Biometrika (1978), **65**, 2, *pp.* 419–27
Printed in Great Britain

Inference based on regression estimator in double sampling

By AJIT C. TAMHANE

Department of Industrial Engineering and Management Sciences,
Northwestern University, Evanston, Illinois

Summary

The problem of hypothesis testing using the regression estimator in double sampling is considered. Test procedures are provided when the covariance matrix between the primary and the auxiliary variables is either partially known or completely unknown. For the latter case a new 'studentized' version of the regression estimator is proposed as a test statistic. The exact null distribution of this statistic is derived in a special case. An approximation to the null distribution is derived in the general case and studied by means of the Monte Carlo method. The problem of choosing between the double sample regression estimator and the single sample mean estimator is also discussed.

Some key words: Bivariate normal distribution; Double sampling; Exact and approximate null distributions; Hypothesis test; Missing observations; Regression estimator.

1. Introduction

In practice it is often the case that the characteristic of principal interest to the investigator is very expensive to measure. However, another characteristic can be identified which is highly correlated with the first one and is relatively inexpensive to measure. These characteristics will be referred to as the primary, Y, and the auxiliary, X, variables respectively. For estimating the mean of Y a single sampling plan which takes observations only on Y may not yield an estimator with desired precision since it may not be feasible to take a sufficiently large number of observations because of budget constraints. Then the precision of the estimator can often be improved by adopting a double sampling plan which takes observations also on the auxiliary variable. In this plan, n_1 observations are taken on both X and Y in the first phase; n_2 additional observations are taken on X alone in the second phase. An estimator of the mean of Y, commonly referred to as the double sample regression estimator or simply as the regression estimator, is then constructed using all the observations. This estimator was first proposed by Bose (1943) and is widely used in sample surveys.

The regression estimator also arises in inference problems regarding means of multivariate populations with missing observations; some recent references on this topic are Rubin (1976) and Little (1976). Some developments in the present paper parallel those considered by Lin (1973) for testing the difference between the means of two correlated variables when some observations on one variable are missing. The calibration problem considered by Scheffé (1973) and Williams (1969) is also related to the problem under study.

In the present paper our main interest centres on the problem of testing hypotheses concerning the mean of Y using its regression estimator. We assume that X and Y are jointly normally distributed. First we state some basic results regarding the regression estimator and compare its performance with the ordinary sample mean of Y obtained by using a single sampling plan, when both the sampling plans are subject to the same budget constraint. Next we provide test procedures based on the regression estimator for situations where some partial knowledge about the covariance matrix between X and Y is available.

Finally, we consider the case where the covariance matrix is completely unknown. This same problem has recently received attention in the papers by Khatri, Bhargava & Shah (1974) and Little (1976). Whereas Khatri *et al.* derived the exact distribution of a certain 'studentized' version of the regression estimator, Little derived a type of t approximation to the distribution of another studentized version of it. The exact distribution derived by Khatri *et al.* is very complicated and depends on ρ, the correlation coefficient between X and Y, which is a nuisance parameter in the present problem. Therefore their results are not particularly useful from a practical viewpoint. We propose a new studentization and derive its exact null distribution in a special case which does not depend on ρ. Next we derive an approximation to the null distribution in the general case which is easy to apply in practice. This approximation also involves a t distribution with nonintegral number of degrees of freedom. We study the empirical size and power of the test based on this approximation by means of the Monte Carlo method and compare it with some competing test statistics, including Little's statistic.

2. PRELIMINARIES

2·1. *Model and some basic results*

We assume that (X, Y) follows a bivariate normal distribution with mean vector (μ_1, μ_2) and covariance matrix

$$\Omega = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Suppose that we have a random sample (X_i, Y_i) $(1 \leqslant i \leqslant n_1)$ from this distribution, and in addition that we have n_2 independent observations X_i $(n_1+1 \leqslant i \leqslant n_1+n_2)$. Let $n = n_1+n_2$ and $\theta = n_2/n$. It can be shown that if Ω is known then the minimum variance unbiased estimator of μ_2 is given by $\hat{\mu}_2 = \bar{Y}_1 + \beta\theta(\bar{X}_2 - \bar{X}_1)$, where $\beta = \rho\sigma_2/\sigma_1$ and

$$\bar{Y}_1 = \sum_{i=1}^{n_1} Y_i/n_1, \quad \bar{X}_1 = \sum_{i=1}^{n_1} X_i/n_1, \quad \bar{X}_2 = \sum_{i=n_1+1}^{n} X_i/n_2.$$

The estimator $\hat{\mu}_2$ is referred to as the regression estimator. It is unbiased and its variance equals $\sigma_2^2(1-\rho^2\theta)/n_1$.

