Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 122 (2004) 27–46
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

On the harm that ignoring pretesting can cause

Dmitry Danilov[a], Jan R. Magnus[b,*]

[a] *Eurandom, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[b] *CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands*

Accepted 23 October 2003

Abstract

In econometrics the same data set is typically used to select the model and to estimate the parameters in the selected model. In applied econometrics practice, however, one typically acts as if the model had been given a priori, thus ignoring the fact that the estimators are in fact *pretest* estimators. Hence one assumes incorrectly that the estimator is unbiased, and that the reported variance, conditional on the selected model, is equal to its unconditional variance.

In this paper, we find the *unconditional* first and second moments of the pretest estimator (in fact, of a more general estimator, the WALS estimator), taking full account of the fact that model selection and estimation are an integrated procedure, and show that the error in not reporting the correct moments can be large. We also show that this error can vary substantially between different model selection procedures. Finally, we ask how the error increases when the number of auxiliary regressors increases.

JEL classification: C13; C51

Keywords: Pretest estimator; Model selection; Mean squared error

1. Introduction

In econometrics, but also in many other disciplines, the same data set is commonly used for model selection and for estimation. Standard statistical theory is therefore not directly applicable, since the properties of most estimators depend not only on the stochastic nature of the selected model, but also on the way the model was selected.

* Corresponding author. Tel.: +31-13-466-3092; fax: +31-13-466-3066.
E-mail addresses: danilov@eurandom.tue.nl (D. Danilov), magnus@uvt.nl (J.R. Magnus).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2003.10.018

The simplest example of this situation is the standard linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}$, where we are uncertain whether to include $\boldsymbol{z}$ or not. [1] The usual procedure is to compute the t-statistic of γ, and then, depending on whether $|t|$ is 'large' or 'small', decide to use the unrestricted or the restricted model. We then estimate $\boldsymbol{\beta}$ from the selected model. This estimator is a *pretest* estimator, but we commonly report its properties as if estimation had not been preceded by model selection. Thus we report no bias and an incorrect variance.

This is clearly wrong. Our view is *not* that we should avoid pretesting, even though it is well-known that pretest estimators have poor properties, inadmissibility being only one of them. This would be near-impossible in applied work. Our view is simply that we should report estimates of the correct bias and variance (or mean squared error), taking full account of the fact that model selection and estimation are an integrated procedure. This paper attempts to do this.

The literature on pretesting starts with Bancroft's (1944) famous article. Bancroft is mostly concerned with the bias introduced by pretests of homogeneity of variances and pretests of a regression coefficient. He considers the simplest case, in our notation $\boldsymbol{y} = \beta\boldsymbol{x} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}$ (one β, one γ), where he wishes to estimate β while being uncertain about whether $\boldsymbol{z}$ should be in the regression or not. He then investigates the bias of the pretest estimator of β. Mosteller (1948) considers the special case $\boldsymbol{x}' = (\boldsymbol{\iota}', \boldsymbol{\iota}')$, $\boldsymbol{z}' = (\boldsymbol{0}', \boldsymbol{\iota}')$, where $\boldsymbol{\iota}$ denotes the vector of ones. Thus, Mosteller considers pooling: if $\gamma = 0$ we pool, otherwise we do not pool. In this context, he calculates the mean squared error of the pretest estimator. Huntsberger (1955) extends Mosteller's paper by explicitly writing the pretest estimator as a (continuous) weighted average of the restricted ($\gamma = 0$) and unrestricted estimator, where the weights are functions of the relevant t-statistic. The fact that the pretest estimator has many undesirable properties is highlighted by Sclove et al. (1972). The early literature is discussed in detail in Judge and Bock's (1978) important monograph.

Lovell (1983) asks what will be the true significance level of a t-test after pretesting, and recommends a simple rule-of-thumb. Roehrig (1984) establishes the relationship between the mean squared error of the pretest estimator and the mean squared error of the estimator of the nuisance parameters, a result later generalized by Magnus and Durbin (1999). Mittelhammer (1984) compares the risk functions of several estimators (including the pretest) under model misspecification, and concludes *inter alia* that all alternatives to OLS can be inferior to OLS in terms of prediction risk. The literature of this period is well summarized in Judge and Bock (1983) and in the special issue of the *Journal of Econometrics* (1984), edited by George Judge.

Asymptotic aspects are considered in Sen (1979), Pötscher (1991), Zhang (1992), and Pötscher and Novak (1998). While most studies, including ours, are confined to the first two moments of the pretest statistics, Giles and Srivastava (1993) and Leeb and Pötscher (2003) derive the distribution of the traditional pretest estimator. Summaries of the latest developments are given in Giles and Giles (1993), Chatfield (1995), and Magnus (1999).

[1] We follow the notation proposed by Abadir and Magnus (2002).

Different model selection strategies are discussed by Hoover and Perez (1999), who favor the general-to-specific procedure. Hendry (2001) advertises computer-automated general-to-specific procedures and claims that these procedures perform well in Monte Carlo experiments.

In spite of all this literature, we are still far removed from having a fully integrated procedure of model selection and parameter estimation. The current paper attempts to narrow this gap. Our main tool is a generalization of the 'equivalence theorem' of Magnus and Durbin (1999). We derive the bias, variance, and mean squared error of the pretest estimator (in fact, of a more general estimator, the so-called WALS estimator), and estimate the error of not reporting the correct moments (the 'underreporting error'). This error can be very substantial. We also show that there can be large differences in underreporting between different model selection procedures. Finally, we investigate (in a special case) how the underreporting error increases when the number of auxiliary regressors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ increases.

The paper is organized as follows. We define the formal framework and the notation in Section 2, where we also provide a generalization of the 'equivalence theorem' of Magnus and Durbin (1999). This generalization forms the basis of the subsequent analysis. In Section 3 we discuss underreporting and its bounds. Section 4 discusses the simplest case, where there is only one auxiliary regressor $\boldsymbol{z}$, and hence only one possible pretest procedure (using the t-statistic). In Sections 5 and 6 we address the more difficult case where we have two auxiliary regressors. Then, there is no unique selection procedure. We show, *inter alia*, that there can be large differences between general-to-specific and specific-to-general model selection. Section 7 concludes. The appendix contains two useful results on constrained least squares.

2. The WALS estimator and the equivalence theorem

In this paper we will consider pretest-type estimators at two levels of generality. In the current section we consider only the most general type, the so-called WALS estimator, to be defined shortly. Our framework is the standard linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ $(n \times 1)$ is the vector of observations, $\boldsymbol{X}$ $(n \times k)$ and $\boldsymbol{Z}$ $(n \times m)$ are matrices of nonrandom regressors, $\boldsymbol{\varepsilon}$ $(n \times 1)$ is a random vector of unobservable disturbances, and $\boldsymbol{\beta}$ $(k \times 1)$ and $\boldsymbol{\gamma}$ $(m \times 1)$ are unknown nonrandom parameter vectors. We assume that $k \geqslant 1$, $m \geqslant 1$, $k + m \leqslant n - 1$, that the design matrix $(\boldsymbol{X} : \boldsymbol{Z})$ has full column-rank $k + m$, and that the disturbances $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. $\mathrm{N}(0, \sigma^2)$.

The reason for distinguishing between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is that $\boldsymbol{X}$ contains explanatory variables which we want in the model on theoretical or other grounds (irrespective of the found t-values of the β-parameters), while $\boldsymbol{Z}$ contains additional explanatory variables of which we are less certain. Our focus is the estimation of $\boldsymbol{\beta}$. Hence the only role for $\boldsymbol{Z}$ is to improve the estimation of $\boldsymbol{\beta}$, while $\boldsymbol{\gamma}$ is a vector of nuisance parameters. The columns of $\boldsymbol{X}$ are called 'focus' regressors, and the columns of $\boldsymbol{Z}$

'auxiliary' regressors. We define the matrices

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \quad \boldsymbol{Q} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}$$

and the scaled and normalized parameter vector $\boldsymbol{\eta} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{1/2}\boldsymbol{\gamma}/\sigma$. The matrix $\boldsymbol{Q}$ can be interpreted as the (scaled) correlation between $\boldsymbol{X}$ and $\boldsymbol{Z}$. Clearly, $\boldsymbol{Q} = \boldsymbol{0}$ if and only if $\boldsymbol{Z}$ is orthogonal to $\boldsymbol{X}$. The least-squares (LS) estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are

$$\boldsymbol{b}_u = \boldsymbol{b}_r - \boldsymbol{Q}\hat{\boldsymbol{\theta}}, \quad \hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{y},$$

where $\boldsymbol{b}_r = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\hat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{1/2}\hat{\boldsymbol{\gamma}}$. The subscripts '$u$' and '$r$' denote 'unrestricted' and 'restricted' (with $\boldsymbol{\gamma} = \boldsymbol{0}$), respectively. Letting $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}/\sigma$, we see that $\hat{\boldsymbol{\eta}} \sim \mathrm{N}(\boldsymbol{\eta}, \boldsymbol{I}_m)$. Notice that $\hat{\boldsymbol{\eta}}$ is only observable when σ is known, while $\hat{\boldsymbol{\theta}}$ is observable whether σ is known or not.

