Sample Size Calculation for Method Validation Using Linear Regression

E. A. Colosimo, F. R. B. Cruz[†], J. L. O. Miranda

Departament of Statistics,
Federal University of Minas Gerais,
31270-901 - Belo Horizonte - MG, Brazil
E-mail: {enricoc,fcruz,jlom}@est.ufmg.br

T. van Woensel

Department of Operations, Planning, Accounting and Control,
Eindhoven University of Technology,
Eindhoven, The Netherlands
E-mail: T.v.Woensel@tm.tue.nl

December 4, 2006

Abstract

In this paper, we present a method for sample size calculation for studies involving both the intercept and slope parameters of a simple linear regression model. Some methods have been proposed in the literature to determine the adequate sample size. However, they are usually based on the line slope only. We propose a method based on the F statistic that involves both the intercept and the slope parameters of the model. The validation process is conducted by fitting

[†]Corresponding author: Prof. Frederico Cruz. E-mail: fcruz@est.ufmg.br. Phone: (+55 31) 3499 5929. Fax: (+55 31) 3499 5929.

a simple linear regression model and by testing a zero intercept and unity slope hypothesis. Compared to a traditional method and using Monte Carlo simulations, encouraging results attest for the clear superiority of the proposed method. The paper ends with a real-life example showing the value of the new method in practice.

Keywords: $\mathcal{F}$ distribution, measurement method, non-central parameter, type II error.

1 Introduction

Sample size determination is a fundamental aspect in scientific research. In this paper, sample size calculation is related to a statistical decision expressed by means of a test of hypotheses. Under this situation, a significance level α controls the probability of a type I error and an adequate sample size generates a certain power for the specified value under the alternative hypothesis. If the sampling regime is however too small, statistical tests may fail in detecting the hypothesized difference.

The motivation of this work lies on various real-life problems. Assessing the accuracy of a new method (e.g., in chemical engineering, Linnet, 1990, production engineering etc.) is a fundamental step in the method validation process. A newly proposed method is often compared with the results of a reference method in order to evaluate its performance (Lin, 1989). The validation consists in testing a zero intercept and unity slope hypothesis under which the two methods are statistically equivalent. Crucially in the comparison between two methods is the sampling process. Collecting the correct sample is a difficult and costly task which should be done with care. Therefore, it is important to have a performing and cost-effective method-

ology to determine the best sample size to assure that potential differences between two methods can be observed. In the literature, a number of statistical techniques for sample size and power calculations in linear regression are proposed. For instance, Hintze (1996) provided a method based on the magnitude of the correlation coefficient. Kraemer & Thiemann (1987), Sokal & Rohlf (1995), and Cohen (1988) also provided methods to detect a correlation coefficient of a certain magnitude. Dupont & Plummer (1998) proposed a method to detect a regression slope of a given size.

However, all these techniques are based on the correlation coefficient or, in a similar fashion, on the slope of the model. Therefore, these results do not apply perfectly well to the core problem of method validation because such a test should consider both the intercept and the slope parameters of the simple regression line. In this paper, we introduce a new method to determine sample sizes for the problem of method validation, which is based on the F statistic for the joint test involving the intercept and the slope of the regression line. Under the alternative hypothesis, we propose an approximation for the non-centrality parameter of the $\mathcal{F}$ distribution in order to obtain the desired sample size.

The outline of the paper is as follows. In Section 2, the notation is expressed in terms of the linear regression model and the problem background is established. Section 3 presents two methods for sample size determination. The first method considers separate calculations for each individual parameters of the linear regression model. Based on this method, two sample sizes are obtained. Then, the choice of the adequate sample size is proposed to be either (1) the largest one or (2) the average between the two sample size estimates. The other is a newly developed method, based on the F statistics. Monte Carlo simulation results comparing both methods appear in Section 4.

Section 5 describes the results of a case study based on a real-life situation. Section 6 ends the paper with final remarks and topics for future research in the area.

2 The Linear Regression Model and the Problem of Method Validation

Consider the well known simple linear regression model

$$y_i = b_0 + b_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{1}$$

in which y_i is the response variable, x_i is the regression variable, b_0 and b_1 are the intercept and the slope, respectively, which are the parameters to be estimated, ε_i is the error such that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, and n is the sample size. In our problem setting, x corresponds to the reference method and y to the newly proposed method.