If Ω is not known then it can be shown (Anderson, 1957) that the maximum likelihood estimator of μ_2 is $\hat{\mu}_2 = \bar{Y}_1 + \hat{\beta}\theta(\bar{X}_2 - \bar{X}_1)$, where $\hat{\beta} = \Sigma(X_i - \bar{X}_1)(Y_i - \bar{Y}_1)/\Sigma(X_i - \bar{X}_1)^2$ is the usual estimate of the regression coefficient from the first n_1 observations. This latter estimator of the mean is commonly referred to as the regression estimator. It is unbiased and its variance equals

$$\sigma_2^2[1 + \{\theta/(n_1-3)\}\{1-(n_1-2)\rho^2\}]n_1^{-1} \tag{2·1}$$

for $n_1 > 3$ (Morrison, 1971).

2·2. *When to use the regression estimator*

A natural competitor to $\hat{\mu}_2$ is $\bar{Y}_1$. This sample mean estimator of μ_2 does not use data on the auxiliary variable. If Ω is known, we have that $\text{var}(\hat{\mu}_2) \leqslant \text{var}(\bar{Y}_1)$. If Ω is unknown we have that $\text{var}(\hat{\mu}_2) \leqslant \text{var}(\bar{Y}_1)$ whenever $\rho^2 \geqslant 1/(n_1-2)$.

The above comparison of $\hat{\mu}_2$ and $\bar{Y}_1$ would be useful when the double sample is already available, either through a plan or through an accident, that is missing observations, and the choice is to be made between the two estimators. However, if the allocation of the observations is within the control of the experimenter and he wishes to plan his experiment to estimate μ_2 then the following comparison would perhaps be more appropriate; see also Matthai (1951).

Suppose that the sampling cost function is linear, with no set-up costs, and that c_1 and c_2 denote the costs per observation on X and Y respectively with $c_2 \geqslant c_1 > 0$. Further assume that the total sampling cost is not to exceed some preassigned positive constant K. Consider the double sampling plan for which n_1 and n_2 must satisfy $(c_1+c_2)n_1+c_2n_2 \leqslant K$. Henceforth in this section we consider n_1 and n_2 to be nonnegative continuous variables. It is easy to verify that, if Ω is known, then it is worth increasing n_2 above zero only if $\rho^2 > c_1/(c_1+c_2)$. When this inequality holds, we should choose n_1 and n_2 so that

$$n_1/(n_1+n_2) = \{c_1(1-\rho^2)/c_2\rho^2\}^{1/2} \qquad (2\cdot2)$$

subject to the constraint that $(c_1+c_2)n_1+c_2n_2 = K$; the corresponding minimum value of $\operatorname{var}(\hat{\mu}_2) = \sigma_2^2[(c_1\rho^2)^{1/2}+\{c_2(1-\rho^2)\}^{1/2}]^2/K$. On the other hand, if a single sampling plan is followed and all the observations are made on Y, then the variance of the resulting sample mean estimator $\bar{Y}$ is $\sigma_2^2c_2/K$. It is easy to verify that $\operatorname{var}(\hat{\mu}_2) \leqslant \operatorname{var}(\bar{Y})$ whenever $\rho^2 \geqslant 4c_1c_2/(c_1+c_2)^2 > c_1/(c_1+c_2)$. Thus consideration of when to use the regression estimator depends crucially on $|\rho|$ and the cost ratio c_1/c_2.

If Ω is not known then we cannot make the optimum allocation as in (2·2); however, if a prior estimate of ρ is available then it should be used in making an approximately optimum allocation.