Magnus and Durbin (1999) considered the estimation of $\boldsymbol{\beta}$ in model (1) and proposed a weighted-average least-squares (WALS) estimator of $\boldsymbol{\beta}$ of the form $\boldsymbol{b} = \lambda\boldsymbol{b}_u + (1-\lambda)\boldsymbol{b}_r$, where $\lambda = \lambda(\hat{\boldsymbol{\theta}}, s_u^2)$ and s_u^2 denotes the estimator for σ^2 in the unrestricted model. This includes the usual pretest estimator as a special case, but only when one restricts the choice of model to the fully restricted and the fully unrestricted case. In the case $m = 1$, there are indeed two models: the unrestricted and the restricted (where $\gamma = 0$). But when $m = 2$, there are already four possible models: the unrestricted model, two partially restricted models (one of the two γ's is zero), and the restricted model (both γ's are zero). In general, there are 2^m models to consider. Thus we need a generalization of the WALS estimator, which includes as special cases not only the unrestricted estimator $\boldsymbol{b}_u$ and the restricted estimator $\boldsymbol{b}_r$ (where *all* γ's are set equal to zero), but also many or all intermediate estimators where *some* of the γ's are set equal to zero.

Let $\boldsymbol{S}_i$ be an $m \times r_i$ selection matrix of rank $r_i \geqslant 0$, so that $\boldsymbol{S}_i' = (\boldsymbol{I}_{r_i} : \boldsymbol{0})$ or a column-permutation thereof. Let $\mathscr{M}_i$ denote the linear model (1) under the restriction $\boldsymbol{S}_i'\boldsymbol{\gamma} = \boldsymbol{0}$, and denote LS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in model $\mathscr{M}_i$ by $\boldsymbol{b}_{(i)}$ and $\boldsymbol{c}_{(i)}$. In Lemma A1 in the appendix we show that the partially restricted estimator $\boldsymbol{b}_{(i)}$ can be written as $\boldsymbol{b}_{(i)} = \boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}_i\hat{\boldsymbol{\theta}}$, where $\boldsymbol{W}_i$ is an idempotent matrix, depending only on $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{S}_i$. That is, $\boldsymbol{b}_{(i)}$ is a linear function of the two *independent* vectors $\boldsymbol{b}_r$ and $\hat{\boldsymbol{\theta}}$.[2] Moreover, $\boldsymbol{c}_{(i)}$ is a linear function of $\hat{\boldsymbol{\theta}}$ only and hence independent of $\boldsymbol{b}_r$.

We now define the WALS estimator of $\boldsymbol{\beta}$ as $\boldsymbol{b} = \sum_i \lambda_i \boldsymbol{b}_{(i)}$, where the sum is taken over all 2^m different models obtained by setting a subset of the γ's equal to zero, and the λ_i are weight-functions satisfying certain minimal regularity conditions, namely

$$\lambda_i \geqslant 0, \quad \sum_{i=1}^{2^m} \lambda_i = 1, \quad \lambda_i = \lambda_i(\boldsymbol{M}\boldsymbol{y}).$$

The WALS estimator can then be written as $\boldsymbol{b} = \boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}\hat{\boldsymbol{\theta}}$, where $\boldsymbol{W} = \sum_i \lambda_i \boldsymbol{W}_i$. Notice that, while $\boldsymbol{W}_i$ is nonrandom, $\boldsymbol{W}$ is random.

[2] The independence of $\boldsymbol{b}_r$ and $\hat{\boldsymbol{\theta}}$ follows immediately from the fact that $\boldsymbol{X}'\boldsymbol{y}$ and $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{y}$ are independent. In fact, even if the observations $y_1, \ldots, y_n$ are not normal and the data-generating process is unknown, $\boldsymbol{b}_r$ and $\hat{\boldsymbol{\theta}}$ will still be uncorrelated, as long as the $\{y_i\}$ are uncorrelated with constant variance (Leeb and Pötscher (2003), Lemma A.1).

A few words about the regularity conditions are in order. If σ^2 is known, then most or all pretest procedures will use statistics (such as t- and F-statistics) which depend on $\hat{\boldsymbol{\theta}}$ only. If σ^2 is not known and estimated by s_u^2, then all t- and F-statistics will depend on $(\hat{\boldsymbol{\theta}}, s_u^2)$. Now, it is a basic result in least-squares theory that s_u^2 is independent of $(\boldsymbol{b}_u, \hat{\boldsymbol{\gamma}})$. It follows that $\boldsymbol{b}_r$ is independent of s_u^2. Hence, $\boldsymbol{b}_r$ will be independent of $(\hat{\boldsymbol{\theta}}, s_u^2)$. Finally, if σ^2 is not known and estimated by $s_{(i)}^2$ (the estimator of σ^2 in model $\mathscr{M}_i$), then it is no longer true that all t- and F-statistics depend only on $(\hat{\boldsymbol{\theta}}, s_u^2)$. However, they still depend only on $\boldsymbol{My}$, because Lemma A1 implies that both $\boldsymbol{c}_{(i)}$ and the residuals $\boldsymbol{e}_{(i)}$ from model $\mathscr{M}_i$ are linear functions of $\boldsymbol{My}$. We conclude that the regularity conditions on λ_i are reasonable and mild.[3]

Theorem 1 (Equivalence Theorem, generalized). *If the regularity conditions on λ_i are satisfied, then*

$$\mathrm{E}(\boldsymbol{b}) = \boldsymbol{\beta} - \sigma\boldsymbol{Q}\mathrm{E}(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \quad \mathrm{var}(\boldsymbol{b}) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\mathrm{var}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}')$$

and hence

$$\mathrm{MSE}(\boldsymbol{b}) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}').$$

Proof. Since $\boldsymbol{b}_r$ and $\boldsymbol{My}$ are independent, we have

$$\mathrm{E}(\boldsymbol{b}_r \mid \boldsymbol{My}) = \mathrm{E}(\boldsymbol{b}_r), \quad \mathrm{var}(\boldsymbol{b}_r \mid \boldsymbol{My}) = \mathrm{var}(\boldsymbol{b}_r).$$

Hence,

$$\begin{aligned}\mathrm{E}(\boldsymbol{b} \mid \boldsymbol{My}) &= \mathrm{E}(\boldsymbol{b}_r \mid \boldsymbol{My}) - \boldsymbol{Q}\mathrm{E}(\boldsymbol{W}\hat{\boldsymbol{\theta}} \mid \boldsymbol{My}) \\ &= \mathrm{E}(\boldsymbol{b}_r) - \sigma\boldsymbol{Q}\boldsymbol{W}\hat{\boldsymbol{\eta}} = \boldsymbol{\beta} - \sigma\boldsymbol{Q}(\boldsymbol{W}\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\end{aligned}$$

and

$$\mathrm{var}(\boldsymbol{b} \mid \boldsymbol{My}) = \mathrm{var}(\boldsymbol{b}_r \mid \boldsymbol{My}) = \mathrm{var}(\boldsymbol{b}_r) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

The unconditional mean and variance of $\boldsymbol{b}$ and hence its mean squared error follow. □

Theorem 1 provides a nontrivial generalization, using a simpler proof, of Theorem 2 in Magnus and Durbin (1999). Apparently, the properties of the complicated WALS estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ depend critically on the properties of the less complicated estimator $\boldsymbol{W}\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$.[4] We notice that neither the bias, nor the variance or the mean squared error of $\boldsymbol{b}$ depend on $\boldsymbol{\beta}$. They do, however, depend on $\boldsymbol{\gamma}$ or, more accurately, on $\boldsymbol{\eta}$.

Theorem 1 allows us to obtain bounds for the mean squared error. If we let $\xi_{\min}$ and $\xi_{\max}$ denote the smallest and largest eigenvalue of $\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})$, respectively, then

$$\sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \xi_{\min}\boldsymbol{Q}\boldsymbol{Q}') \leqslant \mathrm{MSE}(\boldsymbol{b}) \leqslant \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \xi_{\max}\boldsymbol{Q}\boldsymbol{Q}').$$

[3] The regularity conditions allow not only all standard pretest procedures, but also inequality-constrained least squares. Thus, Theorem 1 below explains the 'surprising symmetry' found by Thomson and Schmidt (1982, p. 176).

[4] For the pretest estimators studied in Giles and Srivastava (1993) and Leeb and Pötscher (2003) the MSE can also be obtained from the finite-sample densities derived in these papers.

For the traditional pretest estimator we shall see later (Fig. 1) that, for the case $m = 1$, the smallest and largest eigenvalues are 0.28 and 2.46, respectively. For the case $m \geqslant 2$, however, the bounds can become infinite, depending on the pretest procedure. In this paper our interest is not in the bounds for the true MSE, but in the difference between the true MSE and the reported MSE.

It is also interesting to compare the WALS estimator $\boldsymbol{b}$ (model selection and pretesting properly taken into account) with the unrestricted estimator $\boldsymbol{b}_u$ (no model selection, always use the unrestricted model). Since, $\mathrm{MSE}(\boldsymbol{b}_u) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\boldsymbol{Q}')$, we find that

$$\mathrm{MSE}(\boldsymbol{b}_u) - \mathrm{MSE}(\boldsymbol{b}) = \sigma^2(\boldsymbol{Q}(\boldsymbol{I}_m - \mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}')).$$

Hence, in theory it is possible that $\mathrm{MSE}(\boldsymbol{b})$ is uniformly (that is, for all $\boldsymbol{\eta}$) larger (or smaller) than $\mathrm{MSE}(\boldsymbol{b}_u)$. This happens if all eigenvalues of $\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})$ are larger (smaller) than one. This situation is unlikely to occur in practical situations. In the traditional pretest context, it does not occur for $m = 1$ (see the discussion after Fig. 1) and $m = 2$ (checked numerically, but not reported in the paper).