Least squares estimators (Draper & Smith, 1998) for the parameters b_0 and b_1 and the unbiased estimator for σ^2 are then

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x},$$

$$\widehat{b}_1 = \frac{S_{xy}}{S_{xx}},$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-2},$$

in which $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$, $S_{xy} = \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})$, $\overline{x} = \sum_{i=1}^{n} x_i/n$, $\overline{y} = \sum_{i=1}^{n} y_i/n$, and $\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i$. The variances of the parameter estimators are

$$\mathrm{Var}(\widehat{b}_0) = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}), \tag{2}$$

and

$$\mathrm{Var}(\widehat{b}_1) = \frac{\sigma^2}{S_{xx}}. \tag{3}$$

Evaluating the method validation problem relies on the following hypotheses:

$$\mathrm{H}_0 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ \textit{vs.}\ \mathrm{H}_1 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In other words, under H_0 the two methods are equivalent. The test statistic used for this simultaneous inference is (Seber, 1977):

$$F_{\text{calc}} = \frac{\left[\left(b_0 - \hat{b}_0\right)^2 + 2\bar{x}\left(b_0 - \hat{b}_0\right)\left(b_1 - \hat{b}_1\right) + (\sum_{i=1}^{n} x_i^2/n)\left(b_1 - \hat{b}_1\right)^2\right]}{2s^2/n}$$

which under H_0 is well known to have the central $\mathcal{F}$ distribution

$$F_{\text{calc}} \sim \mathcal{F}_{2,n-2}.$$

3 Methods for Sample Size Calculation

In this section we propose two methods to get an adequate sample size to guarantee a certain power for a specific value under the alternative hypothesis. The first method is naïve in the sense that it is based on sample sizes calculations for each parameter of the model separately. The second proposed method is based on the non-central $\mathcal{F}$ distribution. The following conditions are necessary in order to determine the sample size:

- It is necessary to specify a point in the alternative hypothesis, H_1 to guarantee a power of $1-\beta$. Let us consider a generic point $(\delta_0, 1+\delta_1)$, $\delta_0, \delta_1 > 0$, for developing both methods, that is

$$H_0 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \textit{ vs. } H_1 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \delta_0 \\ 1+\delta_1 \end{pmatrix}.$$

- It is necessary to have past information concerning the parameters of the model or in an alternative way a pilot sample in order to get such information, which is an usual procedure in many practical studies. Just as an example, Cohen's method (Cohen, 1988) is also developed under the assumption of the existence of a pilot data set.

3.1 Method 1: Simple Approach

A very simple way to achieve the proposed goal is to separately detect differences in the intercept and in the slope. Of course, in this case two possible sample sizes are quantified (one for the slope and one for the intercept). As such, one needs to select or combine these two sample sizes.

Slope hypotheses

For the following hypotheses concerning the slope

$$H_0 : b_1 = 1 \textit{ vs. } H_1 : b_1 = 1 + \delta_1,$$

the number of pairs (x_i, y_i) to detect a difference $\delta_1 = b_1 - 1$, with at least power $1-\beta$, according to Equation (1) and (3), is equal to

$$n_{b_1} \geq \frac{\sigma^2 \left(t_{(\alpha/2, n_{b_1}-2)} + t_{(\beta, n_{b_1}-2)} \right)^2}{\delta_1^2 s_x^2} + 1, \tag{4}$$

in which $s_x^2 = S_{xx}/(n-1)$.

Of course, multiple solutions are possible for the expression (4) but we need the smallest n_{b_1} that results in the required power. From Expression (4), it is observed that in order to check for the feasibility of a candidate n_{b_1} (that is, if the inequality holds), one needs to know $t_{(\alpha/2,n_{b_1}-2)}$ and $t_{(\beta,n_{b_1}-2)}$, the quantiles of the student-t distribution, which clearly depend on n_{b_1}. Thus, to solve this dilemma an iterative solution scheme must be employed. We propose to use the well-known bisection method (Burden & Faires, 2005), which is accurate, fast, and considerably easy to implement. We would like to point out however that many other classical optimization algorithms (see, for instance, Burden & Faires, 2005) could be used. The algorithm is presented in Figure 1 in pseudo-code, ready to be implemented in any programming language.

In order to better understand the proposed algorithm, consider Figure 2, which represents the bisectrix of the first quadrant and the following function $f(n)$, defined as

$$f(n) = \frac{\sigma^2 \left(t_{(\alpha/2,n-2)} + t_{(\beta,n-2)}\right)^2}{\delta_1^2 s_x^2} + 1,$$

for one of the cases ($\sigma = 1.0$, $n_{\text{pilot}} = 16$, and $\delta = 0.8$) simulated in Section 4.