3. Test procedures for partially known Ω

3·1. *Preliminaries*

Suppose that it is desired to construct a test using the regression estimator for the two-sided hypothesis testing problem H_0: $\mu_2 = 0$ against H_1: $\mu_2 \neq 0$. We shall consider two special cases; in each we shall assume different but specified levels of partial knowledge about Ω. Clearly for completely known Ω, one uses $\hat{\mu}_2 n_1^{1/2}\sigma_2^{-1}(1-\rho^2\theta)^{-1/2}$ as the test statistic which is distributed as $N(0,1)$ under H_0 and which gives a uniformly most powerful test among all unbiased tests.

3·2. *Ω known up to a multiple*

The situation where Ω is known up to a multiple is equivalent to β and ρ being known and $\sigma_2/\sigma_1 = \beta/\rho$. An unbiased estimate of σ_2^2 with $2n_1+n_2-2$ degrees of freedom can then be obtained as

$$\hat{\sigma}_2^2 = \left\{\sum_{i=1}^{n_1}(Y_i-\bar{Y}_1)^2+(\beta/\rho)^2\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}(2n_1+n_2-2)^{-1},$$

where $\bar{X} = (1-\theta)\bar{X}_1+\theta\bar{X}_2$ is the cumulative sample mean of the X's. The test is based on the statistic

$$T_1 = \{\bar{Y}_1+\beta\theta(\bar{X}_2-\bar{X}_1)\}n_1^{1/2}\{\hat{\sigma}_2(1-\rho^2\theta)^{1/2}\}^{-1},$$

which is distributed as a $t_{2n_1+n_2-2}$ variable under H_0. A two-sided test based on this statistic is uniformly most powerful among all unbiased tests.

3·3. *Only β known*

Define $U_i = Y_i-\beta\theta X_i$ $(1 \leqslant i \leqslant n_1)$ and $V_i = -\beta\theta X_{n_1+i}$ $(1 \leqslant i \leqslant n_2)$ and note that the random variables U_i and V_i are all mutually independent. Further the U_i are $N(\xi_1, \eta_1^2)$ with $\xi_1 = \mu_2-\beta\theta\mu_1$ and $\eta_1^2 = \sigma_2^2(1+\theta^2\rho^2-2\theta\rho^2)$ and the V_i are $N(\xi_2, \eta_2^2)$ with $\xi_2 = -\beta\theta\mu_1$ and $\eta_2^2 = \sigma_2^2\theta^2\rho^2$. Therefore the given testing problem is equivalent to testing H_0: $\xi_1 = \xi_2$ against H_1: $\xi_1 \neq \xi_2$. Since the variances η_1^2 and η_2^2 are unknown and unequal and the samples are independent, this corresponds to the Behrens–Fisher problem. A large number of approximate solutions to this problem are available (Lee & Gurland, 1975), any one of which can be applied in the present context. Here we shall present a modification of Banerjee's (1961) method; this modification is based on the fact that η_1^2 and η_2^2 are not completely unrelated parameters. Naik (1975) has considered a similar modification for Lin's (1973) testing problem.

Define

$$\bar{U} = \sum_{i=1}^{n_1} U_i/n_1, \quad \bar{V} = \sum_{i=1}^{n_2} V_i/n_2, \quad S_1^2 = \sum_{i=1}^{n_1} (U_i-\bar{U})^2/(n_1-1), \quad S_2^2 = \sum_{i=1}^{n_2} (V_i-\bar{V})^2/(n_2-1).$$

The test statistic used is

$$T_2 = \{\bar{Y}_1 + \beta\theta(\bar{X}_2 - \bar{X}_1)\}\,(a_1 S_1^2/n_1 + a_2 S_2^2/n_2)^{-1/2}, \tag{3·1}$$

and the critical region is $T_2^2 > 1$. In (3·1), a_1 and a_2 are constants to be determined so that the size of the critical region is $\leqslant \alpha$ $(0 < \alpha < 1)$, the specified level of significance.