3. Pretesting and underreporting

While we considered the general WALS estimator in the previous section, we now consider the less general (traditional) pretest estimator. In the idealized context of the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, we define a *pretest procedure* as a two-step procedure. In step 1 we select the model. In step 2 we estimate the unknown parameters $\boldsymbol{\beta}$ (and σ^2) from the selected model. This yields the pretest estimators $\boldsymbol{b}$ (and s^2). In a pretest procedure thus defined, all λ_i are zero except one which is one. We require (as in Theorem 1) that the model selection criterion depends on $\boldsymbol{y}$ only through $\boldsymbol{M}\boldsymbol{y}$, the residuals from the restricted model. From here on we shall also assume that σ^2 is known (but see our comments in the conclusion).

The mean squared error of the pretest estimator $\boldsymbol{b}$ is, according to Theorem 1,

$$\mathrm{MSE}(\boldsymbol{b}) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}').$$

In applied econometrics practice the same estimator $\boldsymbol{b}$ is selected, but the effects of pretesting are ignored, the reported bias is zero, and hence the reported MSE equals the reported variance. The reported MSE equals

$$\widetilde{\mathrm{MSE}}(\boldsymbol{b}) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}'),$$

according to Lemma A1, since $\boldsymbol{W} = \boldsymbol{W}_i$ if the ith model is selected, and σ^2 is assumed known. Notice that $\widetilde{\mathrm{MSE}}(\boldsymbol{b})$ is random since $\boldsymbol{W}$ is random. Let $\boldsymbol{\omega}'\boldsymbol{\beta}$ be our focus parameter, where $\boldsymbol{\omega}$ is an arbitrary nonzero $k \times 1$ vector. In order to compare

$$\mathrm{MSE}(\boldsymbol{\omega}'\boldsymbol{b}) = \sigma^2(\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} + \boldsymbol{\omega}'\boldsymbol{Q}\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})\boldsymbol{Q}'\boldsymbol{\omega}) \tag{2}$$

with

$$\widetilde{\mathrm{MSE}}(\boldsymbol{\omega}'\boldsymbol{b}) = \sigma^2(\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega} + \boldsymbol{\omega}'\boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}'\boldsymbol{\omega}), \tag{3}$$

we define the *underreporting ratio* UR as one minus the ratio of (3) and (2). Thus,

$$\mathrm{UR} = 1 - \frac{\widetilde{\mathrm{MSE}}(\boldsymbol{\omega}'\boldsymbol{b})}{\mathrm{MSE}(\boldsymbol{\omega}'\boldsymbol{b})} = \frac{\boldsymbol{q}'(\boldsymbol{R}-\boldsymbol{W})\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{R}\boldsymbol{q} + (1/q_0^2)}, \tag{4}$$

where

$$\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\eta}) = \mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}}), \quad \boldsymbol{q} = \frac{\boldsymbol{Q}'\boldsymbol{\omega}}{\sqrt{\boldsymbol{\omega}'\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{\omega}}}, \quad q_0^2 = \frac{\boldsymbol{\omega}'\boldsymbol{Q}\boldsymbol{Q}'\boldsymbol{\omega}}{\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega}}.$$

Notice that $\boldsymbol{q}'\boldsymbol{q} = 1$. The UR is a random variable, since it depends on $\boldsymbol{W}$, which depends on $\hat{\boldsymbol{\eta}}$. Both the UR and its expectation are unobservable, since they depend on $\boldsymbol{\eta}$ via $\boldsymbol{R}(\boldsymbol{\eta})$.

One would expect that the matrix $\mathrm{MSE}(\boldsymbol{b})$ is at least as large as the matrix $\mathrm{E}(\widetilde{\mathrm{MSE}}(\boldsymbol{b}))$ (in the sense that their difference is positive semidefinite), because pretesting introduces additional noise which is ignored in the reported MSE. Since $\mathrm{E}(\boldsymbol{W}) = \sum_i (\mathrm{E}\lambda_i)\boldsymbol{W}_i$ and

$$\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}}) = \sum_{i=1}^{2^m} \mathrm{E}(\lambda_i(\boldsymbol{W}_i\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\boldsymbol{W}_i\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})'),$$

this is guaranteed if the matrix

$$\sum_{i=1}^{2^m} \mathrm{E}(\lambda_i((\boldsymbol{W}_i\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})(\boldsymbol{W}_i\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})' - \boldsymbol{W}_i)) \tag{5}$$

is positive semidefinite. We will see in Section 4 that it is possible to devise pretest procedures which do not satisfy this requirement. Such procedures, however, tend to be rather silly. We will say that a pretest procedure is *viable* if the matrix in (5) is positive semidefinite over the whole parameter space.[5] For any viable pretest procedure, $\mathrm{E}(\mathrm{UR})$ is a number between zero and one. When q_0^2 (known to the investigator) is zero, then there is no underreporting: $\mathrm{E}(\mathrm{UR}) = 0$.[6] But when q_0^2 is large, $\mathrm{E}(\mathrm{UR})$ can be close to one.

The $m \times m$ matrix $\mathrm{E}(\boldsymbol{W})$ is a weighted average of idempotent matrices, and hence is bounded: all its elements are $\leqslant 1$ in absolute value, and all its diagonal elements (and all its eigenvalues) lie in the interval $[0, 1]$. In fact,

$$0 \leqslant \pi_u \leqslant \xi_j(\mathrm{E}\boldsymbol{W}) \leqslant 1 - \pi_r \leqslant 1 \quad (j = 1, \ldots, m),$$

where $\xi_j(\boldsymbol{A})$ denotes the jth eigenvalue of $\boldsymbol{A}$, π_u is the probability of choosing the unrestricted model ($\boldsymbol{P}_i = \boldsymbol{0}$), and π_r the probability of choosing the restricted model ($\boldsymbol{P}_i = \boldsymbol{I}_m$).

The $\mathrm{E}(\mathrm{UR})$ is a function of $\boldsymbol{q}$ (normalized by $\boldsymbol{q}'\boldsymbol{q} = 1$), q_0^2, $\boldsymbol{\eta}$, and $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ (and m). Maximizing over $\boldsymbol{q}$ gives the inequality

$$\mathrm{E}(\mathrm{UR}) \leqslant q_0^2 \max_{1 \leqslant j \leqslant m} \xi_j((\boldsymbol{I}_m + q_0^2\boldsymbol{R})^{-1/2}(\boldsymbol{R} - \mathrm{E}\boldsymbol{W})(\boldsymbol{I}_m + q_0^2\boldsymbol{R})^{-1/2}).$$

[5] The term 'underreporting ratio' might therefore be criticized since it can become negative. In the class of viable procedures which we study, the expected ratio is always positive.

[6] This happens when $\boldsymbol{X}'\boldsymbol{Z} = \boldsymbol{0}$, but also (more generally and less trivially) when $\boldsymbol{Q}'\boldsymbol{\omega} = \boldsymbol{0}$. In either case $\boldsymbol{b} = \boldsymbol{b}_r$ whatever pretesting we do.

Then, letting $\mathrm{E}^*(\mathrm{UR}) := \max_{q,q_0^2} \mathrm{E}(\mathrm{UR})$, we find, as $q_0^2 \to \infty$,

$$\mathrm{E}^*(\mathrm{UR}) = 1 - \min_{1 \leqslant j \leqslant m} \xi_j(\boldsymbol{R}^{-1/2}(\mathrm{E}\boldsymbol{W})\boldsymbol{R}^{-1/2}) \leqslant 1 - \frac{\pi_u}{\max_j \xi_j(\boldsymbol{R})}, \tag{6}$$

which depends on $\boldsymbol{\eta}$ and $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ (and m). We see from (6) that the expected UR can be arbitrarily close to 1 if the mean squared error $\boldsymbol{R}$ fails to be bounded in $\boldsymbol{\eta}$. This cannot happen when $m = 1$ (unless we choose the restricted model with probability $\pi_r \geqslant \bar{\pi}_r > 0$, where $\bar{\pi}_r$ is constant, and thus does not depend on the observed t-value), but it can happen when $m \geqslant 2$, as we will see in Section 6.

Finally, since E(UR) depends on $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$, we briefly consider the role of this matrix. Without loss of generality, we may scale all $\boldsymbol{z}$-variables so that $\boldsymbol{z}_j'\boldsymbol{M}\boldsymbol{z}_j = 1$ for all $j = 1, \ldots, m$. In the special case where we can choose the $\boldsymbol{z}$-variables to be 'orthogonal' (in the sense that $\boldsymbol{M}\boldsymbol{z}_i$ and $\boldsymbol{M}\boldsymbol{z}_j$ are orthogonal for every $i \neq j$), we have $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$, and major simplifications occur.

