A COMPARISON OF WALS ESTIMATION WITH PRETEST AND MODEL SELECTION ALTERNATIVES WITH AN APPLICATION TO COSTS OF HIP FRACTURE TREATMENTS

Antti Liski[1], *Erkki P. Liski*[2], *Reijo Sund* [3], *Merja Juntunen* [4]

[1] Department of Signal Processing, Tampere University of Technology,
P.O.Box 553, FIN-33101 Tampere, FINLAND, antti.liski@tut.fi
[2] Department of Mathematics and Statistics, University of Tampere,
FIN-33014 Tampere, FINLAND, epl@uta.fi
[3] Service System Research Unit,
National Institute for Health and Welfare, FINLAND
reijo.sund@thl.fi
[4] Centre for Health and Social Economics CHESS,
National Insitute for Health and Welfare, FINLAND
merja.juntunen@thl.fi

ABSTRACT

The paper considers a new model averaging method called weighted average least squares (WALS). The method has good risk profile and its computational burden is light. The WALS technique can be easily applied to large data sets when the number of regressors is large. In the current paper the theory is used to compare the costs of hip fracture treatments between hospital districts in Finland.

1. INTRODUCTION

This paper presents a model averaging technique introduced by Magnus [9] and Magnus & Durbin [11] which is called weighted average least squares (WALS). Secondly, we apply this technique on hip fracture data of 11961 patients aged 50 or over in years 1999-2005. The purpose is to compare treatment costs of hip fracture patients between hospital districts in Finland. WALS is computationally superior over the post model selection (PMS) estimators because computing time of WALS increases only linearly with m, the number of regressors, while computing time of the PMS estimators is of order 2^m. WALS also has better risk profile over PMS estimators, and it avoids an unbounded risk. It is known that the finite-sample distributions of PMS estimators are difficult to estimate and the model selection (MS) step may have a dramatic effect on the sampling properties of PMS estimators [6].

Our framework is the ordinary linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ matrix of explanatory variables that we want to keep in the model on theoretical or other grounds. An $n \times m$ matrix $\mathbf{Z}$ contains m additional explanatory variables which we add in the model only if they are supposed to improve estimation of $\boldsymbol{\beta}$. Following Danilov and Magnus [3] we call the x-variables "focus" regressors and z-variables auxiliary regressors. The matrix $(\mathbf{X}, \mathbf{Z})$ is assumed to be of full column rank.

We have M linear regression models $\mathcal{M}_1, \ldots, \mathcal{M}_M$ such that

$$\mathcal{M}_i : \{\boldsymbol{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n\},$$

where $\mathbf{Z}_i = \mathbf{Z}\mathbf{W}_i$ and $\mathbf{W}_i = \mathrm{diag}(w_{i1}, \ldots, w_{im})$ is an $m \times m$ diagonal matrix with diagonal elements $w_{ij} \in \{0, 1\}, j = 1, \ldots, m$. We may suppose that the models are in increasing order with respect to diagonal elements of $\mathbf{W}_i$ when the diagonal is interpreted as m-digit binary number $w_{i1}, \ldots, w_{im}$. Then the indices $1, \ldots, M$ are associated with the diagonals as follows

$$1 \to 00\ldots0, 2 \to 10\ldots0, 3 \to 01\ldots0, \\ \ldots, M \to 11\ldots1,$$

where the number of models is $M = 2^m$. For $m = 2$ we have the 4 diagonal matrices $\mathbf{W}_1 = \mathrm{diag}(0, 0)$, $\mathbf{W}_2 = \mathrm{diag}(1, 0), \mathbf{W}_3 = \mathrm{diag}(0, 1), \mathbf{W}_4 = \mathrm{diag}(1, 1)$ and the corresponding matrices of auxiliary regressors

$$\mathbf{Z}_i = \mathbf{Z}\mathbf{W}_i; \quad i = 1, 2, 3, 4.$$

Given an MS procedure S selecting from the set of candidate models $\mathcal{M}_1, \ldots, \mathcal{M}_M$, the associated PMS estimator may be represented as

$$\hat{\boldsymbol{\beta}}_S = \sum_{i=1}^{M} I(S = \mathcal{M}_i)\hat{\boldsymbol{\beta}}_i, \tag{2}$$

where $\hat{\boldsymbol{\beta}}_i$ denotes the LS estimator of $\boldsymbol{\beta}$ under $\mathcal{M}_i$ and $I(\cdot)$ is the indicator function with the value 1 for the selected model and and 0 for all other models. There are many well known MS methods such as Akaike's (AIC) and Bayesian (BIC) information criteria, as well the minimum description length (MDL) principle, for example. For (2) we have to evaluate the model goodness criterion c_S for each model $\mathcal{M}_i, 1 \leq i \leq M$. If the number of variables m is large ($M = 2^m$), the task can be computationally prohibitive. By far the most common selection approach in practice is to apply a sequence of the hypothesis tests and attempt to identify the nonzero regression coefficients and select the corresponding regressors. Forward selection, backward elimination, and stepwise regression are the best known examples of these techniques [13]. All major statistical software have procedures for these techniques.