Let us define $\lambda = (\eta_1^2/n_1)/\{(\eta_1^2/n_1) + (\eta_2^2/n_2)\}$. Then we note that under H_0

$$T_2^2 \sim \frac{\chi_1^2}{\lambda a_1 \chi^2_{n_1-1}/(n_1-1) + (1-\lambda)\, a_2 \chi^2_{n_2-1}/(n_2-1)}, \tag{3·2}$$

where χ_ν^2 denotes a central chi-squared variable with ν degrees of freedom. Note that in (3·2) all the χ^2 variables are independent. By using the concavity property of the distribution function of a χ_1^2 variable it can be seen that $\mathrm{pr}\,(T_2^2 > 1 \mid H_0) \leqslant \alpha$ for all λ, $\lambda_{\min} \leqslant \lambda \leqslant \lambda_{\max}$, if this probability is made equal to α at $\lambda = \lambda_{\min}$ and $\lambda = \lambda_{\max}$. Banerjee's solution is obtained by taking $\lambda_{\min} = 0$ and $\lambda_{\max} = 1$, which corresponds to letting $\eta_1^2/\eta_2^2 \to 0$ and $\eta_1^2/\eta_2^2 \to \infty$ respectively. This yields $a_i = F_{1,n_i-1,\alpha}$ for $i = 1, 2$, where $F_{\nu_1,\nu_2,\alpha}$ denotes the upper α point of an F variable with ν_1 and ν_2 degrees of freedom.

In our problem $\eta_1^2/\eta_2^2 = (1+\theta^2\rho^2 - 2\theta\rho^2)/(\theta^2\rho^2)$ ranges between $(1-\theta)^2/\theta^2$ and ∞ as ρ^2 varies from 1 to 0. Therefore the corresponding values of $\lambda_{\min}$ and $\lambda_{\max}$ are $(1-\theta)$ and 1 respectively. We should set the values of a_i so as to make the size of the test equal to α at these extreme values of λ. The resulting new values of a_i would be no bigger than the previous ones and they would lead to a more powerful test statistic. These values are $a_1 = F_{1,n_1-1,\alpha}$, and a_2 is the solution to the equation

$$\mathrm{pr}\,\{\chi_1^2 > (1-\theta)(F_{1,n_1-1,\alpha})\,\chi^2_{n_1-1}/(n_1-1) + \theta a_2 \chi^2_{n_2-1}/(n_2-1)\} = \alpha. \tag{3·3}$$

The solutions to (3·3) for selected values of n_1, n_2 and α have been tabulated by Naik (1975, Table 1); note that our a_2, n_1 and n_2 correspond to Naik's l_2, n and n^* respectively. Thus they can be used to perform the test. Extensions to one-sided hypothesis problems are quite straightforward.

4. TEST PROCEDURES FOR UNKNOWN Ω

4·1. *Test statistic*

In this case, we propose the following studentized regression estimator as the test statistic:

$$T_3 = \frac{\bar{Y}_1 + \hat{\beta}\theta(\bar{X}_2 - \bar{X}_1)}{[\theta\Sigma\,(Y_i - \hat{Y}_i)^2/\{n_1(n_1-3)\} + \Sigma\,(Y_i - \bar{Y}_1)^2/\{n(n_1-1)\}]^{1/2}}, \tag{4·1}$$

where the summations are over $i = 1, \ldots, n_1$ and $\hat{Y}_i = \bar{Y}_1 + \hat{\beta}(X_i - \bar{X}_1)$ for $1 \leqslant i \leqslant n_1$. In (4·1) the squared denominator is an unbiased estimate of $\mathrm{var}\,(\hat{\mu}_2)$ given by (2·1) (Tikkiwal, 1960).

Little's (1976) test statistic differs only slightly from ours; he uses

$$T_4 = \frac{\bar{Y}_1 + \hat{\beta}\theta(\bar{X}_2 - \bar{X}_1)}{[\theta(n_1-2)\Sigma(Y_i - \hat{Y}_i)^2/\{n_1^2(n_1-3)\} + \Sigma(Y_i - \bar{Y}_1)^2/(nn_1)]^{1/2}}. \tag{4·2}$$

Therefore the asymptotic distributions, as $n_1 \to \infty$, $n_2/n = \theta \to \theta^* \in [0, 1]$, of both the test statistics are the same, namely $N\{\mu_2 n_1^{1/2} \sigma_2^{-1}(1-\theta^*\rho^2)^{-1/2}, 1\}$. Large sample tests can be constructed using this fact. However, here we are mainly concerned with small sample theory.