Theorem 2. *Assume that $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$ and that σ^2 is known.*[7] *Then*, (a) *all models which include $\boldsymbol{z}_j$ as a regressor yield the same t-statistic of γ_j*; *and* (b) *suppose we select $\boldsymbol{z}_j$ if and only if the t-statistic $\hat{\eta}_j$ is significant in the sense that $|\hat{\eta}_j| > c$ for some given $c > 0$ (such as $c = 1.96$). Then, letting $\lambda(x) = 1$ if $|x| > c$ and 0 otherwise, $\boldsymbol{W}$ is a diagonal matrix with $w_{jj} = \lambda(\hat{\eta}_j)$, and $\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}}) = \boldsymbol{V} + \boldsymbol{d}\boldsymbol{d}'$, where $\boldsymbol{V}$ is a diagonal $m \times m$ matrix and $\boldsymbol{d}$ an $m \times 1$ vector with typical elements*

$$v_{jj} = \mathrm{var}(\lambda(\hat{\eta}_j)\hat{\eta}_j), \quad d_j = \mathrm{E}(\lambda(\hat{\eta}_j)\hat{\eta}_j - \eta_j).$$

Proof. Lemma A2 directly implies (a). To prove (b) we note that $\boldsymbol{W}$ is diagonal, because all $\boldsymbol{W}_i$ are diagonal. Its jth diagonal element w_{jj} is either 0 (if $\boldsymbol{z}_j$ is excluded from the model) or 1 (if $\boldsymbol{z}_j$ is included), that is, $w_{jj} = \lambda(\hat{\eta}_j)$. Also, the components of $\boldsymbol{W}\hat{\boldsymbol{\eta}}$ are independent of each other, and hence the variance matrix is diagonal. □

Since we will see that the choice of model selection procedure may matter a lot for the properties of the estimated focus parameters, it is advisable—if at all possible—to choose the auxiliary regressors such that $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$. In the most common cases, this will make the pretest estimator independent of the chosen model selection procedure. It also allows us to obtain explicit analytical expressions for the moments of the estimator, and it guarantees bounded risk for any value of m. This is important, because, in general, risk is not necessarily bounded when $m \geqslant 2$; see Section 6.

4. Underreporting: one nuisance parameter

In the case of one nuisance parameter, the model becomes $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{z} + \boldsymbol{\varepsilon}$, where the nuisance parameter γ is a scalar. We have only two models to compare: the unrestricted

[7] This is too restrictive. If σ^2 is always estimated by s_u^2, the LS estimator in the unrestricted model, then the theorem still holds. In more general cases, an approximate version of the theorem holds.

($\boldsymbol{W}_1 = 1$, $\boldsymbol{b}_{(1)} = \boldsymbol{b}_u$, $\lambda_1 = \lambda$) and the restricted ($\boldsymbol{W}_2 = 0$, $\boldsymbol{b}_{(2)} = \boldsymbol{b}_r$, $\lambda_2 = 1 - \lambda$). As a result we find

$$\boldsymbol{b} = \lambda \boldsymbol{b}_u + (1 - \lambda)\boldsymbol{b}_r, \quad \boldsymbol{W} = \lambda,$$

and

$$\mathrm{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}}) = \mathrm{MSE}(\lambda\hat{\eta}) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \quad \mathrm{E}(\boldsymbol{W}) = \mathrm{E}(\lambda).$$

The underreporting ratio is thus

$$\mathrm{UR}(\hat{\eta}, \eta) = \frac{R(\eta) - \lambda(\hat{\eta})}{R(\eta) + (1/q_0^2)},$$

where $\lambda(\hat{\eta}) = 1$ if $|\hat{\eta}| > c$ for some $c > 0$, and 0 otherwise, and

$$R(\eta) = \mathrm{E}(\lambda\hat{\eta} - \eta)^2, \quad q_0^2 = \frac{(\boldsymbol{z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega})^2}{(\boldsymbol{z}'\boldsymbol{M}\boldsymbol{z})(\boldsymbol{\omega}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\omega})}.$$

Assuming again that σ^2 is known and that c is given (say, $c = 1.96$), the λ-function depends only on $\hat{\eta}$, R depends only on η, and hence the UR depends on q_0^2 and $\hat{\eta}$ (both known to the investigator), and η (unknown).

It is easy to see that the larger is $R(\eta)$, the larger is UR. The random variable $\lambda\hat{\eta}$, considered as an estimator of η, thus plays a crucial role in determining the amount of underreporting. We consider its squared bias, variance and MSE in Fig. 1.[8]

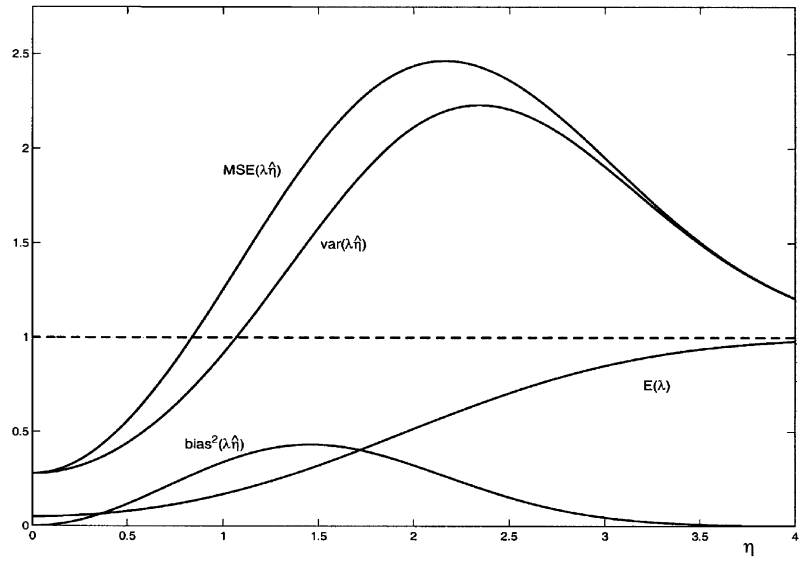

Fig. 1. Moments of $\lambda\hat{\eta}$ and λ compared ($m = 1$, $c = 1.96$).

[8] All results reported in Figs. 1 and 2 are based on exact analytical formulas, with the exception of the locus curve in Fig. 2 which was obtained numerically (in Matlab 5.3) using the Nelder-Mead simplex method.

The bias of $\lambda\hat{\eta}$ is negative for $\eta > 0$ and reaches its minimum -0.66 at $\eta = 1.46$. The variance reaches its minimum 0.28 at $\eta = 0$ and its maximum 2.23 at $\eta = 2.34$. The MSE $R(\eta)$ is shaped similarly to the variance. It reaches its minimum at $\eta = 0$ and its maximum 2.46 at $\eta = 2.16$. The variance of $\lambda\hat{\eta}$ is large relative to its bias, suggesting that variance-reduction is more important than bias-reduction. We see that $\mathrm{MSE}(\lambda\hat{\eta}) < 1$ if and only if $|\eta| < 0.84$. This means that the unrestricted estimator $\boldsymbol{b}_u$ is better (has lower mean squared error) than the pretest estimator $\boldsymbol{b}$ as soon as $|\eta| > 0.84$.

We also graph the expectation of the reported MSE of $\lambda\hat{\eta}$, that is, $\mathrm{E}(\lambda)$, as a function of η for $c = 1.96$, and the MSE of the unrestricted estimator of η, that is $\mathrm{MSE}(\hat{\eta})$ (the dashed line, constant at 1). Since λ only takes the values 0 and 1, $\mathrm{E}(\lambda)$ denotes the probability of choosing the unrestricted model ($\lambda = 1$). But λ also denotes the reported variance (MSE). We see that $\mathrm{E}(\lambda) \equiv \Pr(|\hat{\eta}| > c)$ increases monotonically between 0.05 at $\eta = 0$ and 1 at $\eta = \infty$. Since $\mathrm{MSE}(\lambda\hat{\eta}) > \mathrm{E}(\lambda)$, the pretest procedure is viable.[9]

In Fig. 2 we graph the expected underreporting ratio E(UR) for five different values of q_0^2: 0, 0.1, 1, 10, and ∞. At $q_0^2 = 0$ there is no underreporting and $\mathrm{E(UR)} = 0$. At $q_0^2 = 1$, we see that E(UR) is 0.18 at $\eta = 0$, reaches a maximum 0.57 at $\eta = 1.73$, and varies substantially with η (from 0 to 0.57), implying that on average the pretest MSE can be 2.3 times the reported MSE ($1/(1-0.57) = 2.3$). At $q_0^2 = \infty$, E(UR) is large; the maximum occurs at $\eta = 0.82$ where $\mathrm{E(UR)} = 0.87$, so that the reported variance should be multiplied by about 7.5 in order to obtain the true MSE of the pretest estimator.

Clearly, the maximum of E(UR) depends on q_0. This dependence is graphed as the dashed line in Fig. 2. We conclude that the effect of not reporting the true bias and variance of the pretest estimator can lead to serious misrepresentation of the results, *even in the case* $m = 1$. The larger is q_0^2 (known to the investigator), the larger will be the expected UR.

5. Model selection: general-to-specific and specific-to-general

When $m = 1$ pretesting is simple: look at the t-statistic for γ in the unrestricted model. If $|t| > c$, choose the unrestricted model (leading to $\boldsymbol{b}_u$); otherwise choose the restricted model (leading to $\boldsymbol{b}_r$). When $m > 1$ there are many ways to pretest. We consider the case $m = 2$ under the following conditions: model selection is based on t-statistics only, in the selected model all t-statistics are 'significant' (as defined in Theorem 2), and σ^2 is known.

Without loss of generality we normalize $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, the regressors associated with the nuisance parameters γ_1 and γ_2, by setting $\boldsymbol{z}_i'\boldsymbol{M}\boldsymbol{z}_i = 1$ for $i = 1, 2$. Then,

$$\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

[9] However, not all λ-functions lead to a viable procedure. For example, the—admittedly silly—procedure defined by $\lambda = 1$ if $|\hat{\eta}| \leqslant c$ and 0 otherwise is not viable, since $\mathrm{MSE}(\lambda\hat{\eta}) < \mathrm{E}(\lambda)$ at $\eta = 0$ for any $c > 0$.