Clearly

$$\mathcal{M}_0 : \{\boldsymbol{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\} \text{ and } \mathcal{M}_M : \{\boldsymbol{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I}_n\},$$

where $\mathcal{M}_0$ is the fully restricted model without any auxiliary regressors and $\mathcal{M}_M$ is the unrestricted model containing all auxiliary regressors. The least squares (LS) estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ under the model $\mathcal{M}_M$ are [15] (Section 3.7)

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\hat{\boldsymbol{\gamma}} \quad \text{and} \quad \hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{M}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}\boldsymbol{y}, \qquad (3)$$

respectively, where $\hat{\boldsymbol{\beta}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ is the LS estimator of $\boldsymbol{\beta}$ under the model $\mathcal{M}_0$, $\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It is known that dropping z-variables from the model decreases the variance of the LS estimator of the remaining regression parameters [7]. However, after elimination of variables, the estimates of the remaining parameters are biased if the full model is correct.

The traditional approach to select between the models $\mathcal{M}_0$ and $\mathcal{M}_M$ is to test the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ and to include $\mathbf{Z}$ if the hypothesis "$\boldsymbol{\gamma} = \mathbf{0}$" is rejected and exclude z-variables otherwise. Then inference on $\boldsymbol{\beta}$ is made as if the resulting model were correct. In this approach the alternative estimators under consideration are the restricted LS estimator $\hat{\boldsymbol{\beta}}_0$ under the restriction $\boldsymbol{\gamma} = \mathbf{0}$ and the unbiased unrestricted LS estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ in the model (1). An another traditional approach is to compare the estimators $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_M$ with respect to the mean squared error (MSE) criterion. Then we test the hypothesis

$$MSE(\hat{\boldsymbol{\beta}}_0) \leq MSE(\hat{\boldsymbol{\beta}}_M) \qquad (4)$$

and choose $\hat{\boldsymbol{\beta}}_0$ if the hypothesis is accepted. Here the inequality "$\leq$" refers to to Löwner ordering of nonnegative definite matrices. Toro-Vizcarrondo and Wallace [18] made this point and developed a test for the hypothesis (4). A review of the general theory of comparing estimators under exact or stochastic linear restrictions with respect to the MSE criterion can found e.g. in Rao et.al. [14] and in Judge and Bock [5].

We shall more generally consider estimators of the model average form

$$\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{M} c(\mathcal{M}_i)\hat{\boldsymbol{\beta}}_i, \qquad (5)$$

where the weights $c(\mathcal{M}_i) \geq 0, 1 \leq i \leq M$, sum to one and are allowed to be random, as in the post-selection estimator class. Buckland [1] suggested using weights proportional to $\exp(-AIC_i/2)$, where AIC_i is the AIC score for the candidate model $\mathcal{M}_i$. Similar weighting can be derived from other model selection criteria as well. Liski et al. [8] propposed a model average estimator using weights derived from the MDL criterion. However, for large values of m the estimator (5) is infeasible unless the set of candidate models is somehow restricted.

2. PRETESTING

For simplicity we assume for a moment that $m = 1$, and consequently $\mathbf{Z}$ is a single $n \times 1$ vector $\boldsymbol{z}$. Then we have two alternative models $\mathcal{M}_0, \mathcal{M}_1$, where $\mathcal{M}_0$ is the restricted model as before and

$$\mathcal{M}_1 : \ \boldsymbol{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma, \sigma^2\mathbf{I})$$

is the unrestricted model. At first we assume that σ^2 is known, but later in WALS implementation (Section 4) σ^2 is replaced by its usual unbiased estimator obtained from the unrestricted model. Using the notation

$$\boldsymbol{q} = \frac{\sigma}{\sqrt{\boldsymbol{z}'\mathbf{M}\boldsymbol{z}}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{z} \quad \text{and} \quad \theta = \frac{\gamma}{\sigma/\sqrt{\boldsymbol{z}'\mathbf{M}\boldsymbol{z}}},$$

we can write the LS estimators for $\boldsymbol{\beta}$ and γ under $\mathcal{M}_1$ as