Clearly, the solutions for Expression (4) are the points below the bisectrix, that is, points for which $n \geq f(n)$, as seen in Figure 2. Additionally, $n - f(n)$ has opposite signs at a very low n_{lower} and at a very high n_{upper}. It follows that there exists a minimum such that $n \geq f(n)$ holds. Once $f(n)$ is a non-increasing function and a solution has bracketed between these two values of n (the lower and upper bounds), the bisection algorithm can be used to narrow this interval very effectively. The lowest upper bound n^* is the sample size sought.

algorithm

{read input}

read α, β, δ_1, *and a pilot sample* (x_i, y_i) *of size* n_0

$\sigma^2 \leftarrow \sum_{i=1}^{n_0}(y_i - \hat{y}_i)^2/(n_0 - 2)$

$\bar{x} \leftarrow \sum_{i=1}^{n_0} x_i/n_0$

$s_x^2 \leftarrow \sum_{i=1}^{n_0}(x_i - \bar{x})^2/(n_0 - 1)$

{get lower bound}

$n_{\mathrm{lower}} \leftarrow n_0$

$f(n_{\mathrm{lower}}) \leftarrow \left\lceil \sigma^2 \left(t_{(\alpha/2, n_{\mathrm{lower}}-2)} + t_{(\beta, n_{\mathrm{lower}}-2)} \right)^2 / (\delta_1^2 s_x^2) + 1 \right\rceil$

{get upper bound}

$n_{\mathrm{upper}} \leftarrow n_0$

$f(n_{\mathrm{upper}}) \leftarrow \left\lceil \sigma^2 \left(t_{(\alpha/2, n_{\mathrm{upper}}-2)} + t_{(\beta, n_{\mathrm{upper}}-2)} \right)^2 / (\delta_1^2 s_x^2) + 1 \right\rceil$

while $n_{\mathrm{upper}} < f(n_{\mathrm{upper}})$

$n_{\mathrm{upper}} \leftarrow 2n_{\mathrm{upper}}$

$f(n_{\mathrm{upper}}) \leftarrow \left\lceil \sigma^2 \left(t_{(\alpha/2, n_{\mathrm{upper}}-2)} + t_{(\beta, n_{\mathrm{upper}}-2)} \right)^2 / (\delta_1^2 s_x^2) + 1 \right\rceil$

end while

{narrow interval by bisection}

repeat

$n_{\mathrm{cnd}} \leftarrow \lceil (n_{\mathrm{lower}} + n_{\mathrm{upper}})/2 \rceil$

$f(n_{\mathrm{cnd}}) \leftarrow \left\lceil \sigma^2 \left(t_{(\alpha/2, n_{\mathrm{cnd}}-2)} + t_{(\beta, n_{\mathrm{cnd}}-2)} \right)^2 / (\delta_1^2 s_x^2) + 1 \right\rceil$

if $n_{\mathrm{cnd}} > f(n_{\mathrm{cnd}})$ **then**

$n_{\mathrm{upper}} \leftarrow n_{\mathrm{cnd}}$

else

$n_{\mathrm{lower}} \leftarrow n_{\mathrm{cnd}}$

end if

until $n_{\mathrm{upper}} - n_{\mathrm{lower}} \leq 1$

{print results}

print n_{upper}

end algorithm

Figure 1: Iterative algorithm to compute n_{b_1}.

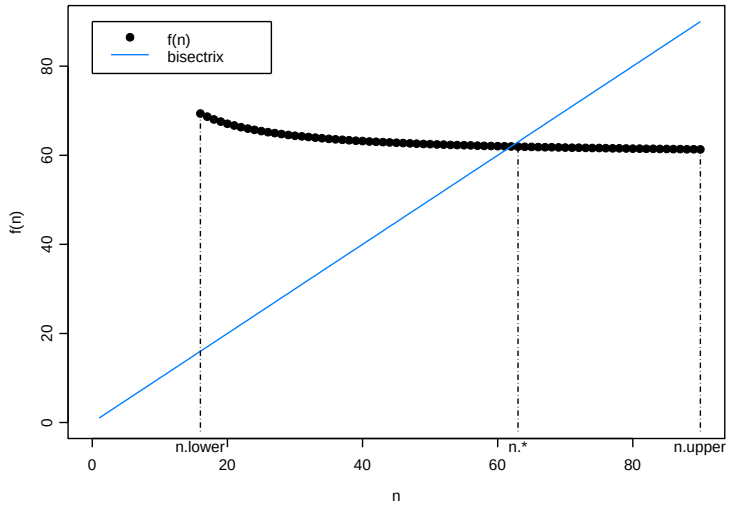

Figure 2: Graphical interpretation of the iterative method.