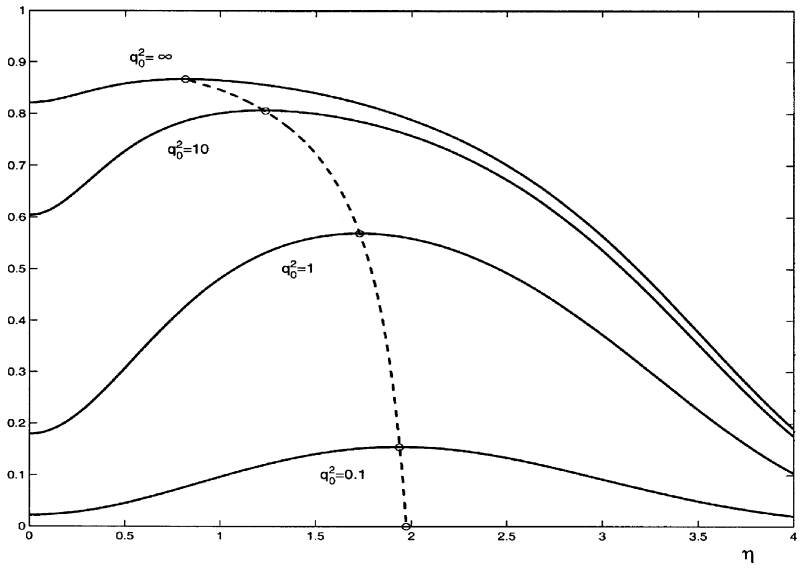

Fig. 2. E(UR) and locus of max(E(UR)) ($m = 1$, $c = 1.96$).

where $|r| < 1$, and

$$(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2} = \frac{1}{\sqrt{1-r^2}} \begin{pmatrix} \alpha & -\rho \\ -\rho & \alpha \end{pmatrix},$$

with

$$\alpha = \frac{\sqrt{1+r}+\sqrt{1-r}}{2}, \quad \rho = \frac{\sqrt{1+r}-\sqrt{1-r}}{2}.$$

There are four t-statistics to consider: two in the unrestricted model (denoted t_1 and t_2), one in the model where $\gamma_2 = 0$ (denoted $t_{(1)}$), and one in the model where $\gamma_1 = 0$ (denoted $t_{(2)}$). Let $\hat{\eta}_1$ and $\hat{\eta}_2$ denote the components of $\hat{\boldsymbol{\eta}}$. Then, each of the four t-statistics is a linear function of $\hat{\eta}_1$ and $\hat{\eta}_2$ in accordance with Lemma A1:

$$t_1 = \alpha\hat{\eta}_1 - \rho\hat{\eta}_2, \quad t_2 = -\rho\hat{\eta}_1 + \alpha\hat{\eta}_2$$

and

$$t_{(1)} = \alpha\hat{\eta}_1 + \rho\hat{\eta}_2, \quad t_{(2)} = \rho\hat{\eta}_1 + \alpha\hat{\eta}_2.$$

Of course, since $\alpha^2 + \rho^2 = 1$, all four t-statistics are normally distributed with unit variance and, under the appropriate null hypothesis, mean zero. Also, $t_{(1)}$ is independent of t_2 and $t_{(2)}$ is independent of t_1, for the same reason that $\boldsymbol{b}_r$ and $\hat{\boldsymbol{\eta}}$ are independent. Further,

$$\operatorname{corr}(t_1, t_{(1)}) = \operatorname{corr}(t_2, t_{(2)}) = \sqrt{1-r^2} > 0$$

and

$$\operatorname{corr}(t_1, t_2) = -r, \quad \operatorname{corr}(t_{(1)}, t_{(2)}) = r.$$

Finally,

$$|t_1| > |t_2| \Leftrightarrow |t_{(1)}| > |t_{(2)}| \Leftrightarrow |\hat{\eta}_1| > |\hat{\eta}_2|.$$

We will investigate two pretest procedures that are in common use: 'general-to-specific' and 'specific-to-general'. Let $\mathscr{M}_0$ denote the restricted model, $\mathscr{M}_1$ the model with only z_1 ($\gamma_2 = 0$), $\mathscr{M}_2$ the model with only z_2 ($\gamma_1 = 0$), and $\mathscr{M}_{12}$ the unrestricted model. Then we define the general-to-specific (or 'backward' or 'top–down') procedure as follows: (a) Estimate the unrestricted model $\mathscr{M}_{12}$. This yields t-statistics t_1 and t_2, (b) choose $\mathscr{M}_{12}$ if both t_1 and t_2 are significant, (c) otherwise, (i) if $|t_1| > |t_2|$ estimate $\mathscr{M}_1$, yielding $t_{(1)}$. If $t_{(1)}$ is significant choose $\mathscr{M}_1$, otherwise choose $\mathscr{M}_0$; (ii) if $|t_1| \leqslant |t_2|$ estimate $\mathscr{M}_2$, yielding $t_{(2)}$. If $t_{(2)}$ is significant choose $\mathscr{M}_2$, otherwise choose $\mathscr{M}_0$.

Similarly, we define the specific-to-general (or 'forward' or 'bottom–up') procedure as follows: (a) Estimate both partially restricted models $\mathscr{M}_1$ and $\mathscr{M}_2$. This yields t-statistics $t_{(1)}$ and $t_{(2)}$; (b) choose $\mathscr{M}_0$ if neither $t_{(1)}$ nor $t_{(2)}$ is significant; (c) otherwise, estimate the unrestricted model yielding t_1 and t_2, and choose $\mathscr{M}_{12}$ if t_1 and t_2 are both significant; (d) in all other cases choose $\mathscr{M}_1$ (if $|t_{(1)}| > |t_{(2)}|$) or $\mathscr{M}_2$ (if $|t_{(1)}| \leqslant |t_{(2)}|$).

Since the two cases ($|t_{(1)}| \leqslant c < |t_1|$, $|t_2| \leqslant c < |t_{(2)}|$) and ($|t_{(2)}| \leqslant c < |t_2|$, $|t_1| \leqslant c < |t_{(1)}|$) cannot occur, we see that both procedures are identical, except for the case where t_1 and t_2 are both significant, while $t_{(1)}$ and $t_{(2)}$ are both not significant. In that case, the general-to-specific procedure chooses the unrestricted model and the specific-to-general procedure chooses the restricted model. In the special case $r = 0$, we find $t_1 = t_{(1)} = \hat{\eta}_1$ and $t_2 = t_{(2)} = \hat{\eta}_2$, and the two pretest procedures coincide. When $|r| \to 1$, the difference between the two procedures is at its largest. In spite of the seemingly small difference between the two pretest procedures, the effect of pretesting on underreporting will be surprisingly different for the two procedures.

6. Underreporting: two nuisance parameters

In the case $m = 1$ the expected underreporting ratio E(UR) depends (for fixed c) on two parameters: q_0^2 (known to the investigator) and η (unknown). In the case $m = 2$, E(UR) depends, after normalization, on five parameters: q_0^2, q_1 (the first component of q) and r (known), and η_1 and η_2 (unknown). In addition, E(UR) depends on the procedure.

We have four models to compare: the unrestricted $\mathscr{M}_{12}$, the partially restricted $\mathscr{M}_1$ ($\gamma_2 = 0$) and $\mathscr{M}_2$ ($\gamma_1 = 0$), and the restricted $\mathscr{M}_0$ ($\gamma_1 = \gamma_2 = 0$). This implies selection matrices $\boldsymbol{S}_0 = \boldsymbol{I}_2$, $\boldsymbol{S}_1 = (0,1)'$, and $\boldsymbol{S}_2 = (1,0)'$ (the matrix $\boldsymbol{S}_{12}$ has no columns), and hence $\boldsymbol{W}_0 = \boldsymbol{0}$, $\boldsymbol{W}_{12} = \boldsymbol{I}_2$,

$$\boldsymbol{W}_1 = \frac{1}{2}\begin{pmatrix} 1+\sqrt{1-r^2} & r \\ r & 1-\sqrt{1-r^2} \end{pmatrix}$$

and

$$\boldsymbol{W}_2 = \frac{1}{2}\begin{pmatrix} 1-\sqrt{1-r^2} & r \\ r & 1+\sqrt{1-r^2} \end{pmatrix}.$$

Since $\boldsymbol{W} = \lambda_0 \boldsymbol{W}_0 + \lambda_1 \boldsymbol{W}_1 + \lambda_2 \boldsymbol{W}_2 + \lambda_{12} \boldsymbol{W}_{12}$, we thus find

$$\boldsymbol{W} = \frac{1}{2} \begin{pmatrix} \operatorname{tr}(\boldsymbol{W}) + \sqrt{1-r^2}(\lambda_1 - \lambda_2) & r(\lambda_1 + \lambda_2) \\ r(\lambda_1 + \lambda_2) & \operatorname{tr}(\boldsymbol{W}) - \sqrt{1-r^2}(\lambda_1 - \lambda_2) \end{pmatrix},$$

where $\operatorname{tr}(\boldsymbol{W}) = \lambda_1 + \lambda_2 + 2\lambda_{12}$. As before, let $\lambda(x) = 1$ if $|x| > c$ and 0 otherwise. Then,

$$\lambda_0 = (1 - \lambda(t_{(1)}))(1 - \lambda(t_{(2)})) - \delta B_1, \quad \lambda_1 = \lambda(t_{(1)})(1 - \lambda(t_2)) - (1 - \mu)B_2,$$

$$\lambda_2 = \lambda(t_{(2)})(1 - \lambda(t_1)) - \mu B_2, \quad \lambda_{12} = \lambda(t_1)\lambda(t_2) - (1 - \delta)B_1,$$

with

$$B_1 = \lambda(t_1)\lambda(t_2)(1 - \lambda(t_{(1)}))(1 - \lambda(t_{(2)})),$$

$$B_2 = \lambda(t_{(1)})\lambda(t_{(2)})(1 - \lambda(t_1))(1 - \lambda(t_2)).$$

Here, $\mu = 1$ if $|\hat{\eta}_1| > |\hat{\eta}_2|$ and 0 otherwise, and $\delta = 1$ if the pretest procedure is general-to-specific and 0 if the procedure is specific-to-general.

