Power of double-sampling tests for General Linear Hypotheses

David Causeur, François Husson

Power of double-sampling tests for General Linear Hypotheses

DAVID CAUSEUR*†‡ and FRANÇOIS HUSSON†

†IRMAR, Agrocampus Rennes, 65 rue de St-Brieuc, CS 84215, 35042 Rennes Cedex, France
‡CREST-ENSAI, France

In this paper, testing procedures based on double-sampling are proposed that yield gains in terms of power for the tests of General Linear Hypotheses. The distribution of a test statistic, involving both the measurements of the outcome on the smaller sample and of the covariates on the wider sample, is first derived. Then, approximations are provided in order to allow for a formal comparison between the powers of double-sampling and single-sampling strategies. Furthermore, it is shown how to allocate the measurements of the outcome and the covariates in order to maximize the power of the tests for a given experimental cost.

Keywords: Auxiliary covariate; Double-sampling; F-test; Missing data; Optimal sampling design

MSC: 62D05; 62J10; 62K05

1. Introduction

Since the first works on the so-called regression estimator of a total or a mean by Cochran [1, 2], double-sampling is widely used in the context of sample survey to reduce the variance of the estimation by use of measurements of an auxiliary covariate that are available on a wider sample. The same idea can be found, for instance in Conniffe and Moran [3] , transposed to the estimation of the parameters of a multivariate normal regression model. In the agricultural applications handled by Conniffe and Moran [3] and later by Engel and Walstra [4], a surrogate and also cheaper version of the outcome is expected to be a beneficial auxiliary covariate for use in a double-sampling design.

The purpose is then to find the optimal allocation of the measurements of the outcome and of its approximate version, namely the double-sampling scheme that minimizes the variance of the estimators of the regression coefficients subject to a given experimental cost. In the presence of many auxiliary covariates, Conniffe [5], and more recently Causeur and Dhorne [6], show the enhancements that are made possible by accounting for the joint distribution of the concomitant variables when optimizing the sampling design. Furthermore, aiming

*Corresponding author. Tel.: +33-2-23-48-58-84; Fax: +33-2-23-48-58-71; Email: david.causeur@agrocampus-rennes.fr

at a more detailed definition of the optimal sampling plan by accounting for every marginal experimental cost, some extensions based on multiple-phase monotone sampling designs were also described by Causeur [7]. Analogous ideas can also be found in Breslow *et al.* [8], where the properties of estimation in non-parametric models are investigated in a double-sampling framework.

Probably because they are mostly motivated by regression examples in which the main issue is to derive a prediction formula, the previously cited works focus on the improvement of estimation that is made possible by a double-sampling design. On the contrary, very less is said about the testing procedures that traditionally complete the estimation results in a regression analysis. Such issues are encountered for instance when comparing types of soils with respect to some physical or chemical property when a few number of soil samples can be used for a reliable analysis of the property under study whereas much more numerous indirect measurement of the same property can be obtained by remote sensing. Especially in this case involving a qualitative explanatory variable, explicit expressions for the estimators of the parameters of the analysis of variance model are straightforward deduced from the existing works but, here, testing procedures for the effects of the different factors are usually much more important. The first aim of this paper is therefore to extend the double-sampling analysis by testing strategies for General Linear Hypotheses. By analogy with the well-known desirable properties of the double-sampling estimator, it is shown in the present paper how to include the measurements of the auxiliary covariates in the test statistic in order to improve its power relative to the traditional F-test used in the single-sampling context. Furthermore, in such situations where the main goal is the test of an effect rather than the estimation of parameters, a strategy is proposed for an optimal allocation of the measurements of the outcome and the auxiliary covariate, which consists, for a given experimental cost, in maximizing the power of the test with respect to the sample sizes. Calculation tools to derive asymptotic approximations of the power of the double-sampling test are provided in order to make this approach numerically possible.

In section 2, the paper introduces the sampling framework together with notations for the different models involved in the procedure. Some results on the functional relationships between the parameters of the models are recalled and the case of qualitative explanatory variables is also specifically examined in this section. Section 3 is devoted to the definition of a testing procedure for General Linear Hypotheses that takes advantage of the double-sampling framework. An asymptotic approximation of the distribution of the test statistic is given. In section 4, strategies to obtain the most powerful sampling design for a given experimental cost are presented.

2. Regression settings in the double-sampling framework

First, the main model describes linearly the conditional distribution of a continuous outcome Y given p predictors $x = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, with $p \geq 1$:

$$Y \sim \mathcal{N}\left(\beta_y^{(0)} + x\beta_y; \sigma_y^2\right),$$

where $\beta_y^{(0)}$ is the constant of the model and $\beta_y = (\beta_y^{(1)}, \ldots, \beta_y^{(p)})'$ is the p-vector of the slope parameters.

When a double-sampling scheme is used, auxiliary covariates $Z = (Z^{(1)}, \ldots, Z^{(q)})$ are introduced in the model by means of two auxiliary regression models. In the following, in order to improve the readability of the arguments, the presentation will focus on the case $q = 1$. An extension to the multivariate context can be achieved on the basis of Causeur and

Dhorne [6]. However, accounting for multiple-phase monotone sampling schemes that could be inspired by Causeur [7] is more complex and out of the purpose of this paper.