Intercept hypotheses

In a similar way, the hypotheses for the intercept are:

$$\mathrm{H}_0 : b_0 = 0 \ \textit{vs.}\ \mathrm{H}_1 : b_0 = 0 + \delta_0.$$

Similarly, it can be shown that the number of pairs (x_i, y_i) to detect a difference $\delta_0 = b_0 - 0$, with at least power $1 - \beta$ according to Equation (1) and (2) is

$$n_{b_0} \geq \frac{\sigma^2 \left(t_{(\alpha/2, n_{b_0}-2)} + t_{(\beta, n_{b_0}-2)}\right)^2 (s_x^2 + \bar{x}^2)}{{\delta_0}^2 s_x^2}. \tag{5}$$

Here we also need to use an iterative scheme, similar to the algorithm in Figure 1. Indeed, in order to compute the value n_{b_0} we need to know quantities, $t_{(\alpha/2, n_{b_0}-2)}$ and $t_{(\beta, n_{b_0}-2)}$, that depend on it, as seen from Eq. (5).

Determine the sample size

Because two individual tests (slope and intercept separately) are employed, two sample sizes usually are found. At least two strategies are possible. First, the sample size to be used is the largest of the two, that is:

$$n_{\text{sample}} = \max(n_{b_1}, n_{b_0}). \tag{6}$$

or, secondly, one can take the average between these two quantities, that is:

$$n_{\text{sample}} = (n_{b_1} + n_{b_0})/2. \tag{7}$$

We remark that Equation (6) and (7) are naïve strategies to determine the sample size of interest. Note that Equations (4) and (5) are well-known expressions from the literature.

3.2 Method 2: Joint Approach

Under the alternative point $(b_0, b_1) = (\delta_0, 1 + \delta_1)$, F_{calc} has a non-central $\mathcal{F}$ distribution (Graybill, 1976)

$$F_{\text{calc}} \sim \mathcal{F}_{2,n-2,\lambda},$$

in which

$$\lambda = \frac{\sum_{i=1}^{n}(\delta_0 + \delta_1 x_i)^2}{\sigma^2}. \tag{8}$$

A natural way to turn Equation (8) into a more useful expression in practical situations is to replace x_i by its expectation $\mathrm{E}(x)$. Using this approximation for the particular case in which $\delta_0 = \delta_1 = \delta$, we get:

$$\lambda = \frac{n\delta^2 \left[1 + \mathrm{E}(x)\right]^2}{\sigma^2}, \tag{9}$$

in which $\mathrm{E}(x)$ will be replaced by the average of the pilot sample $\bar{x}$.

The proposed algorithm to compute the sample size is shown in Figure 3. Note that the lower and upper bounds must be determined for the sample size and again, an iterative scheme must be used in order to compute the lowest n that results in the given power β. Because of its simplicity and efficiency, we also choose the bisection method (Burden & Faires, 2005), but many other similar algorithms could be used.

Note that the estimation of the effective power $\beta_{\text{effective}}$ for a given sample size n can be obtained by means of the following expression:

$$\beta(n, \lambda)_{\text{effective}} = 1 - \text{pf}\Big[\text{qf}(1 - \alpha, 2, n - 2), 2, n - 2, \lambda\Big],$$

in which pf is the cumulative probability of the non-central $\mathcal{F}$ distribution, with non-central parameter λ, as defined in Equation (9), and qf is the quantile of the central $\mathcal{F}$ distribution.

4 Simulation Study

The algorithms were implemented in S-PLUS (2000) and are available from the authors upon request. The goal of the simulation study is to evaluate the performance of the algorithms. Simulations were set for 100 pilot samples of size n_{pilot}, each of them to detect a deviation δ from H_0. That is

$$\mathrm{H}_0 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \textit{ vs. } \mathrm{H}_1 : \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \delta \\ 1 + \delta \end{pmatrix}.$$

Each sample size should guarantee a pre-specified significance level α and power $1 - \beta$, that were estimated by Monte Carlo simulations. Thus, in order to verify the type I error, 500 samples were generated under H_0, for each pilot sample considered, and the proportion of rejections was computed.

algorithm

{read input}

read α, β, δ, *and a pilot sample* (x_i, y_i) *of size* n_0

$\mathrm{E}(x) \leftarrow \sum_{i=1}^{n_0} x_i / n_0$

$\sigma^2 \leftarrow \sum_{i=1}^{n_0} (y_i - \hat{y}_i)^2 / (n_0 - 2)$

{find lower and upper bounds}

$i \leftarrow 0$

$\lambda \leftarrow n_i \delta^2 \left[1 + \mathrm{E}(x)\right]^2 / \sigma^2$

compute $\beta(n_i, \lambda)_{\text{effective}}$

$n_{\text{lower}} \leftarrow n_0$

while $\beta(n_i, \lambda)_{\text{effective}} < \beta$

$i \leftarrow i + 1$

$n_i \leftarrow 2n_{(i-1)}$

$\lambda \leftarrow n_i \delta^2 \left[1 + \mathrm{E}(x)\right]^2 / \sigma^2$

compute $\beta(n_i, \lambda)_{\text{effective}}$

end while

$n_{\text{upper}} \leftarrow n_i$

{reduce interval by bisection}

repeat

$n_{\text{cnd}} \leftarrow \text{ceiling}\left[(n_{\text{lower}} + n_{\text{upper}})/2\right]$