It can be checked that in the case of small samples, the distribution of T_3 is free of all nuisance parameters except ρ. The exact distribution of T_3 is extremely hard to derive in the general case. We derive the exact null distribution of T_3 in a special case in § 4·2 and a t type of approximation in the general case in § 4·3.

4·2. *Exact null distribution in a special case*

The main difficulty in deriving the exact distribution of T_3 arises because its numerator and denominator are correlated. However, in one special case of some practical interest this difficulty disappears and the distribution can be easily derived. This is the situation where a double sample is planned and $c_2 \gg c_1$ and therefore $n_2 \gg n_1$; see (2·2). This leads us to the following theorem.

THEOREM 4.1. *Suppose that for fixed* n_1 $(3 < n_1 < \infty)$, $n_2 \to \infty$ *and therefore* $\theta \to 1$; *then*

$$\mathrm{pr}(T_3^2 \leqslant a \mid H_0) \to \int_0^\infty G_{1,n_1-2}\left[\frac{(n_1-2)a}{(n_1-3)\{1+z/(n_1-1)\}}\right] dG_{1,n_1-1}(z), \tag{4·3}$$

where $G_{\nu_1,\nu_2}(.)$ *denotes the distribution function of an* F_{ν_1,ν_2} *variable.*

Proof. First we note that under the specified limiting conditions T_3 converges stochastically to the random variable

$$T = \{\bar{Y}_1 - \hat{\beta}(\bar{X}_1 - \mu_1)\}\{n_1(n_1-3)\}^{1/2}\left\{\sum_{i=1}^{n_1}(Y_i - \hat{Y}_i)^2\right\}^{-1/2}.$$

We shall henceforth restrict attention to the null distribution of T. We shall first condition on the X_i $(1 \leqslant i \leqslant n_1)$ and therefore on the sufficient statistics $\bar{X}_1$ and $S^2 = \Sigma(X_i - \bar{X}_1)^2$. Then using results of Anderson (1958, p. 64), we have that, say,

$$\begin{aligned} \bar{Y}_1/\{\sigma_2(1-\rho^2)^{1/2}\} &= U \sim N\{\beta(\bar{X}_1 - \mu_1)\sigma_2^{-1}(1-\rho^2)^{-1/2}, 1\}, \\ \hat{\beta}S/\{\sigma_2(1-\rho^2)^{1/2}\} &= V \sim N\{\beta S\sigma_2^{-1}(1-\rho^2)^{-1/2}, 1\}, \\ \sum_{i=1}^{n_1}(Y_i - \hat{Y}_i)^2/\{\sigma_2^2(1-\rho^2)\} &= W \sim \chi^2_{n_1-2}, \end{aligned} \tag{4·4}$$

and U, V and W are mutually independent. Therefore

$$\frac{T^2}{n_1(n_1-3)} = \frac{\{U - V(\bar{X}_1 - \mu_1)/S\}^2}{W} \sim \frac{\{1/n_1 + (\bar{X}_1 - \mu_1)^2/S^2\} F_{1,n_1-2}}{n_1 - 2}.$$

Hence

$$\mathrm{pr}(T^2 \leqslant a \mid H_0) = E_{(\bar{X}_1, S^2)}\,\mathrm{pr}(F_{1,n_1-2} \leqslant \{(n_1-2)a\}/[(n_1-3)\{1 + (\bar{X}_1 - \mu_1)^2 n_1/S^2\}]).$$

Now by using the fact that $(\bar{X}_1 - \mu_1)^2 n_1/S^2 \sim (n_1-1)^{-1} F_{1,n_1-1}$, it is easy to see that (4·3) follows immediately; the result does not involve the nuisance parameter ρ. It is also easy to evaluate on a computer and therefore can be used to construct the tests when $n_2 \gg n_1$.