First, common observations of the outcome, the predictors and the auxiliary covariates are supposed to be observable on a sample of size n, and the following linear model is assumed for the distribution of the outcome given the predictors and the auxiliary covariate:

$$Y \sim \mathcal{N}\left(\beta_{y|z}^{(0)} + x\beta_{y|z} + \gamma Z; \sigma_{y|z}^2\right),$$

where $\beta_{y|z}^{(0)}$ is the constant of the model, $\beta_{y|z} = (\beta_{y|z}^{(1)}, \ldots, \beta_{y|z}^{(p)})'$ is the p-vector of the slope parameters for the predictors and γ is the slope parameter for the covariate.

Now, common observations of the covariate and the predictors are also available on a sample of size $N \geq n$, including the first sample, and the following linear model is assumed for the distribution of the covariate given the predictors:

$$Z \sim \mathcal{N}\left(\beta_z^{(0)} + x\beta_z; \sigma_z^2\right),$$

where $\beta_z^{(0)}$ is the constant, $\beta_z = (\beta_z^{(1)}, \ldots, \beta_z^{(p)})$ is the p-vector of slope parameters and σ_z^2 is the residual variance.

A closed-form expression of the maximum-likelihood estimators is now based on a kind of transitiveness relationship that enables to express the parameters of the main model as a combination of the parameters of the auxiliary models:

$$\begin{aligned}\beta_y &= \beta_{y|z} + \gamma\beta_z,\\ \sigma_y^2 &= \sigma_{y|z}^2 + \gamma^2\sigma_z^2.\end{aligned}$$

Although such a double-sampling procedure has mainly been used in situations where the predictors are quantitative variables, the preceding definitions are formally valid in analysis of variance contexts where the explanatory variables are qualitative.

In an illustrative purpose, let us examine the case a one way analysis of variance model with a factor taking three levels. The double-sampling design is summarized in table 1.

The preceding presentation of the model still holds where the explanatory variables are now dummy variables for the factor: for the ith level of the factor

$$Y \sim \mathcal{N}\left(\mu_y + \alpha_y^{(i)}; \sigma_y^2\right), \tag{1}$$

with, for instance, the cornerstone restriction on the main effects: $\alpha_y^{(1)} = 0$.

Table 1. An example of a double sampling design for a one way analysis of variance model with a factor taking three levels. Missing values for the response variable are identified by ?.

Sample sizes	Y	Z	X
n_1			firstst level
$N_1 - n_1$	?		
n_2			Second level
$N_2 - n_2$	?		
n_3			Third level
$N_3 - n_3$	?		

Similarly, given the covariate and the factor's level, it can be tempting to assume an analysis of covariance model with interaction:

$$Y \sim \mathcal{N}\big(\mu_{y|z} + \alpha_{y|z}^{(i)} + (\eta + \delta^{(i)})Z; \sigma_{y|z}^2\big),$$

with $\alpha_{y|z}^{(1)} = 0$ and $\delta^{(1)} = 0$. However, it is straightforward checked that the assumption of homoscedasticity in model (1) holds if and only if all the interaction parameters $\delta^{(i)}$ are zero.

In that case, the analogy between the analysis of variance model and the regression settings can be used to express the parameters of the main model as a function of the parameters of the auxiliary models:

$$\begin{aligned}
\mu_y &= \mu_{y|z} + \eta\mu_z, \\
\alpha_y &= \alpha_{y|z} + \eta\alpha_z, \\
\sigma_y^2 &= \sigma_{y|z}^2 + \eta^2\sigma_z^2.
\end{aligned}$$

Simple conditions for the estimability of the parameters of interest $(\mu_y, \alpha_y^{(i)}, \sigma_y^2)$ are deduced from the above functional relationships. Indeed, to ensure that all parameters in model (2) are estimable, both response Y and covariate Z should be observed on a sample of size $n \geq 4$, namely the number of levels plus 1, with $n = n_1 + n_2 + n_3$ and no n_i should be zero.

Moreover, as mentioned in the next section, the sampling fractions n_i/N_i will be assumed to have the same limit for large values of the sample sizes.

3. Testing procedures

In the settings introduced in the previous section, we are interested in the following test of a General Linear Hypothesis:

$$\begin{cases} \mathrm{H}_0 : \beta_y = 0 \\ \mathrm{H}_1 : \beta_y \neq 0 \end{cases}.$$

Results for tests of a particular subset of coefficients in β_y can be deduced from those that are presented thereafter.