$\lambda \leftarrow n_{\text{cnd}} \delta^2 \left[1 + \mathrm{E}(x)\right]^2 / \sigma^2$

compute $\beta(n_i, \lambda)_{\text{effective}}$

if $\beta(n_i, \lambda)_{\text{effective}} > \beta$

$n_{\text{upper}} \leftarrow n_{\text{cnd}}$

else

$n_{\text{lower}} \leftarrow n_{\text{cnd}}$

end if

until $(n_{\text{upper}} - n_{\text{lower}} \leq 1)$

{print results}

print n_{upper}

end algorithm

Figure 3: Algorithm to compute n.

Conversely, to verify the type II error, 500 samples were generated under H_1 and the proportion of non-rejections was computed. In the simulations the nominal values were fixed at $\alpha = 0.05$ and $\beta = 0.10$. In order to study the effects of the parameters in the sample size calculations, the pilot sample sizes were assumed to be $n_{\text{pilot}} = \{8, 16\}$, the error variance, $\sigma = \{1.0, 2.0\}$, and $\delta = \{0.2, 0.4, 0.8\}$. The values of x were generated as a gamma distribution with shape and rate parameters 2.0 and 2.0.

The results are presented in Table 1 and summarized in Figures 4 and 5. It is seen that the sample sizes obtained by methods #1 and #1-b are in general higher than those obtained by method #2, as seen in Figure 4. Although all methods guarantee the established nominal type I error and only in one case (**0.0477**) the 95% confidence interval did not cover the nominal value, $\alpha = 0.05$, the empirical type II errors for methods #1 and #1-b are found to be way off. However, for method #2 the errors are quite close to the nominal level, as seen in Figure 5. Notice from Table 1 that only for method #2 the 95% confidence intervals cover the established nominal type II error ($\beta = 0.10$). Additionally, as expected, the mean and median sample sizes increase for both methods when σ increases and δ decreases, as shown in Figure 4. The sample sizes decrease slightly when the pilot sample size n_{pilot} increases for methods #1 and #1-b but remain almost unchanged for method #2. Both sample sizes strategies of method #1 presented similar results. However there is a little improvement in using the average, Eq. (7), over the largest values, Eq. (6). The distribution of the sample size is slightly skewed since the median is always smaller than the mean. Concluding, methods #1 and #1-b obtained much larger samples sizes than it is necessary to attain the target power. This is an important insight as, in real-life situations, sampling is usually expensive.

Table 1: Monte Carlo simulation results for the sample size calculations.