4·3. *Approximate null distribution in the general case*

To fit a t distribution to T_3 under H_0 we first fit, as Little (1976), a scaled chi-squared ($g\chi_f^2$) distribution to the unbiased estimate of var $(\hat{\mu}_2)$ used in (4·1) by matching their first two moments. To compute the moments of estimated var $(\hat{\mu}_2)$ we note that $\Sigma(Y_i-\bar{Y}_1)^2 \sim \sigma_2^2\chi_{n_1-1}^2$, where the summation is over $i = 1, \ldots, n_1$. Further $\Sigma(Y_i-\bar{Y}_1)^2 = \Sigma(Y_i-\hat{Y}_i)^2+\hat{\beta}^2\Sigma(X_i-\bar{X}_1)^2$ and conditional on the X_i $(1 \leqslant i \leqslant n_1)$ these two terms are independent and their distributions are as given in (4·4). The second moment of the estimate of var $(\hat{\mu}_2)$ can then be found by some routine calculations; the first moment is already known and is given in (2·1). Equating the two moments, we obtain

$$\frac{\sigma_2^2}{n_1}[1+\{\theta/(n_1-3)\}\{1-(n_1-2)\rho^2\}] = gf, \tag{4·5}$$

$$\sigma_2^4\left\{\frac{1}{(n_1-1)n^2}+\frac{\theta^2(n_1-2)(1-\rho^2)^2}{n_1^2(n_1-3)^2}+\frac{2\theta(n_1-2)(1-\rho^2)^2}{n_1(n_1-1)(n_1-3)n}\right\} = g^2 f. \tag{4·6}$$

Equations (4·5) and (4·6) yield

$$f = [1-\theta+\{\theta(n_1-2)/(n_1-3)\}(1-\rho^2)]^2$$
$$\times\left\{\frac{(1-\theta)^2}{(n_1-1)}+\frac{\theta^2(n_1-2)(1-\rho^2)^2}{(n_1-3)^2}+\frac{2\theta(1-\theta)(n_1-2)(1-\rho^2)^2}{(n_1-1)(n_1-3)}\right\}^{-1} \tag{4·7}$$

Now if we regard $\hat{\beta}$ as a fixed constant then the numerator of (4·1) is normally distributed and therefore under H_0, T_3 is approximately distributed as t_f. However, there are two difficulties in using this approximation: (i) as remarked above $\hat{\beta}$ is not a fixed constant but is a random variable, and (ii) f depends on the unknown parameter ρ. In spite of the first difficulty, which would become less serious for large n_1, we shall still attempt to fit the t distribution but with degrees of freedom $\hat{f}$, where $\hat{f}$ is obtained from (4·7) with ρ replaced by its usual estimate $\hat{\rho}$. In this manner we take care of the second difficulty.

We point out that at $\rho^2 = 1, f = n_1-1$ which agrees with the exact null distribution of T_3, since, in this case, $Y_i = \delta+\beta X_i$ almost surely for some $\beta \neq 0$, and T_3 becomes

$$\{(\delta+\beta\bar{X})\sqrt{n}\}/\{\beta^2\Sigma(X_i-\bar{X}_1)^2/(n_1-1)\}^{1/2},$$

which is distributed as t_{n_1-1}. At $\rho^2 = 0$ it can be checked using some tedious algebra that $n_1-2 \leqslant f \leqslant n_1-1$; for fixed n_1 as $\theta \to 0, f \to n_1-1$ and as $\theta \to 1, f \to n_1-2$. Further, f is concave in ρ^2 and the minimum value of $f = n_1-2$ which is attained at $\rho^2 = 0$ in the above limiting case. Also the maximum value of f is attained when

$$\rho^2 = 1-(1-\theta)/\{\theta(n_1-1)/(n_1-3)+2(1-\theta)\},$$

and this value equals

$$(n_1-1)+\{\theta(n_1-2)/(n_1-3)\}\{\theta/(n_1-3)+2(1-\theta)/(n_1-1)\},$$

which is increasing in θ for fixed n_1 and tends to $2n_1-3$ as $\theta \to 1$.

The above discussion gives some idea about the range of values of f. It also indicates that, since f is always greater than n_1-2, one can use t_{n_1-2} as a conservative approximation to the null distribution of T_3. Although this will lead to less powerful tests, it might still be useful in practice because of its ease in application. We study both these approximations by means of the Monte Carlo method in the next section.