Call Y_n and Z_n the n-vectors of the observations of the outcome and the covariate respectively on the sample of size n. Call also Z_N the N-vector of the observations of the covariate on the whole sample of size N. Finally, call $X_{(n)}$ and $X_{(N)}$ the matrices with n and N rows respectively and p columns, containing the observations of the predictors on the samples of size n and N respectively. Up to now, the same notation will be used as a sub or super scripts for other quantities, such as estimators, when they are derived on the small or large sample of size n or N respectively. Furthermore, $X_{(n)}$ is assumed to be full rank and such that the corresponding variance-covariance matrix $S_{(n)} = (1/n)X'_{(n)}(I_n - (1/n)J_n)X_{(n)}$, where J_n is the $n \times n$ matrix with all entries equal to 1, tends to a positive definite matrix Σ_x for large values of n. It is also assumed that $S_{(N)} = (1/N)X'_{(N)}(I_N - (1/N)J_N)X_{(N)}$ tends to the same limit for large values of N.

A useful way of defining the model in a double-sampling context consists in working with the $(n + N)$-vector obtained by concatenating Y_n and Z_N. Under the assumptions of section 2,

$U = (Y_n', Z_N')'$ is normally distributed with the following expectation and variance:

$$\begin{aligned}
\mathbb{E}(U) &= \begin{bmatrix} \mathbf{1}_n & X_{(n)} & \mathbf{0}_n & \mathbf{0}_{n,p} \\ \mathbf{0}_N & \mathbf{0}_{N,p} & \mathbf{1}_N & X_{(N)} \end{bmatrix} \begin{pmatrix} \beta_y^{(0)} \\ \beta_y \\ \beta_z^{(0)} \\ \beta_z \end{pmatrix}, \\
&= X\beta, \\
\mathrm{Var}(U) &= \begin{bmatrix} \sigma_y^2 I_n & \gamma\sigma_z^2 I_n & \mathbf{0}_{1,N-n} \\ \gamma\sigma_z^2 I_n & \sigma_z^2 I_N & \\ \mathbf{0}_{N-n,1} & & \end{bmatrix}, \\
&= V, \qquad (2)
\end{aligned}$$

where $\mathbf{1}_k$ and $\mathbf{0}_k$ stand for the vectors of size k which all entries equal 1 and 0 respectively and $\mathbf{0}_{k,p}$ stands for the $k \times p$ matrix with all elements equal to 0.

First, let us examine the testing procedure in the situation of known variance parameters before extending the results to the general case of unknown variance parameters. Note that assuming the parameters in V are known implies not only that σ_y and σ_z are known but also γ.

3.1 *Exact test for known variance parameters*

According to Causeur and Dhorne [6], the generalized least squares estimator $\hat{\beta}$ of β can be explicitly derived as follows:

$$\begin{aligned}
\hat{\beta} &= (X'V^{-1}X)^{-1}X'V^{-1}U, \\
&= \begin{pmatrix} \hat{\beta}_y^{(n)} + \gamma\left[\hat{\beta}_z^{(N)} - \hat{\beta}_z^{(n)}\right] \\ \hat{\beta}_z^{(N)} \end{pmatrix},
\end{aligned}$$

where $\hat{\beta}_y^{(n)}$, $\hat{\beta}_z^{(n)}$ and $\hat{\beta}_z^{(N)}$ are ordinary least squares estimators of β_y and β_z.

Now, call $\mathrm{SS}_{\beta_y} = \mathrm{SS}_{\beta_y}(\gamma, \sigma_y^2, \sigma_z^2)$ the sum-of-squares traditionally involved when testing the nullity of β_y:

$$\mathrm{SS}_{\beta_y} = \hat{\beta}_y' \ \mathrm{Var}\ (\hat{\beta}_y)^{-1}\hat{\beta}_y.$$

It can be deduced, for instance from Conniffe [5], that the former test statistic can also be expressed as follows:

$$\mathrm{SS}_{\beta_y} = \left[\frac{\hat{\beta}_y^{(n)}}{\sigma_y} + \rho\frac{\hat{\beta}_z^{(N)} - \hat{\beta}_z^{(n)}}{\sigma_z}\right]' \left[\frac{\rho^2}{N}S_{(N)}^{-1} + \frac{1-\rho^2}{n}S_{(n)}^{-1}\right]^{-1} \left[\frac{\hat{\beta}_y^{(n)}}{\sigma_y} + \rho\frac{\hat{\beta}_z^{(N)} - \hat{\beta}_z^{(n)}}{\sigma_z}\right],$$

where $\rho = \gamma\sigma_z/\sigma_y$ is the partial correlation coefficient between Y and Z given the predictors. The test statistic appears therefore as a continuum indexed by ρ between the classical standardized sum-of-squares derived on the small sample for $\rho = 0$ and on the wider sample for $\rho = 1$. Therefore, it can be expected that the present testing strategy will always be at least as good as the usual F-test based on the available measurements of the outcome and always worse than the F-test that could be derived if the outcome was observed on the whole sample.

According to well-known results in the theory of linear models, the former quadratic form is distributed according to a noncentral χ_p^2 distribution with non-centrality parameter $\beta_y[(\rho^2/N)S_{(N)}^{-1} + ((1-\rho^2)/n)S_{(n)}^{-1}]^{-1}\beta_y/\sigma_y^2$, which enables a classical testing procedure based on the exact distribution of the test statistic.