Because E(UR) depends on 5 parameters, only a six-dimensional plot would do full justice to its behavior. This task being beyond us, let us first consider the mean squared error $\boldsymbol{R} = \operatorname{MSE}(\boldsymbol{W}\hat{\boldsymbol{\eta}})$ and the expected reported variance $\operatorname{E}(\boldsymbol{W})$ for the two procedures. Both functions depend on η_1, η_2, and r. The $\operatorname{E}(\boldsymbol{W})$ is always bounded, as noted in Section 3. The matrix $\boldsymbol{R}$ is also bounded in the general-to-specific procedure, but $\boldsymbol{R}$ can be unbounded in the specific-to-general procedure. More specifically,

$$\max_{\eta_1,\eta_2} \boldsymbol{R}(\eta_1, \eta_2, r) \to \infty \quad \text{as } r \to 1,$$

when the procedure is specific-to-general. This very different behavior of $\boldsymbol{R}$ in the two procedures is reflected in Fig. 3, where we consider

$$\mathrm{E}^{**}(\mathrm{UR}) := \max_{\boldsymbol{\eta}} \mathrm{E}^{*}(\mathrm{UR}) = 1 - \min_{\boldsymbol{\eta}} \min_{1 \leqslant j \leqslant m} \xi_j(\boldsymbol{R}^{-1/2}(\operatorname{E}\boldsymbol{W})\boldsymbol{R}^{-1/2}),$$

as a function of r.[10]

For both procedures the function $\mathrm{E}^{**}(\mathrm{UR})$ is symmetric around $r = 0$. For $r = 0$ the two procedures are the same and the function value is almost 0.90. In the specific-to-general procedure, $\mathrm{E}^{**}(\mathrm{UR})$ increases monotonically to 1 as r increases from 0 to 1. The general-to-specific procedure has a uniformly lower $\mathrm{E}^{**}(\mathrm{UR})$, its behavior is non-monotonic, and it converges to 0.87 as $r \to 1$, the same maximum value as in the case $m = 1$ (depicted as a horizontal line in the figure). The difference between the two procedures is especially large when r is close to 1, that is when $\boldsymbol{M}\boldsymbol{z}_1$ and $\boldsymbol{M}\boldsymbol{z}_2$ are strongly correlated. This can be understood as follows. Let $r = 1$ and let $\eta_1 = -\eta_2 = \bar{\eta}$, say. Then, for large $\bar{\eta}$, the probability of choosing one of the partially restricted models $\mathscr{M}_1$ or $\mathscr{M}_2$ approaches 0. In the specific-to-general case, we will choose the restricted model $\mathscr{M}_0$ with probability approaching 0.95 and model $\mathscr{M}_{12}$

[10] The E(UR) in Figs. 3, 5 and 6 was computed using routines D01FCF and D01GBF of the Fortran NAG Library, Mark 18. The (very computer-intensive) maximization in Fig. 3 is achieved by a grid method. Fig. 4 was obtained numerically in Matlab 5.3 using the Nelder-Mead simplex method.

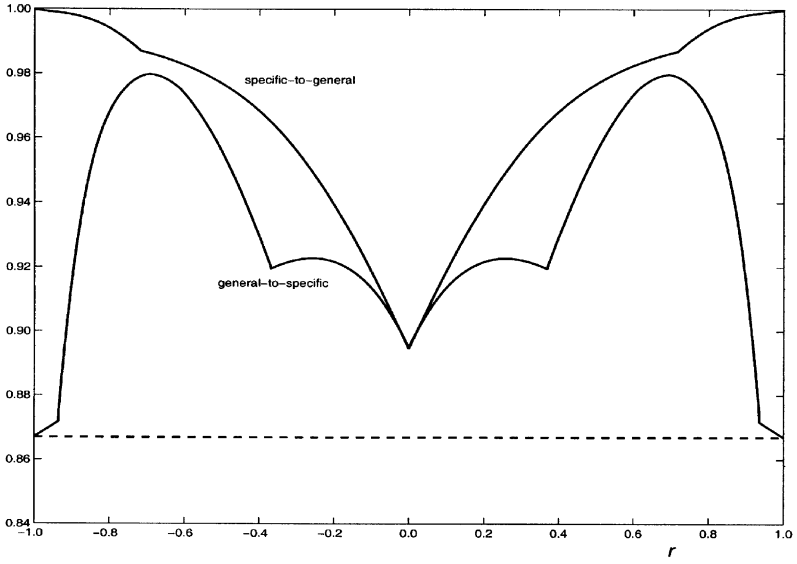

Fig. 3. max(E(UR)) as a function of r ($m = 2$).

with probability approaching 0.05. Hence, for $r = 1$ and $\bar{\eta} \to \infty$, we find that E(UR) approaches 1 for any q_0^2. (In fact, the MSE of the pretest estimator is unbounded and proportional to $\bar{\eta}^2$ when $\bar{\eta}$ approaches ∞.) But in the general-to-specific case, the MSE is always bounded and hence $\mathrm{E}^*(\mathrm{UR}) < 1$, using (6).

Although the functions are continuous, there are various kinks. This is the result of the fact that there exist various local maxima. At a kink we move from one local maximum to another local maximum. Clearly, underreporting can be a very serious problem and, for $m \geqslant 2$, can be essentially unbounded, depending on the chosen pretest procedure.

For $r = 0$ the worst case gives $\mathrm{E}^{**}(\mathrm{UR}) = 0.8669$ for $m = 1$ and 0.8953 for $m = 2$. We now ask how underreporting depends on m. There are 2^m models to consider and one may think therefore that 'badness' increases by a factor of 2^m. On the other hand, all t-statistics are functions of only m random variables $\hat{\eta}_1, \ldots, \hat{\eta}_m$, so that 'badness' increases possibly only by a factor of m. We consider only the special case where $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$. Then all vectors $\boldsymbol{M}\boldsymbol{z}_i$ are orthogonal, and the m-dimensional problem collapses to m one-dimensional problems (Theorem 2). The maximum $\mathrm{E}^{**}(\mathrm{UR})$ is plotted in Fig. 4 as a function of m.

The figure reveals that $\mathrm{E}^{**}(\mathrm{UR})$ increases with m but less than linearly. Although this result is valid only when $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$, it nevertheless suggests that the increase in 'badness' is not as fast as one might have feared.

In a practical situation, we know q_0^2, $\boldsymbol{q}$, and r, but not η_1 and η_2. Let us analyze a typical and representative situation where $q_0^2 = 2$, $\boldsymbol{q} = (1/3, (2/3)\sqrt{2})'$ (so that $\boldsymbol{q}'\boldsymbol{q} = 1$), and $r = 0.8$.

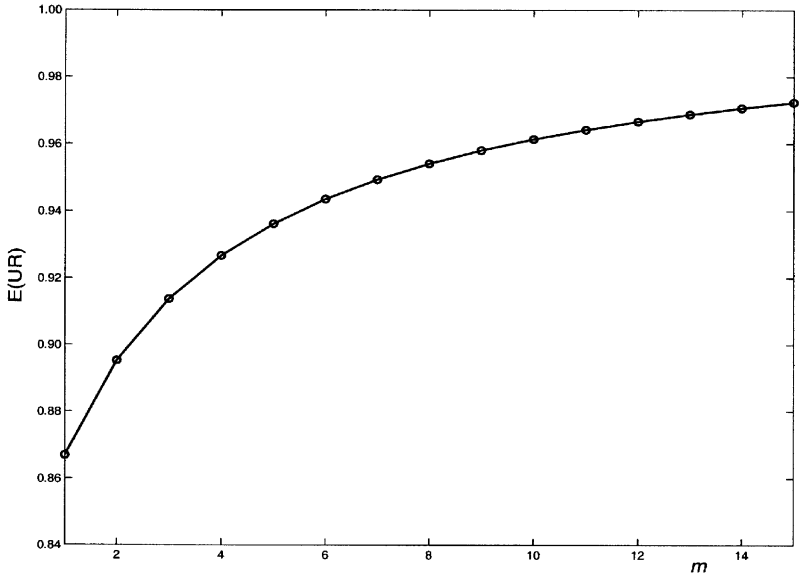

Fig. 4. max(E(UR)) as a function of m ($\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$).