				n_{sample}		type I error			type II error		
method	σ	n_{pilot}	δ	mean	median	mean	$\Delta^{(*)}$	median	mean	$\Delta^{(*)}$	median
1	1.0	8	0.2	976	694	0.0502	0.0021	0.0500	0.0005	0.0009	$< 10^{-4}$
			0.4	230	182	0.0501	0.0019	0.0500	0.0008	0.0011	$< 10^{-4}$
			0.8	76	60	0.0509	0.0018	0.0500	0.0005	0.0007	$< 10^{-4}$
		16	0.2	841	723	0.0487	0.0020	0.0480	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.4	225	183	0.0505	0.0020	0.0500	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.8	63	53	0.0497	0.0017	0.0500	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
	2.0	8	0.2	3708	2645	0.0490	0.0018	0.0490	$< 10^{-4}$	0.0001	$< 10^{-4}$
			0.4	976	694	0.0502	0.0021	0.0500	0.0005	0.0009	$< 10^{-4}$
			0.8	230	182	0.0501	0.0019	0.0500	0.0008	0.0011	$< 10^{-4}$
		16	0.2	3031	2499	0.0519	0.0019	0.0520	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.4	841	723	0.0487	0.0020	0.0480	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.8	225	183	0.0505	0.0020	0.0500	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
1-b	1.0	8	0.2	914	740	0.0493	0.0018	0.0490	0.0025	0.0030	$< 10^{-4}$
			0.4	223	178	0.0487	0.0018	0.0480	0.0073	0.0122	$< 10^{-4}$
			0.8	60	42	0.0494	0.0019	0.0480	0.0004	0.0008	$< 10^{-4}$
		16	0.2	753	611	**0.0477**$^{(\dagger)}$	0.0018	0.0480	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.4	176	155	0.0499	0.0017	0.0500	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.8	48	44	0.0509	0.0019	0.0510	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
	2.0	8	0.2	3518	2843	0.0514	0.0019	0.0500	0.0058	0.0101	$< 10^{-4}$
			0.4	914	740	0.0493	0.0018	0.0490	0.0025	0.0030	$< 10^{-4}$
			0.8	223	178	0.0487	0.0018	0.0480	0.0073	0.0122	$< 10^{-4}$
		16	0.2	2634	2340	0.0490	0.0020	0.0480	$< 10^{-4}$	0.0001	$< 10^{-4}$
			0.4	753	611	0.0477	0.0018	0.0480	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
			0.8	176	155	0.0499	0.0017	0.0500	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
2	1.0	8	0.2	93	80	0.0492	0.0019	0.0480	0.1623	0.0375	0.0720
			0.4	25	23	0.0489	0.0018	0.0490	0.1322	0.0286	0.0850
			0.8	10	9	0.0485	0.0020	0.0490	0.0671	0.0100	0.0660
		16	0.2	84	79	0.0503	0.0018	0.0500	**0.1240**$^{(\ddagger)}$	0.0248	0.0740
			0.4	23	22	0.0498	0.0018	0.0500	**0.1111**$^{(\ddagger)}$	0.0169	0.0860
			0.8	16	16	0.0515	0.0018	0.0500	0.0007	0.0002	$< 10^{-4}$
	2.0	8	0.2	333	261	0.0515	0.0024	0.0510	0.2082	0.0459	0.1190
			0.4	93	80	0.0492	0.0019	0.0480	0.1623	0.0375	0.0720
			0.8	25	23	0.0489	0.0018	0.0490	0.1322	0.0286	0.0850
		16	0.2	318	289	0.0489	0.0018	0.0480	**0.1283**$^{(\ddagger)}$	0.0276	0.0900
			0.4	84	79	0.0503	0.0018	0.0500	**0.1240**$^{(\ddagger)}$	0.0248	0.0740
			0.8	23	22	0.0498	0.0018	0.0500	**0.1111**$^{(\ddagger)}$	0.0169	0.0860

$^{(*)}$ refers to the half-width of the 95% confidence interval.

$^{(\dagger)}$ 95% c.i. **does not** cover the nominal type I error, $\alpha = 0.05$.

$^{(\ddagger)}$ 95% c.i. **does** cover the nominal type II error, $\beta = 0.10$.

