3)v_3}\, dv_3\, db_2,$$

where K is a constant. We make the transformation of variables

$$u = \frac{v_3}{b_2^2}, \qquad dv_3 = b_2^2\, du;$$

then the joint distribution of b_2 and u is

$$Ke^{-(b_2-\beta_2)^2/2c_{22}} (ub_2^2)^{\frac{1}{2}(n-5)} e^{-\frac{1}{2}(n-3)b_2^2u} b_2^2\, du\, db_2.$$

Taking the expected value when $u \leq \frac{1}{\lambda c_{22}}$ we have

$$E(b_2) = \frac{K}{P\left(u \leq \frac{1}{\lambda c_{22}}\right)} \cdot \int_{-\infty}^{\infty}\int_{0}^{1/\lambda c_{22}} b_2 \,|\, b_2 \,|^{n-3} u^{\frac{1}{2}(n-5)} \exp\left[-\frac{(b_2-\beta_2)^2}{2c_{22}} - \frac{(n-3)}{2} b_2^2 u\right] du\, db_2,$$

where $v_2 = s_2^2$, and $P\left(u \leq \frac{1}{\lambda c_{22}}\right)$ is the probability that u be less than or equal to. $\frac{1}{\lambda c_{22}}$.

Dropping subscripts for convenience and expanding the factor which involves e to the first power of b, we have

$$E(b) = \frac{Ke^{-\beta^2/2c}}{P\left(u \leq \frac{1}{\lambda c}\right)} \int_{-\infty}^{\infty}\int_{0}^{1/\lambda c} b \,|\, b \,|^{n-3} u^{\frac{1}{2}(n-5)} \cdot \exp\left\{-\tfrac{1}{2}b^2\left[\frac{1+(n-3)cu}{c}\right]\right\}\left[1 + \left(\frac{b\beta}{c}\right) + \frac{1}{2!}\left(\frac{b\beta}{c}\right)^2 + \cdots\right] du\, db,$$

where

$$-\infty < b < \infty, \qquad 0 < u < \frac{1}{\lambda c}.$$

Now, clearly the even terms of the series vanish whether n is odd or even when b is integrated out.

After integration with respect to b, we have an infinite series in which the typical term (apart from constants) is of the form

$$u^{\frac{1}{2}(n-5)}/[1 + (n - 3)cu]^{\frac{1}{2}(n+r)}$$

where r is an even positive integer. Subsequent integration with respect to u leads to an infinite series of incomplete integrals of the F distribution. By transforming the integrals, the series may be expressed in terms of incomplete beta functions as follows:

$$E(b_2) = \frac{\beta_2 e^{-\beta_2^2/2c_{22}}}{P\left(u \leq \dfrac{1}{\lambda c_{22}}\right)} \cdot \left[I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2}\right) + \left(\frac{\beta_2^2}{2c_{22}}\right) I_{x_0}\left(\frac{n-3}{2}, \frac{5}{2}\right) + \frac{1}{2!}\left(\frac{\beta_2^2}{2c_{22}}\right)^2 I_{x_0}\left(\frac{n-3}{2}, \frac{7}{2}\right) + \cdots \right].$$

Let

$$a = \frac{\beta_2^2}{2c_{22}} \quad \text{or} \quad a = \frac{(1 - \rho^2)(n - 1)\beta_2^2}{2}.$$

Then we have

$$E(b_2) = \frac{\beta_2}{P\left(u \leq \dfrac{1}{\lambda c_{22}}\right)} \sum_{i=0}^{\infty} \frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right), \tag{12}$$

where $x_0 = \dfrac{1}{\dfrac{\lambda}{n-3} + 1}$,

and λ is the desired % point of the F-distribution for 1 and $(n - 3)$ degrees of freedom. Now from (10) we have

$$E(b_1) = \beta_1 + \beta_2\rho - \rho E(b_2)$$

which enables us to obtain $E(b_1)$ from (12).