Here we note that Little provides the approximation $t_{\hat{h}}$ to the null distribution of his statistic T_4 where

$$\hat{h} = (n-1)\{1+(1/n_1-1/n)(1-\hat{\rho}^4)\}^{-1}. \tag{4·8}$$

5. Monte Carlo study

The sampling studies were carried out for the one-sided hypothesis testing problem H_0: $\mu_2 = 0$ against H_1: $\mu_2 > 0$ at levels $\alpha = 5\%$ and 10%. The following tests, that is statistics and their associated critical regions, were compared in the study.

Test I: $T_3 > t_{\hat{f},\alpha}$, where T_3 is our test statistic given by (4·1) and $\hat{f}$ is given by (4·7);

Test II: $T_4 > t_{\hat{h},\alpha}$, where T_4 is Little's test statistic given by (4·2) and $\hat{h}$ is given by (4·8);

Test III: $T_5 > t_{n_1-1,\alpha}$, where T_5 is the usual t statistic obtained by ignoring the data on the X's, namely $T_5 = \bar{Y}_1\sqrt{n_1}/\{\Sigma(Y_i-\bar{Y}_1)^2/(n_1-1)\}^{1/2}$;

Test IV: $T_3 > t_{n_1-2,\alpha}$.

Empirical sizes and powers were computed for the following parameter values: $\mu_2 = 0$, 0·5, 1·0; $\rho = 0\,(0·3)\,0·9$ and $(n_1, n_2) = (10, 20)$, $(10, 30)$, $(10, 40)$, $(20, 20)$, $(20, 30)$, $(20, 40)$. Thus for fixed α, we have 24 observations on the empirical size of each test. Since the distributions of the statistics are free of μ_1, σ_1^2 and σ_2^2, these were taken to be 0, 1 and 1 respectively, without loss of generality. Also note that it suffices to consider only positive values of ρ for the distributions of T_3 and T_4.

We generated 1000 samples in each case. To obtain a pair of bivariate normal variables, first a pair of independent normal variables was generated using the Box–Müller algorithm; correlation was then introduced by the usual transformation. The percentage points corresponding to nonintegral t values were obtained by linear interpolation. For lack of space, we report in Table 1 detailed results only for $\alpha = 0·05$ regarding empirical sizes at for tests I, II and III at $\rho = 0$, 0·3 and 0·9, and powers at $\mu_2 = 0·5$, $\rho = 0$ and 0·9. The detailed results regarding test IV were omitted since they were quite close to that of test I although consistently lower as would be expected. The sample mean of the empirical sizes and their standard errors over 24 problems for all four tests are reported in Table 2.

Table 1. *Empirical size and power at $\mu_2 = 0·5$, of tests I, II and III*; $\alpha = 5\%$*

	$\rho = 0·0$						$\rho = 0·3$			$\rho = 0·9$					
	Size (%)			Power (%)			Size (%)			Size (%)			Power (%)		
(n_1, n_2)	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
(10, 20)	5·1	6·1	5·3	40·6	46·4	44·2	4·1	6·3	4·7	6·0	6·1	5·3	68·2	69·7	43·8
(10, 30)	5·0	6·7	5·8	40·5	44·2	42·5	4·6	6·3	4·0	5·3	6·9	4·9	72·4	76·3	43·0
(10, 40)	4·2	5·0	4·6	41·5	47·6	43·9	4·4	5·0	4·1	6·0	6·3	4·9	75·3	78·5	45·2
(20, 20)	5·8	5·0	5·4	66·9	70·0	67·4	4·5	4·9	5·1	4·8	5·0	4·7	86·9	87·5	69·3
(20, 30)	4·2	4·9	4·5	68·1	70·4	69·3	6·2	5·3	6·0	6·1	5·3	4·7	91·5	90·2	68·4
(20, 40)	4·6	5·9	5·0	66·9	71·3	67·9	5·3	5·8	5·0	6·8	5·3	4·7	93·0	93·8	71·9

* The standard error of any entry is given by $\{P(100-P)/1000\}^{1/2}$.