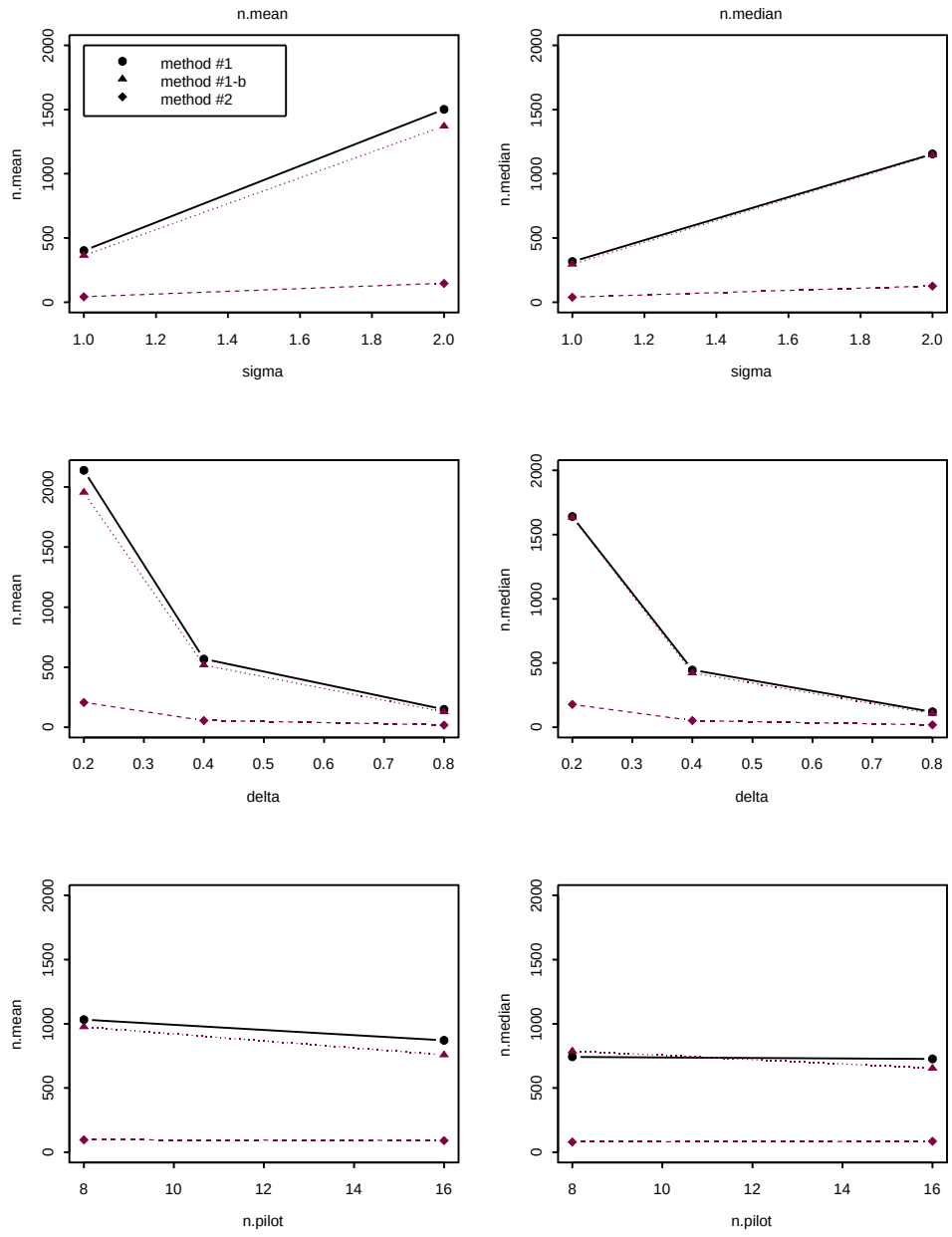

Figure 4: Mean and median sample sizes as functions of the simulation parameters.

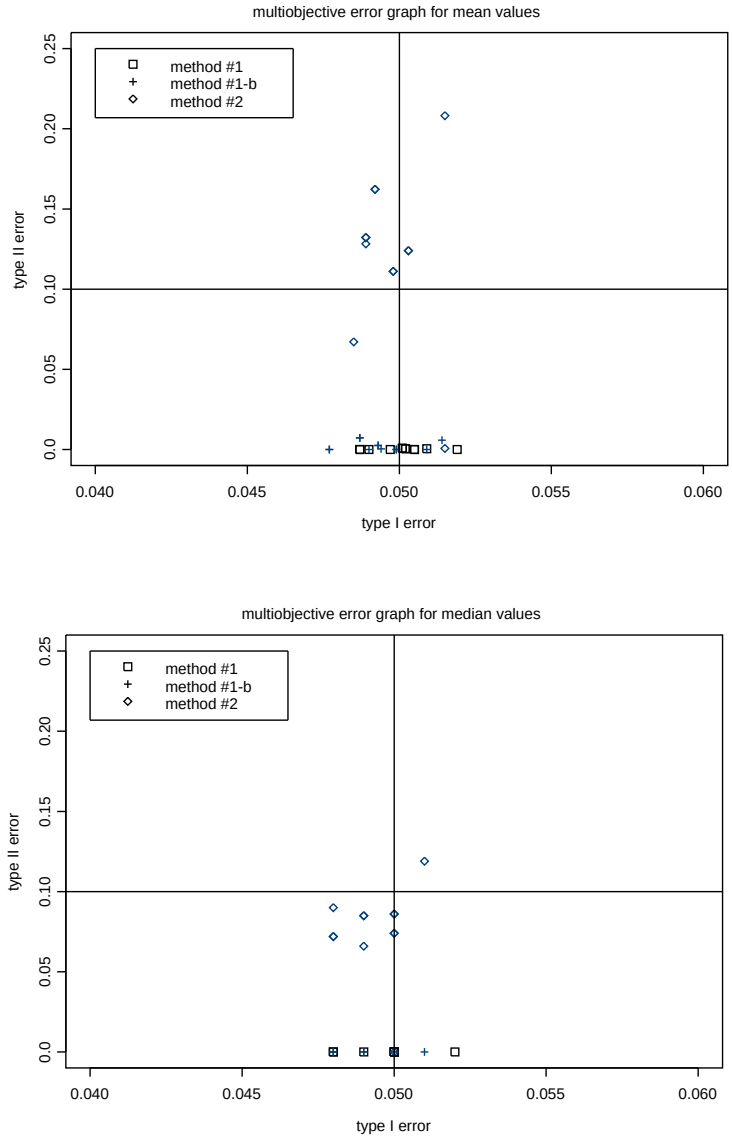

Figure 5: Type I *versus* type II error graphs.