6. Final result. From (10), (11) and (12) we have

$$E(b^*) = P\left(\frac{v_2}{v_3} < \lambda\right)(\beta_1 + \beta_2\rho) + \left[1 - P\left(\frac{v_2}{v_3} < \lambda\right)\right][\beta_1 + \rho\{\beta_2 - E(b_2)\}]$$

$$= \beta_1 + \rho\beta_2 - \left[1 - P\left(\frac{v_2}{v_3} < \lambda\right)\right]\rho E(b_2).$$

The bias in b^* is

$$\rho\left[\beta_2 - \left\{1 - P\left(\frac{v_2}{v_3} < \lambda\right)\right\}E(b_2)\right].$$

Substituting the value of $E(b_2)$, we obtain

$$\text{Bias} = \rho\beta_2\left[1 - \sum_{i=0}^{\infty}\frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right)\right],$$

where $x_0 = \dfrac{1}{\dfrac{\lambda}{n-3} + 1}$, $a = (1 - \rho^2)\left(\dfrac{n-1}{2}\right)\beta_2^2$.

7. Checks. From section 5 we have

$$E(b_2) = \frac{\beta_2}{P\left(\frac{v_2}{v_3} \geq \lambda\right)}\sum_{i=0}^{\infty}\frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right),$$

where $x_0 = \dfrac{1}{\dfrac{\lambda}{n-3} + 1}$.

If $\lambda = 0$, then $x_0 = 1$, and $E(b_2) = \beta_2$.

Also from section 5 we have

$$\text{Bias} = \rho\beta_2\left[1 - \sum_{i=0}^{\infty}\frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right)\right].$$

If $\lambda = 0$, then $x_0 = 1$, and Bias $= 0$.

If $\lambda = \infty$, then $x_0 = 0$, and Bias $= \rho\beta_2$.

8. Discussion. From the mathematical form of the bias,

$$\text{Bias} = \rho\beta_2\left[1 - \sum_{i=0}^{\infty}\frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right)\right],$$

where $x_0 = \dfrac{1}{\dfrac{\lambda}{n-3} + 1}$,

four deductions follow immediately: (i) There is no bias in estimating β_1, if ρ or β_2 is zero. (ii) The coefficient of β_2 in the formula is less than or at most equal to one in absolute value. (iii) The sign of the bias depends upon the signs of ρ and β_2; it is positive if both are positive or both negative, it is negative if ρ and β_2 have opposite signs. (iv) The bias is estimating β_1 is independent of β_1.

We shall discuss the bias in a few special cases by means of selected values of n, ρ, β_2 and λ. In Table III are exhibited the values of the bias for n equal to 5, 11, 21, each at ρ equal to .2, .4, .6, .8, and β_2 equal to .1, .4, 1.0, 2.0, and 4.0. These values have been computed at the 5% point for λ, and at $\lambda = 1$. These special cases seem to indicate: (i) If we fix ρ, β_2, and λ and increase n, then the bias decreases. (ii) If we fix ρ, β_2, and n and change λ from the 5% point to $\lambda = 1$, the bias decreases considerably. (iii) If we fix ρ, n, λ and increase β_2 the bias increases and then decreases. This may be explained in the following manner. From section 6, the bias may be written in the form