3.2 *Approximate test for unknown variance parameters*

According to Causeur and Dhorne [6], the maximum-likelihood estimators of the variance parameters are given by the following expressions:

$$\begin{aligned}\hat{\gamma} &= \frac{\hat{\sigma}_{yz}^{(n)}}{\left\{\hat{\sigma}_z^2\right\}^{(n)}},\\ \hat{\sigma}_z^2 &= \left\{\hat{\sigma}_z^2\right\}^{(N)},\\ \hat{\sigma}_y^2 &= \left\{\hat{\sigma}_y^2\right\}^{(n)} + \hat{\gamma}^2\left[\left\{\hat{\sigma}_z^2\right\}^{(N)} - \left\{\hat{\sigma}_z^2\right\}^{(n)}\right],\end{aligned}$$

where $\hat{\sigma}_{yz}^{(n)}$, $\{\hat{\sigma}_y^2\}^{(n)}$, $\{\hat{\sigma}_z^2\}^{(n)}$ and $\{\hat{\sigma}_z^2\}^{(N)}$ are the usual maximum-likelihood estimators of σ_{yz}, σ_y^2 and σ_z^2 based on residual sum-of-squares. Note that, due to the functional invariance of the maximum-likelihood estimation, ρ is estimated by $\hat{\rho} = \hat{\gamma}\hat{\sigma}_z/\hat{\sigma}_y$.

In the general framework of linear mixed models, Kenward and Roger [9] proposes a small sample inference method to account for the estimation of the variance components in the distribution of the test statistics. We propose an alternative method that is made possible by the specific covariance structure involved in our missing-data problem. Properties of this structure will for instance avoid us some technical assumptions that are used by Kenward and Roger [9].

We propose hereafter to study the testing procedure based on $\widehat{\mathrm{SS}}_{\beta_y} = \mathrm{SS}_{\beta_y}(\hat{\gamma}, \hat{\sigma}_y^2, \hat{\sigma}_z^2)$ obtained by replacing the variance parameters by their maximum-likelihood estimator in the expression of SS_{β_y}.

Example In order to see how estimating the variance parameters modifies the distribution of the test statistic, let us examine an artificial situation. Suppose here that $\sigma_y = \sigma_z = 1$. Suppose also that $p = 2$, $\beta_y = (0, 0)'$ and $\beta_z = (3, 3)'$. The sample sizes can either be ($n = 15$, $N = 30$) or ($n = 30$, $N = 80$) and the partial correlation ρ is either 0.2 or 0.8. With such values for the input parameters, 10,000 datasets are simulated as follows: first, predictors x are randomly drawn from a bivariate normal distribution with mean 0 and variance–covariance identity. Then U is randomly simulated 10,000 times according to the normal conditional distribution introduced in expression (2). In figure 1, the empirical quantiles of $\widehat{\mathrm{SS}}_{\beta_y}$ are plotted against the theoretical quantiles of a χ_2^2 distribution. Whatever the sample sizes and the partial correlation, this figure shows departures of the distribution of $\widehat{\mathrm{SS}}_{\beta_y}$ from the χ_2^2 distribution. As in the single-sampling case where the Fisher distribution replaces the χ^2-distribution when the variance parameter is estimated, the distribution of $\widehat{\mathrm{SS}}_{\beta_y}$ is more heavy-tailed than the χ_2^2 distribution. Moreover, the departures from the χ^2 distribution are more important in small-sample conditions and especially in the case of a poor partial correlation between the outcome and the covariate.

Providing an approximation of the distribution of $\widehat{\mathrm{SS}}_{\beta_y}$ by means of a combination of classical distributions is not as straightforward as in the single-sampling framework. First, the test statistic appears to be a quadratic form which kernel matrix

$$\widehat{\mathrm{Var}}(\hat{\beta}_y)^{-1} = \left[\frac{\hat{\sigma}_y^2\hat{\rho}^2}{N}S_{(N)}^{-1} + \frac{\hat{\sigma}_y^2(1-\hat{\rho}^2)}{n}S_{(n)}^{-1}\right]^{-1}$$

is the inverse of a linear combination of matrices with random coefficients. Furthermore, $\widehat{\mathrm{SS}}_{\beta_y}$ is also a quadratic function of the random vector $\tilde{\beta}_y = \hat{\beta}_y^{(n)} + \hat{\gamma}\left[\hat{\beta}_z^{(N)} - \hat{\beta}_z^{(n)}\right]$, which is not