Figs. 5 and 6 give the E(UR) as a function of η_1 and η_2, first for the general-to-specific procedure, then for the specific-to-general procedure. The E(UR) lies always between 0 and 1, and is symmetric around the point $(\eta_1, \eta_2) = (0, 0)$. The functional dependence on (η_1, η_2) is quite complicated, and also quite different for the two procedures. In the general-to-specific procedure (Fig. 5), E(UR) is very close to 0 at $(\eta_1, \eta_2) = (4, -4)$, but can be as large as 0.6551 at $(0.4, 1.6)$. In the specific-to-general procedure (Fig. 6), E(UR) varies from around 0 at $(4, 4)$ to 0.8798 around the point $(4, -4)$. In this case (and in general), the specific-to-general is more sensitive to underreporting than the general-to-specific procedure. The contours in the (η_1, η_2) plane are iso-value curves.

7. Conclusions

In this paper, we have analyzed the effect of ignoring the model selection procedure in reporting the bias, variance and mean squared error of the commonly used least-squares estimator. We conclude that underreporting can be a very serious problem. The pretest bias appears to be less of a problem than the pretest variance.

When we have m auxiliary regressors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, there are 2^m models to choose between. There are many different possible (viable) procedures to select the model. We find that the choice of model selection procedure matters a lot. Our results for $m = 2$ suggest that the general-to-specific procedure has more desirable properties than the specific-to-general procedure, in the sense that max(E(UR)) is smaller for general-to-specific than for the specific-to-general. The influence of the selection

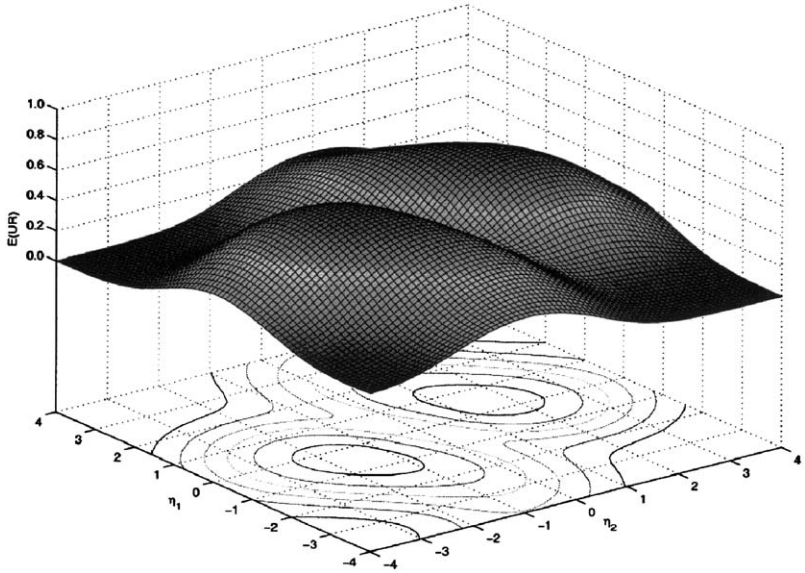

Fig. 5. E(UR) as a function of η_1 and η_2: general-to-specific.

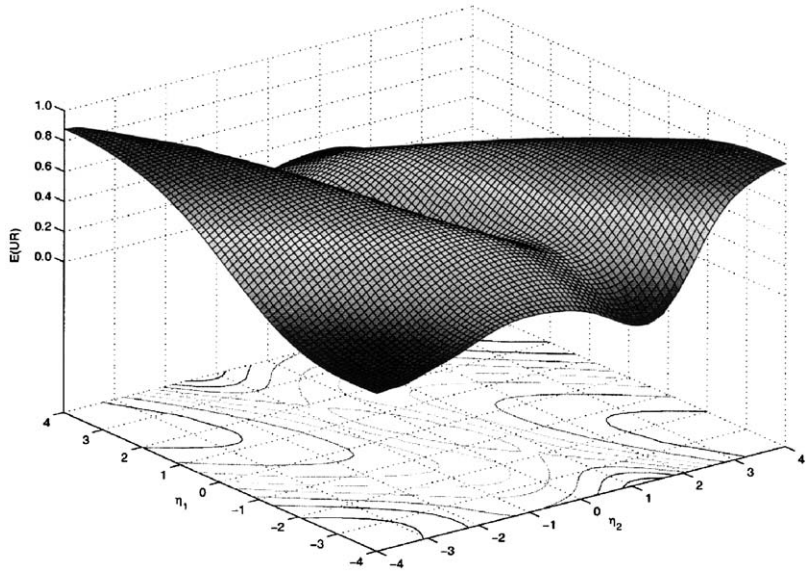

Fig. 6. E(UR) as a function of η_1 and η_2: specific-to-general.

Table 1
E(UR) as a function of the d.f. $n-k-m$ (σ^2 unknown)

$n-k-m$	η			
	0	1	2	4
10	0.76	0.83	0.77	0.26
30	0.80	0.85	0.78	0.22
50	0.81	0.86	0.79	0.21
∞	0.82	0.86	0.79	0.19

procedure is higher when the correlation between the z-variables (measured by $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$) is high, than when it is low. If we can choose the auxiliary regressors such that they are 'orthogonal' (that is, $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}=\boldsymbol{I}_m$), then the general-to-specific and specific-to-general pretest procedures are the same (among others), and hence the sampling properties of the estimators do not depend on the model selection procedure.

As the number of auxiliary regressors m grows, the dangers of underreporting grow as well. In the special 'orthogonal' case ($\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}=\boldsymbol{I}_m$) this growth is somewhat less than linear.

Although Theorem 1 is valid whether or not σ^2 is known, Sections 3–6 assume that σ^2 is known. This is of course unrealistic and we need to address the question how the results are affected when σ^2 is unknown. As an example, let us consider the case of Fig. 2 where $m=1$, $q_0^2=\infty$ and $c=1.96$. When σ^2 is known, the E(UR) takes the values 0.82, 0.86, 0.79, and 0.19 for η equal to 0, 1, 2 and 4, respectively. When σ^2 is not known the calculations are more involved and depend on the degrees of freedom $n-k-m$. The results are summarized in Table 1.[11]

We see that the effects of estimating σ^2 are relatively small, especially in the region of interest where $|\eta|$ is around 1 or 2. Although this example is typical for the behavior of the E(UR), more work is needed in this direction, especially for $m \geqslant 2$.

Future work will also have to clarify various other issues not covered in this paper. For example, there are selection procedures other than general-to-specific and specific-to-general. How does the E(UR) depend on these? All these pretest procedures are discontinuous; they choose one of 2^m models based on the values of t- and F-statistics. A continuous procedure can also be defined. Theorem 1 allows for this situation, and it will probably lead to better sampling properties; see Magnus (2002) for an analysis of the case $m=1$.

Also, the E(UR) depends on $\boldsymbol{\eta}$ which is unknown. What is the best way to estimate the expected underreporting ratio? Simplest is to replace $\boldsymbol{\eta}$ by $\hat{\boldsymbol{\eta}}$. The properties of this estimator will have to be analyzed. Another possibility is to report the worst possible case, given the observed $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$. That is, to calculate $\max_{\boldsymbol{\eta}}\mathrm{E(UR)}$. In Figs. 5 and 6, for example, the maxima are 0.6551 (general-to-specific) and 0.8798

[11] Table 1 was obtained by numerical integration (in Matlab 5.3, using NAG routine D01AMF) of $\boldsymbol{R}(\boldsymbol{\eta}/s_u)$, where s_u^2 is the LS estimator of σ^2 in the unrestricted model.

(specific-to-general), and they give an indication of how bad underreporting can be in a specific situation.

Acknowledgements

We are grateful to seminar participants at the University of Michigan, Michigan State University, the University of Wisconsin at Madison, Tilburg University, Eindhoven University of Technology, the University of Amsterdam, the University of Vienna, ESEM 2001 in Lausanne, and to two referees for constructive and useful comments.

Appendix. Constrained least squares

Consider the linear regression model $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\gamma}+\boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{X}$ $(n\times k)$ and $\boldsymbol{Z}$ $(n\times m)$ are matrices of nonrandom regressors, $\boldsymbol{\varepsilon}$ is a random vector of unobservable disturbances, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown nonrandom parameter vectors. Assume that $k\geqslant 1$, $m\geqslant 1$, $k+m\leqslant n-1$, that $(\boldsymbol{X}:\boldsymbol{Z})$ has full column-rank, and that the disturbances $\{\varepsilon_i\}$ are i.i.d. $\mathrm{N}(0,\sigma^2)$. Define $\boldsymbol{M}=\boldsymbol{I}_n-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, $\boldsymbol{Q}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}$, and $\boldsymbol{\eta}=(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{1/2}\boldsymbol{\gamma}/\sigma$. Let $\boldsymbol{S}_i$ be an $m\times r_i$ selection matrix of rank $r_i\geqslant 0$, so that $\boldsymbol{S}_i'=(\boldsymbol{I}_{r_i}:\boldsymbol{0})$ or a column-permutation thereof. We are interested in the constrained LS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the constraint being $\boldsymbol{S}_i'\boldsymbol{\gamma}=\boldsymbol{0}$.