TABLE III

The Bias in Estimating β_1

	$\lambda_{.05} = 18.513$ $n = 5$				$\lambda_{.05} = 5.318$ $n = 11$				$\lambda_{.05} = 4.414$ $n = 21$			
β_2 \ ρ	.2	.4	.6	.8	.2	.4	.6	.8	.2	.4	.6	.8
0.1	.017	.034	.051	.069	.015	.030	.046	.061	.014	.029	.044	.059
0.4	.067	.134	.202	.272	.049	.101	.159	.227	.033	.071	.122	.193
1.0	.142	.292	.455	.640	.028	.072	.164	.350	.001	.006	.025	.132
2.0	.162	.358	.627	1.038	.000	.000	.001	.083	.000	.000	.000	.001
4.0	.035	.101	.282	.898	.000	.000	.000	.000	.000	.000	.000	.000
	$\lambda = 1$ $n = 5$				$\lambda = 1$ $n = 11$				$\lambda = 1$ $n = 21$			
β_2 \ ρ	.2	.4	.6	.8	.2	.4	.6	.8	.2	.4	.6	.8
0.1	.004	.008	.011	.015	.004	.008	.011	.015	.004	.008	.011	.015
0.4	.012	.026	.040	.057	.009	.019	.032	.051	.005	.011	.022	.041
1.0	.011	.025	.049	.095	.001	.003	.010	.039	.000	.000	.001	.009
2.0	.000	.002	.008	.043	.000	.000	.000	.000	.000	.000	.000	.000
4.0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

$$\text{Bias} = \rho\beta_2 \frac{P\left(\frac{v_2}{v_3} < \lambda\right)}{P\left(\frac{v_2}{v_3} > \lambda\right)} \sum_{i=0}^{\infty} \frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right).$$

Now if ρ, n, λ are held constant and β_2 is relatively small, $P\left(\frac{v_2}{v_3} < \lambda\right)$ is relatively large and $\sum_{i=0}^{\infty} \frac{a^i e^{-a}}{i!} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right)$ is relatively large, but $P\left(\frac{v_2}{v_3} > \lambda\right)$ is relatively small. Hence, for a while as we increase β_2 the bias will increase, but as β_2 gets larger $P\left(\frac{v_2}{v_3} < \lambda\right)$ and $\sum_{i=0}^{\infty} a^i e^{-a} I_{x_0}\left(\frac{n-3}{2}, \frac{3}{2} + i\right)$ becomes smaller while $P\left(\frac{v_2}{v_3} > \lambda\right)$ becomes larger. Hence, a value of β_2 will be reached at which the bias begins to decrease. (iv) If we fix n, β_2, and λ and increase ρ, the bias increases without exception.

The above results were obtained under the assumption that a test of significance criterion is used in making a choice as to the number of independent variables to be retained after the regression $y = b_1x_1 + b_2x_2$ has been fitted. If this test of significance criterion is used, we may wish to have a means of controlling the bias. From a study of Table III we note that the bias may be decreased by increasing n and by using $\lambda = 1$. We also notice that as β_2 increases from 0.1 to 4.0 the bias increases and then decreases; and so passes through a maximum value. Hence, if we have a regression in which β_2 is fairly well below or above this maximum value, we would expect a smaller bias.

The bias in estimating β_1 is "unstudentized," i.e., is a function of the population parameters ρ and β_2. In any particular analysis in hand, it would be necessary to know the values of ρ and β_2 or be willing to use estimates of them obtained from the data.

It is realized that only a beginning has been made on the regression problem: an investigation should be undertaken of the more general problem of the use of a test of significance criterion in making a choice as to the number of independent variables to be retained after the regression

$$y = b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

has been fitted.

REFERENCES

[1] J. Wishart and A. R. Clapham, "A study in sampling techniques: the effect of artificial fertilizers on the yield of potatoes," *Jour. Agri. Sci.*, Vol. 19 (1929), Part 4, p. 605.

[2] W. G. Cochran, "The omission or addition of an independent variate in multiple linear regression," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 5 (1938), p. 171.

[3] G. W. Snedecor, *Statistical Methods*, Third edition. Iowa State College Press, Ames, Iowa. 1940 Sec. 14.3.

[4] M. S. Bartlett, "Square root transformation in analysis of variance," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 3 (1936), pp. 76–77.

[5] Joseph Berkson, "Tests of significance considered as evidence," *Jour. Amer. Stat. Assn.*, Vol. 37 (1942), pp. 325–335.

[6] Karl Pearson (Editor), *Tables of the Incomplete Beta Function*, Cambridge Univ. Press. *Biometrika*. Univ. Coll., London, 1934.