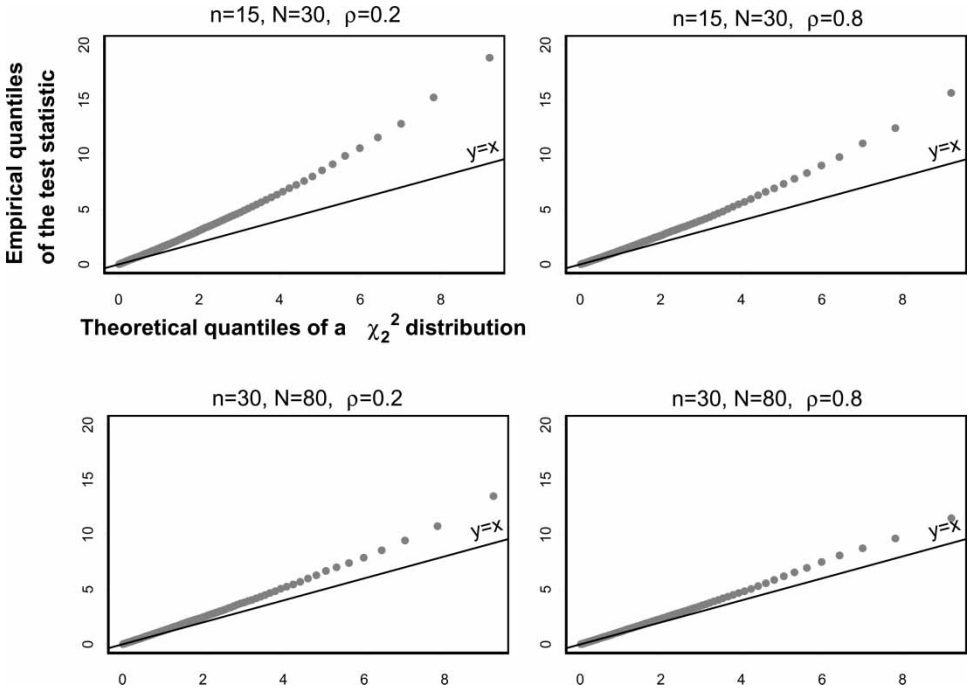

Figure 1. Empirical quantiles of $\widehat{\mathrm{SS}}_{\beta_y}$ versus the quantiles of SS_{β_y}.

independent of the kernel matrix, due to the presence of $\hat{\gamma}$ in the expression of the estimator of the regression coefficients. These two points are addressed in the following proposition giving a more tractable asymptotic approximation of $\widehat{\mathrm{SS}}_{\beta_y}$.

PROPOSITION 1 *Let* $\widehat{\mathrm{SS}}^*_{\beta_y}$ *be defined by* $\widehat{\mathrm{SS}}^*_{\beta_y} = n\hat{\beta}_y' S_{(N)} \hat{\beta}_y / \hat{\sigma}_y^2 \mathrm{K}(\hat{\rho}^2)$, *where* $\mathrm{K}(\hat{\rho}^2) = 1 - (1-\varphi)\hat{\rho}^2$ *and* $\varphi = n/N$ *is the sub-sampling fraction.*

For large n and N, $\widehat{\mathrm{SS}}_{\beta_y}$ *and* $\widehat{\mathrm{SS}}^*_{\beta_y}$ *have the same asymptotic distributions.*

Proof Let us first consider the following re-expression of $\widehat{\mathrm{Var}}(\hat{\beta}_y)^{-1}$:

$$\widehat{\mathrm{Var}}(\hat{\beta}_y)^{-1} = \frac{N}{\hat{\sigma}_y^2 \hat{\rho}^2} \left[I_p + \frac{1-\hat{\rho}^2}{\varphi \hat{\rho}^2} S_{(N)} S_{(n)}^{-1} \right]^{-1} S_{(N)}.$$

Indeed, replacing $S_{(N)} S_{(n)}^{-1}$ by its asymptotic limit I_p, at least provided that the distribution of the predictors has two finite first moments, leads to the asymptotic approximation of $\widehat{\mathrm{Var}}(\hat{\beta}_y)^{-1}$ by $nS_{(N)}/\hat{\sigma}_y^2 \mathrm{K}(\hat{\rho}^2)$. Moreover, the asymptotic expansion of the distribution of $\tilde{\beta}_y$ given in Causeur [10] shows that it is not markedly different, even in small sample conditions, of the distribution of $\hat{\beta}_y$, which proves the proposition. ■

Note that, when $n = N$, $\mathrm{K}(\hat{\rho}^2) = 1$. Both $\widehat{\mathrm{SS}}_{\beta_y}$ and $\widehat{\mathrm{SS}}^*_{\beta_y}$ coincide then with the classical F-statistic used in the single-sampling framework.

The following proposition takes advantage of the preceding results to give an approximation for the distribution of $\widehat{\mathrm{SS}}_{\beta_y}$ in terms of classical distributions.

PROPOSITION 2 *Let us define the random variate* $T_{(n,N)}(\delta_{(n,N)}, \rho^2)$ *as follows*:

$$T_{(n,N)}(\delta_{(n,N)}, \rho^2) = n\left[1 + \varphi\frac{\rho^2}{1-\rho^2}\right]\frac{T_1}{T_2 + \varphi^2 T_3},$$

where T_1, T_2 *and* T_3 *are mutually independent. In addition,* T_1 *is distributed according to a non-central* χ^2_p *with non-centrality parameter* $\delta_{(n,N)}/\mathrm{K}(\rho^2)$ *with* $\delta_{(n,N)} = n\beta'_y S_{(N)}\beta_y/\sigma^2_y$ *and* T_2 *is distributed according to a* χ^2_{n-p-2} *distribution. Suppose now that B and S are independent random variates following respectively a beta distribution* $\mathcal{B}([n-p-1]/2, [N-n]/2)$ *and a* χ^2_{N-p-1}, *then* T_3 *is conditionally distributed, given B and S, as the ratio between a non-central chi-square variable with* 1 *degree of freedom and non-centrality parameter* $[\rho^2/(1-\rho^2)]BS$ *and B.*