Lemma A1. *Let* $\boldsymbol{b}_r=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ *and* $\hat{\boldsymbol{\theta}}=(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{y}$. *Then the constrained* LS *estimators of* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$ *are given by*

$$\boldsymbol{b}_{(i)}=\boldsymbol{b}_r-\boldsymbol{Q}\boldsymbol{W}_i\hat{\boldsymbol{\theta}},\quad \boldsymbol{c}_{(i)}=(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i\hat{\boldsymbol{\theta}},$$

where

$$\boldsymbol{P}_i:=(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{S}_i(\boldsymbol{S}_i'(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i'(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}$$

is a symmetric idempotent $m\times m$ *matrices of rank* r_i, *and* $\boldsymbol{W}_i:=\boldsymbol{I}_m-\boldsymbol{P}_i$. *(If* $r_i=0$ *then* $\boldsymbol{P}_i=\boldsymbol{0}$.*) The distribution of* $\boldsymbol{b}_{(i)}$ *is given by*

$$\boldsymbol{b}_{(i)}\sim \mathrm{N}(\boldsymbol{\beta}+\sigma\boldsymbol{Q}\boldsymbol{P}_i\boldsymbol{\eta},\ \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}+\boldsymbol{Q}\boldsymbol{W}_i\boldsymbol{Q}')),$$

the distribution of $\boldsymbol{c}_{(i)}$ *by*

$$\boldsymbol{c}_{(i)}\sim \mathrm{N}(\sigma(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i\boldsymbol{\eta},\ \sigma^2((\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2})),$$

and the covariance of $\boldsymbol{b}_{(i)}$ *and* $\boldsymbol{c}_{(i)}$ *is* $\mathrm{cov}(\boldsymbol{b}_{(i)},\boldsymbol{c}_{(i)})=-\sigma^2\boldsymbol{Q}\boldsymbol{W}_i(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}$. *The residual vector is* $\boldsymbol{e}_{(i)}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}_{(i)}-\boldsymbol{Z}\boldsymbol{c}_{(i)}=\boldsymbol{D}_i\boldsymbol{y}$, *where*

$$\boldsymbol{D}_i=\boldsymbol{M}-\boldsymbol{M}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{W}_i(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{Z}'\boldsymbol{M}$$

is a symmetric idempotent matrix of rank $n-k-m+r_i$, *and the distribution of* $s_{(i)}^2=\boldsymbol{e}_{(i)}'\boldsymbol{e}_{(i)}/(n-k-m+r_i)$ *is given by*

$$\frac{(n-k-m+r_i)s_{(i)}^2}{\sigma^2}\sim\chi^2(n-k-m+r_i,\boldsymbol{\eta}'\boldsymbol{P}_i\boldsymbol{\eta}).$$

Proof. Let $\boldsymbol{X}_* = (\boldsymbol{X} : \boldsymbol{Z})$, $\boldsymbol{\beta}_*' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')$, and $\boldsymbol{R}' = (\boldsymbol{0} : \boldsymbol{S}_i')$. The LS estimator of $\boldsymbol{\beta}_*$ in the model $\boldsymbol{y} = \boldsymbol{X}_*\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}$ under the restriction $\boldsymbol{R}'\boldsymbol{\beta}_* = \boldsymbol{0}$ is then given by

$$\boldsymbol{b}_* = (\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*'\boldsymbol{y} - (\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{R}(\boldsymbol{R}'(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}'(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*'\boldsymbol{y}.$$

Noting that

$$(\boldsymbol{X}_*'\boldsymbol{X}_*)^{-1} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{Z} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\boldsymbol{Q}' & -\boldsymbol{Q}(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2} \\ -(\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1/2}\boldsymbol{Q}' & (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z})^{-1} \end{pmatrix},$$

and simplifying, the results follow. □

Lemma A2. *Under the same conditions as in Lemma* A1, *assume*, *in addition*, *that* $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{I}_m$. *The constrained* LS *estimators of* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$ *are then given by* $\boldsymbol{b}_{(i)} = \boldsymbol{b}_r - \boldsymbol{Q}\boldsymbol{W}_i\hat{\boldsymbol{\theta}}$ *and* $\boldsymbol{c}_{(i)} = \boldsymbol{W}_i\hat{\boldsymbol{\theta}}$, *where* $\boldsymbol{W}_i := \boldsymbol{I}_m - \boldsymbol{S}_i\boldsymbol{S}_i'$ *is a diagonal* $m \times m$ *matrix with* $m - r_i$ *ones and* r_i *zeros on the diagonal*, *such that the jth diagonal element of* $\boldsymbol{W}_i$ *is* 0 *if* γ_j *is constrained to be zero*, *and* 1 *otherwise. The distribution of* $\boldsymbol{b}_{(i)}$ *is given by*

$$\boldsymbol{b}_{(i)} \sim \mathrm{N}(\boldsymbol{\beta} + \sigma\boldsymbol{Q}\boldsymbol{S}_i\boldsymbol{S}_i'\boldsymbol{\eta},\ \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{Q}\boldsymbol{W}_i\boldsymbol{Q}')),$$

the distribution of $\boldsymbol{c}_{(i)}$ *by* $\boldsymbol{c}_{(i)} \sim \mathrm{N}(\sigma\boldsymbol{W}_i\boldsymbol{\eta},\ \sigma^2\boldsymbol{W}_i)$, *and the covariance of* $\boldsymbol{b}_{(i)}$ and $\boldsymbol{c}_{(i)}$ *is* $\mathrm{cov}(\boldsymbol{b}_{(i)}, \boldsymbol{c}_{(i)}) = -\sigma^2\boldsymbol{Q}\boldsymbol{W}_i$. *Hence all models which include* $\boldsymbol{z}_j$ *as a regressor will have the* same *estimator of* γ_j, *namely* $\hat{\theta}_j$, *irrespective which other* γ*'s are estimated. Moreover*, *the estimators* $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ *are independent.*

Proof. We have $\boldsymbol{P}_i = \boldsymbol{S}_i(\boldsymbol{S}_i'\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i'$, and, since $\boldsymbol{S}_i'$ is a selection matrix of the form $(\boldsymbol{I}_{r_i} : \boldsymbol{0})$ or a column-permutation thereof, it follows that $\boldsymbol{S}_i'\boldsymbol{S}_i = \boldsymbol{I}_{r_i}$ and hence that $\boldsymbol{P}_i$ is a diagonal matrix with r_i ones and $m - r_i$ zeros on the diagonal, and that $\boldsymbol{W}_i$ is a diagonal matrix with $m - r_i$ ones and r_i zeros on the diagonal. The results now follow from Lemma A1. □

References

Abadir, K.M., Magnus, J.R., 2002. Notation in econometrics: a proposal for a standard. The Econometrics Journal 5, 76–90.

Bancroft, T.A., 1944. On biases in estimation due to the use of preliminary tests of significance. Annals of Mathematical Statistics 15, 190–204.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference (with discussion). Journal of the Royal Statistical Society, Series A 158, 419–466.

Giles, J.A., Giles, D.E.A., 1993. Pre-test estimation and testing in econometrics: recent developments. Journal of Economic Surveys 7, 145–197.

Giles, J.A., Srivastava, V., 1993. The exact distribution of a least-squares regression coefficient after a preliminary *t*-test. Statistics and Probability Letters 16, 59–64.

Hendry, D.F., 2001. Achievements and challenges in econometric methodology. Journal of Econometrics 100, 7–10.

Hoover, K.D., Perez, S.J., 1999. Data mining reconsidered; encompassing and the general-to-specific approach to specification search. The Econometrics Journal 2, 1–25.

Huntsberger, D.V., 1955. A generalization of a preliminary testing procedure for pooling data. Annals of Mathematical Statistics 26, 734–743.

Judge, G.G., Bock, M.E., 1978. The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics. North-Holland, Amsterdam.

Judge, G.G., Bock, M.E., 1983. Biased estimation. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. 1. North-Holland, Amsterdam (Chapter 10).

Leeb, H., Pötscher, B.M., 2003. The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. Econometric Theory 19, 100–142.

Lovell, M.C., 1983. Data mining. The Review of Economics and Statistics 65, 1–12.

Magnus, J.R., 1999. The traditional pretest estimator. Theory of Probability and its Applications 44, 293–308.

Magnus, J.R., 2002. Estimation of the mean of a univariate normal distribution with known variance. The Econometrics Journal 5, 225–236.

Magnus, J.R., Durbin, J., 1999. Estimation of regression coefficients of interest when other regression coefficients are of no interest. Econometrica 67, 639–643.

Mittelhammer, R.C., 1984. Restricted least squares, pre-test, OLS and Stein-rule estimators: Risk comparisons under model misspecification. Journal of Econometrics 25, 151–164.

Mosteller, F., 1948. On pooling data. Journal of the American Statistical Association 43, 231–242.

Pötscher, B.M., 1991. Effects of model selection on inference. Econometric Theory 7, 163–185.

Pötscher, B.M., Novak, A.J., 1998. The distribution of estimators after model selection: large and small sample results. Journal of Statistical Computation and Simulation 60, 19–56.

Roehrig, C.S., 1984. Optimal critical regions for pre-test estimators using a Bayes risk criterion. Journal of Econometrics 25, 3–14.

Sclove, S.L., Morris, C., Radhakrishnan, R., 1972. Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. The Annals of Mathematical Statistics 43, 1481–1490.

Sen, P.K., 1979. Asymptotic properties of maximum likelihood estimators based on conditional specification. Annals of Statistics 7, 1019–1033.

Thomson, M., Schmidt, P., 1982. A note on the comparison of the mean square error of inequality constrained least squares and other related estimators. The Review of Economics and Statistics 64, 174–176.

Zhang, P., 1992. On the distributional properties of model selection criteria. Journal of the American Statistical Association 87, 732–737.