Table 2. *Means and standard errors of empirical sizes over 24 problems*

		Empirical size (%)				Empirical size (%)	
Test	Size (%)	Mean	Std error	Test	Size (%)	Mean	Std error
I	5	5·24	0·74	III	5	4·95	0·69
	10	10·30	0·96		10	9·90	0·94
II	5	5·64	0·73	IV	5	5·05	0·69
	10	11·37	1·00		10	10·05	0·95

In Table 2, under binomial sampling, the mean empirical sizes over 24 observations have standard errors 0·14% and 0·19% for the levels 5% and 10% respectively. Thus only Little's test, that is test II, appears to yield inflated sizes; all the other tests appear to control the sizes fairly well. Little's own study gives lower estimates for the sizes of his test; this could be partly because he used only the integral part of $\hat{h}$ in finding the critical points in his simulation study. The standard errors of the empirical sizes for levels 5% and 10% should be compared, respectively, with 0·69% and 0·95% which are the corresponding standard deviations. The high values of standard errors are indicative of sensitivity of the size to ρ and (n_1, n_2). The powers for $\rho = 0{\cdot}3$ differ little from those at $\rho = 0$.

Table 1 gives more detailed results. The size of test I tends to get slightly larger than the desired level with $|\rho|$. Also the inflation in the size of test II mainly arises at small n_1 values, but the size appears to be relatively stable with respect to ρ. The larger power of test II compared to that of test I must be discounted in view of its inflated sizes. Test I is less powerful than test III for $|\rho| \leqslant 0{\cdot}3$ but is substantially more powerful for large values of $|\rho|$; this is in agreement with the discussion in § 2·2 where we found that the regression estimator is preferred to the sample mean for large values of $|\rho|$. For fixed ρ, α and n_1, the power of test I increases with n_2 and this increase is small for small values of $|\rho|$ and moderately large for large values of $|\rho|$. On the other hand, for fixed ρ, α and n_2, the power increases rapidly with n_1 and this increase is stable relative to $|\rho|$. Finally test IV, whose results are not displayed in Table 1, is only slightly less powerful than test I. Thus in some cases the t_{n_1-2} approximation might be preferred instead of the $t_{\hat{f}}$ approximation because of its ease of application and relatively small loss of power.

Our general recommendation would be to use the test based on the $t_{\hat{f}}$ approximation to T_3 when $|\rho|$ is likely to be at least 0·3; otherwise use the ordinary t test obtained by ignoring the data on the auxiliary variables.

This research was supported in part by the Office of Research and Sponsored Programs, Northwestern University, and the National Institute of Education. The author wishes to thank Professor Robert Boruch for the N.I.E. support, Mr Marc Goodfriend for carrying out the Monte Carlo study and two referees for useful suggestions.

REFERENCES

ANDERSON, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Am. Statist. Assoc.* **52**, 200–4.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

BANERJEE, S. K. (1961). On confidence interval for two-means problem based on separate estimates of variances and tabulated values of *t*-variable. *Sankhyā* A **23**, 359–78.

BOSE, C. (1943). Note on the sampling error in the method of double sampling. *Sankhyā* **6**, 330.

KHATRI, C. G., BHARGAVA, R. P. & SHAH, K. R. (1974). Distribution of regression estimate in double sampling. *Sankhyā* C **36**, 3–22.

LEE, A. F. S. & GURLAND, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *J. Am. Statist. Assoc.* **70**, 933–41.

LIN, P. E. (1973). Procedures for testing the difference of means with incomplete data. *J. Am. Statist. Assoc.* **68**, 699–703.

LITTLE, R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrika* **63**, 593–604.

MATTHAI, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhyā* **11**, 145–52.

MORRISON, D. F. (1971). Expectations and variances of maximum likelihood estimates of the multivariate normal distribution parameters with missing data. *J. Am. Statist. Assoc.* **66**, 602–4.

Naik, U. D. (1975). On testing equality of means of correlated variables with incomplete data. *Biometrika* **62**, 615–22.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.

Scheffé, H. (1973). A statistical theory of calibration. *Ann. Statist.* **1**, 1–37.

Tikkiwal, B. D. (1960). On the theory of classical regression and double sampling estimation. *J. R. Statist. Soc.* B **22**, 131–8.

Williams, E. J. (1969). Regression methods in calibration problems. *Bull. Inst. Int. Statist.* **43**, 17–28.

[*Received May* 1977. *Revised January* 1978]