$\widehat{\mathrm{SS}}_{\beta_y}$ *and* $T_{(n,N)}(\delta_{(n,N)}, \rho^2)$ *have the same limiting distributions when n and N are large.*

Proof First, it is deduced from proposition 1 that the asymptotic distributions of $\widehat{\mathrm{SS}}_{\beta_y}$ and $\widehat{\mathrm{SS}}^*_{\beta_y}$ are the same. Moreover,

$$\widehat{\mathrm{SS}}^*_{\beta_y} = n\left[1 + \varphi\frac{\rho^2}{1-\rho^2}\right]\frac{T_1}{T_2 + \varphi^2 T_3},$$

with $T_1 = n\hat{\beta}'_y S_{(N)}\hat{\beta}_y/\sigma^2_y\mathrm{K}(\rho^2)$, $T_2 = [Y'_{(n)}(P_{(n)} - P^*_{(n)})Y_{(n)}]/\sigma^2_y(1-\rho^2)]$ and $T_3 = [Y'_{(n)}\, P^*_{(n)}\, Y_{(n)}/\sigma^2_y(1-\rho^2)]/[Z'_{(n)}P_{(n)}Z_{(n)}/Z'_{(N)}P_{(N)}Z_{(N)}]$. In the former definitions, $P_{(n)}$ and $P_{(N)}$ are the orthogonal projectors onto the linear spaces orthogonal to the spaces spanned by the columns of $X_{(n)}$ and $X_{(N)}$ respectively and $P^*_{(n)} = P_{(n)}Z_{(n)}Z'_{(n)}P_{(n)}/Z'_{(n)}P_{(n)}Z_{(n)}$.

With the same arguments that were used to achieve proposition 1, T_1 is shown to follow asymptotically a noncentral χ^2_p with non-centrality parameter $\delta_{(n,N)}/\mathrm{K}(\rho^2)$. The former statistic depending only on estimators of expectation parameters, it is independent of T_2 and T_3. Moreover, the independence of T_2 and T_3 is made obvious by considering their conditional joint distribution given the covariate and the predictors and the distribution of T_2 is straightforward deduced from the classical theory of the distributions of quadratic forms. Note finally that $[Y'_{(n)}P^*_{(n)}Y_{(n)}/\sigma^2_y(1-\rho^2)]$ is distributed as a χ'^2_1 variate with non-centrality parameter $[\rho^2/(1-\rho^2)][Z'_{(n)}P_{(n)}Z_{(n)}/\sigma^2_z]$. Finally, it is deduced from Johnson *et al.* [11, p. 349] that $B = Z'_{(n)}P_{(n)}Z_{(n)}/Z'_{(N)}P_{(N)}Z_{(N)}$ and $S = Z'_{(N)}P_{(N)}Z_{(N)}/\sigma^2_z$ are independently distributed according to a $\mathcal{B}([n-p-1]/2, [N-n]/2)$ and a χ^2_{N-p-1} respectively. ■

The former proposition provides calculation tools to derive approximations of the quantiles or the probability distribution function of $\widehat{\mathrm{SS}}_{\beta_y}$ by means of Monte-Carlo simulations. This is therefore very useful to maximize the power of the present test with respect to the sample sizes. This point is addressed in the next section. Note that $\delta_{(n,N)}$ is a non-centrality parameter, which expression is similar to the equivalent parameters encountered when testing the global significance of the parameters in the classical framework. By the way, if $n = N$, the distribution of $(1/p)T_{(n,N)}(\delta_{(n,N)}, \rho^2)$ coincides with the non-central Fisher distribution involved for the same test in a single sampling context.

Example In the simulation feature introduced above, the quantiles of the distribution given in proposition 3.2 are plotted in figure 2 against the theoretical quantiles of a χ^2_2 distribution, together with the quantiles of $\widehat{\mathrm{SS}}_{\beta_y}$. Obviously, this figure confirms the good approximation of the distribution of $\widehat{\mathrm{SS}}_{\beta_y}$ by the distribution of $T_{(n,N)}(\sigma^2_y, \rho^2)$ even in small-sample conditions.