5 Case study

In retail stores, handling of products typically forms the largest share of the operational costs (van Zelst et al., 2006). The handling activities are mainly driven by the shelf stacking process. In van Zelst et al. (2006), a study of

the shelf stacking process is presented. The data is collected by means of a motion and time study. In the experiment, for each item to be stacked, the time needed by a store clerk is measured. Empirical data on the stacking process in two grocery retail companies is collected.

The store clerks were followed during the shelf stacking with a camcorder. After the recording process, the times for stacking the items was digitized using a computerized time registration tool, resulting in an extensive database. The entire process resulted in over 70 hours of video tapes, which took a significant number of days to process into a useable digital format (van Zelst et al., 2006). Needless to say that the data collection process was a large investment for the companies involved.

Both retail chains use a different method for stacking the shelves (depending upon their strategy as a retailer). More specifically, Chain A uses more students and does allow for mirroring the items on the shelf (tidying the shelf such that all items are lined up in front of the shelf). Chain B however has full time employees who do the shelf stacking but does not allow for mirroring the items on the shelf. The main question is whether these different stacking methods lead to different stacking times and what size the sample should be in order to observe statistically significant differences.

Using 50 observations for each chain (available from the authors upon request), we obtained the sample sizes presented in Table 2. The nominal values were fixed at $\alpha = 0.05$ and $\beta = 0.10$, and δ at 0.2, 0.4, and 0.8. Note that similar to the simulation experiments, the lower sample sizes for method #2 (compared to method #1) should suffice to detect the differences with the given power.

Table 2: Case study sample sizes.

method	δ	n_{sample}
1	0.2	1,395,265
	0.4	348,818
	0.8	87,207
1-b	0.2	698,394
	0.4	174,601
	0.8	43,653
2	0.2	281
	0.4	73
	0.8	51

6 Conclusion and Final Remarks

Two methods were compared to determine the best sample size in a simple linear regression situation. The first method determines two sample sizes, one for each parameter of the model and takes its maximum or the average value as the proposed sample size. The second method uses the non-central $\mathcal{F}$ distribution in order to make a joint sample size calculation.

As expected the first method is conservative but the simulation results showed that this effect is too drastic. In general, the sample sizes are too large (the empirical type II error is basically zero in contrast to the nominal of 0.10). This fact holds for both sample size strategies of method #1. In the opposite direction the second method presents empirical values very close to the nominal ones. The simulation results show the effectivity of the second method. Moreover, in situations where collecting a sample is expensive or difficult to perform, a smaller sample resulting in the same insights as a larger sample is preferred. A practical case example is presented showing the merits of the new method.

Some questions remains unanswered as for instance whether or not the

proposed method would perform differently for different setting of the parameters and different distributions for x. These are only a few topics for future research in this area.

Acknowledgments

The research of prof. Enrico Colosimo has been partially funded by the CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*) of the Ministry for Science and Technology of Brazil, grant 300582/2003-0. The research of prof. Frederico Cruz has been partially funded by the CNPq, grants 201046/1994-6, 301809/1996-8, 307702/2004-9, 472066/2004-8, and 472877/2006-2, the FAPEMIG (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais*), grants CEX-289/98 and CEX-855/98, and PRPq-UFMG, grant 4081-UFMG/RTR/FUNDO/PRPq/99.

References

Burden, R. L. & Faires, J. D. (2005). *Numerical Analysis*, 8th edn, Thomson Learning, Belmont, CA.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, New Jersey: Lawrence Erlbaum.

Draper, N. & Smith (1998). *Applied Regression Analysis*, John wiley & Sons, New York.

Dupont, W. D. & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression, *Controlled Clinical Trials* **19**(6): 589–601.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*, Duxbury Press, North Scituate, MA.

Hintze, J. L. (1996). *PASS 6.0 User's Guide*, NCSS, Kaysville, UT.

Kraemer, H. C. & Thiemann, S. (1987). *How Many Subject? Statistical Power Analysis in Research*, Sage Publications, Inc., Newbury Park, CA.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility, *Biometrics* **45**: 255–268.

Linnet, K. (1990). Estimation of the linear relationship between the measurements of two methods with proportional errors, *Statistics in Medicine* **9**: 1463–1473.

S-PLUS (2000). *Programmers Guide*, Data Analysis Products Division, MathSoft, Seattle.

Seber, G. A. F. (1977). *Linear Regression Analysis*, John Wiley & Sons, New York.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*, 3th edn, W. H. Freeman and Co., New York.

van Zelst, S., van Donselaar, K., van Woensel, T., Broekmeulen, R. & Fransoo, J. (2006). Logistics drivers for shelf stacking in grocery retail stores: Potential for efficiency improvement, *International Journal of Production Economics* (in press).