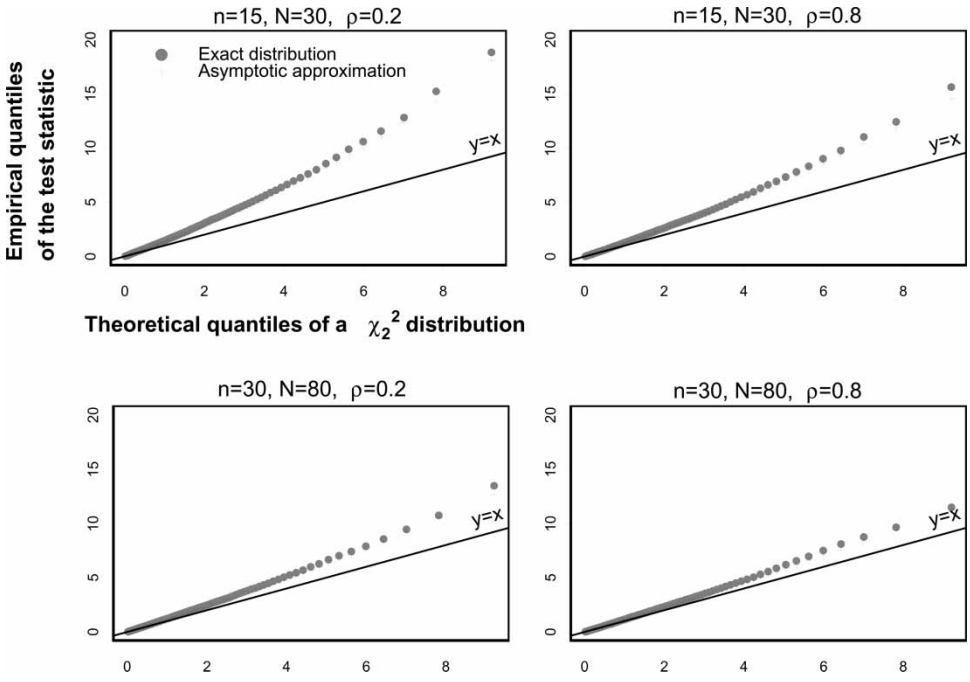

Figure 2. Approximation of the quantiles of $\widehat{\mathrm{SS}}_{\beta_y}$ by the quantiles of $T_{(n,N)}(\sigma_y^2, \rho^2)$.

4. Optimal sampling designs

In the single-sampling context, where the sample size is say m, the power of the F-test is defined with respect to a given standardized distance from zero $\delta = \beta_y' S_{(m)} \beta_y / \sigma_y^2$ of the vector of expectation parameters. This distance from zero, that appears in the non-centrality parameter of the F-distribution under the alternative hypothesis, is chosen as the minimal value over which the test is expected to reject the null hypothesis with a high probability. The task consisting in giving a relevant value for δ is by the way much easier in the case of a one-way analysis of variance model where it can be expressed more intuitively in terms of the sum of the squares of the marginal effects parameters.

Let c_y be the unitary experimental cost for the joint measurements of the outcome and the predictors. It is interesting to keep in mind that double-sampling designs are beneficial in the situations where c_y is much larger than c_z, namely the unitary experimental cost for the measurement of the covariate and the predictors. The global cost for a regression experiment involving a single sample, which size is m, is therefore mc_y. Under these experimental conditions, the power of the F-test for the significance of β_y is denoted $\pi_m(\delta)$.

Now, let c_{yz} denote the unitary experimental cost for the joint measurements of the outcome, the covariate and the predictors. Note that, in many situations where the covariate is an intermediate measurement of the outcome, $c_{yz} = c_y$. The global cost for a double-sampling design is therefore $c_{yz}n + c_z(N - n)$. The issue is now to provide the double-sampling design which maximizes the power $\pi_{(n,N)}(\delta)$ of the test subject to the following restriction on the subsequent experimental cost:

$$c_{yz}n + c_z(N - n) = mc_y.$$

In other words, the object is to seek for the optimal sampling design among those which experimental cost is equivalent to a single-sampling strategy with m experimental units.

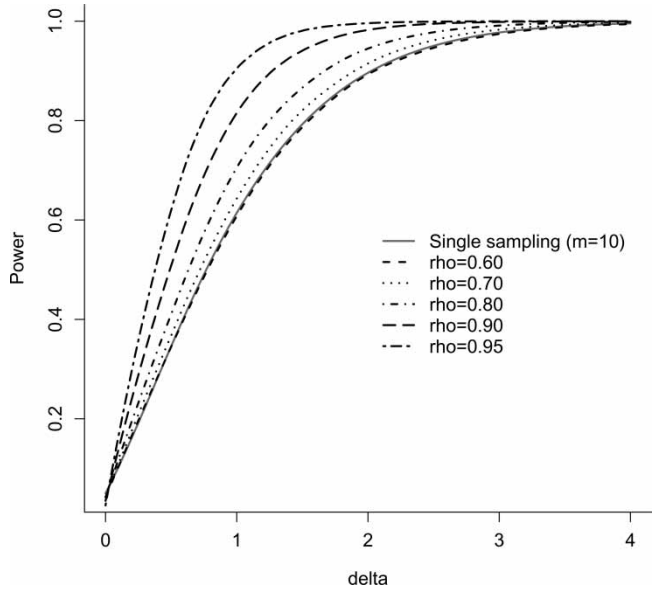

Figure 3. Powers for the optimal double-sampling tests.

On the basis of the approximation provided in proposition 2, the power $\pi_{(n,N)}(\delta)$ can be evaluated by Monte-Carlo simulations. Given δ, the non-centrality parameter $\delta_{(n,N)}$ that appears in the distribution of $T_{(n,N)}(\delta_{(n,N)}, \rho^2)$ can be approximated by $N\delta/\mathrm{K}(\rho^2)$ for large N and m. Therefore, as for the optimization of double-sampling designs with respect to the efficiency of estimation, a prior estimation of ρ^2 is also needed here to obtain the optimal sample sizes.

In order to illustrate the derivation of optimal double-sampling designs, consider the following situation: the unitary costs involving the outcome are supposed to be $c_{yz} = c_y = 20$ and $c_z = 4$. For equally spaced value of δ between 0 and 4, the optimal double-sampling design, with the same experimental cost as the single sampling design with $m = 10$, is calculated. Figure 3 shows the power functions at level $\alpha = 0.05$ for different values of ρ. For large values of ρ, the figure shows that important gains in terms of power of the test can be expected from a double-sampling strategy here. However, when ρ becomes poorer, the optimal double-sampling design coincides with the single-sampling feature.

Conversely, suppose now that the objective power is fixed to 0.9 for a given δ, say $\delta = 1$. By a single-sampling approach, the problem is quite usual and consists in solving the following equation in m: $\pi_m(0.1) = 0.9$. Transposed in the double-sampling context, the issue is similarly to find (n, N) such that $\pi_{(n,N)}(0.1) = 0.9$. In the situations presented above, the solutions are

Table 2. Experimental cost needed to reach the power 0.9 with $\delta = 1$ by single-sampling and double-sampling strategies.

	Single sampling test	Double-sampling tests				
		$\rho = 0.95$	$\rho = 0.90$	$\rho = 0.80$	$\rho = 0.70$	$\rho = 0.60$
Sample sizes	$m = 17$	$(n, N) =$ (6, 28)	$(n, N) =$ (8, 26)	$(n, N) =$ (10, 28)	$(n, N) =$ (12, 26)	$(n, N) =$ (14, 24)
Cost	340	208 (−39%)	232 (−32%)	272 (−20%)	292 (−13%)	320 (−6%)

given in table 2. Here again, the results show that, provided c_z is low enough, the double-sampling approach can help saving experimental cost relative to the single-sampling plan.

5. Discussion

The present paper gives the asymptotic distribution of a test statistic for General Linear Hypotheses in a double-sampling context. Such a result is first useful after a double-sampling experiment to complete the analysis based on estimates by tests of the significance of the predictors. Moreover, in situations where measurements of the outcome are very expensive but can be approximated by a cheaper and highly correlated version, sampling designs aiming either at a reduction of the global experimental cost or at an improvement of the power of the test can be deduced from a closed form expression of the asymptotic distribution of the test statistic in terms of classical distributions.

The double-sampling framework and its multivariate monotone extension are particularly studied patterns of missing data because they enable explicit expressions for the maximum-likelihood estimators of the parameters of a joint multinormal distribution.

The present results take advantage of this desirable property to achieve testing procedures, at least in the case of a two-phase sampling. For convenience, the results are indeed presented in the case of only one covariate but they can straightforward be extended to a multivariate context on the basis of Causeur and Dhorne [6]. However, for more complex sampling schemes in the presence of many covariates such as the monotone designs proposed by Causeur [7], the joint distribution of the variance parameters is far more tedious to be obtained in terms of known distributions. Equivalent studies in the former context are however worthwhile to face specific applications for example to time series or growth curve modelling where monotone patterns of missing data are encountered.

References

[1] Cochran, W.G., 1963, *Sampling Techniques* (2nd edn) (New-York: Wiley).
[2] Cochran, W.G., 1977, *Sampling Techniques* (3rd edn) (New-York: Wiley).
[3] Conniffe, D. and Moran, M.A., 1972, Double sampling with regression in comparative studies of carcass composition. *Biometrics*, **28**, 1011–1023.
[4] Engel, B. and Walstra, P., 1991, Increasing precision or reducing expense in regression by using information from a concomitant variable. *Biometrics*, **47**, 13–20.
[5] Conniffe, D., 1985, Estimating regression equations with common explanatory variables but unequal numbers of observations. *Journal of Econometrics*, **27**, 179–196.
[6] Causeur, D. and Dhorne, T., 1998, Finite-sample properties of a multivariate extension of double-regression. *Biometrics*, **54**(4), 1591–1601.
[7] Causeur, D., 2005, Optimal sampling from concomitant variables for regression problems. *Journal of Statistical planning and Inference*, **128**, 289–301.
[8] Breslow, N., McNeney, B. and Wellner, J.A., 2003, Large sample theory for semi-parametric regression models with two-phase, outcome-dependent sampling. *Annals of Statistics*, **31**(4), 1110–1139.
[9] Kenward, M.G. and Roger, J.H., 1997, Small sample inference for fixed effects estimators from restricted maximum likelihood. *Biometrics*, **53**, 983–997.
[10] Causeur, D., 1999, Exact distribution of the regression estimator in double-sampling. *Statistics*, **32**, 297–315.
[11] Johnson, N.L., Kotz, S. and Balakrishnan, N., 1994, *Continuous Univariate Distributions*, Vol. 1 (2nd edn) (John Wiley and Sons).