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0 - \hat{\theta}\boldsymbol{q}, \qquad \hat{\gamma} = \frac{\boldsymbol{z}'\mathbf{M}\boldsymbol{y}}{\boldsymbol{z}'\mathbf{M}\boldsymbol{z}},$$

where $\hat{\theta} = \frac{\hat{\gamma}}{\sigma}\sqrt{\boldsymbol{z}'\mathbf{M}\boldsymbol{z}}$ denotes the t-ratio, which follows the normal distribution $\mathrm{N}(\theta, 1)$ because σ^2 is known. Note that $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}_0$ are independent.

The traditional approach to choose between $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ is to use the t-ratio. If $\hat{\theta}$ is large, we choose the unrestricted LS estimator $\hat{\boldsymbol{\beta}}_1$, and if $\hat{\theta}$ is small, we choose the restricted LS estimator $\hat{\boldsymbol{\beta}}_0$. This leads to the estimator

$$\tilde{\boldsymbol{\beta}}_{pre} = \begin{cases} \hat{\boldsymbol{\beta}}_0 & \text{if} \quad |\hat{\theta}| \leq c; \\ \hat{\boldsymbol{\beta}}_1 & \text{if} \quad |\hat{\theta}| > c, \end{cases}$$

where c is some positive constant. For example, $c = 1.96$ corresponds to the 5% significance level.

Given that we are interested in the best possible estimation on $\boldsymbol{\beta}$, not in γ, the proper question of interest is, "Is $\hat{\boldsymbol{\beta}}_0$ better estimator than $\hat{\boldsymbol{\beta}}_1$?" When assessing estimators with respect to their MSE, we should compare their MSE matrices, which are

$$MSE(\hat{\boldsymbol{\beta}}_0) = \mathrm{Cov}(\hat{\boldsymbol{\beta}}_0) + \theta^2\boldsymbol{q}\boldsymbol{q}' \quad \text{and}$$
$$MSE(\hat{\boldsymbol{\beta}}_1) = \mathrm{Cov}(\hat{\boldsymbol{\beta}}_0) + \mathrm{Var}(\hat{\theta})\boldsymbol{q}\boldsymbol{q}' = \mathrm{Cov}(\hat{\boldsymbol{\beta}}_0) + \boldsymbol{q}\boldsymbol{q}'.$$

Then

$$MSE(\hat{\boldsymbol{\beta}}_0) - MSE(\hat{\boldsymbol{\beta}}_1) = (\theta^2 - 1)\boldsymbol{q}\boldsymbol{q}',$$

where $\boldsymbol{q}$ is a known vector and θ is the usual non-centrality parameter associated with the t-ratio for testing $\gamma = 0$.

Hence (cf. [9])

$$\begin{aligned} MSE(\hat{\boldsymbol\beta}_0) &\le MSE(\hat{\boldsymbol\beta}_1) \quad \text{if} \quad \theta^2 < 1 \\ MSE(\hat{\boldsymbol\beta}_0) &= MSE(\hat{\boldsymbol\beta}_1) \quad \text{if} \quad \theta^2 = 1 \\ MSE(\hat{\boldsymbol\beta}_0) &\ge MSE(\hat{\boldsymbol\beta}_1) \quad \text{if} \quad \theta^2 > 1. \end{aligned}$$

Toro-Vizcarrondo and Wallace [18] obtained a uniformly most powerful test for $H_0 : \theta^2 \le 1$ vs. $H_1 : \theta^2 > 1$ from the probability

$$P(|\hat\theta| \le c \,|\, \theta^2 = 1) = 1 - \alpha,$$

where α denotes the significance level. E.g. a 5% test corresponds to $c = 2.65$.

There are two problems in applying either the usual t-test or the test of Toro-Vizcarrondo and Wallace. The first is that the choice of significance is largely arbitrary. In the preliminary test we decide whether to use $\hat{\boldsymbol\beta}_0$ or $\hat{\boldsymbol\beta}_1$. The second problem is that after the test neither $\hat{\boldsymbol\beta}_0$ nor $\hat{\boldsymbol\beta}_1$ is actually used. The estimator actually used is the pretest estimator

$$\tilde{\boldsymbol\beta}_{pre} = \lambda\hat{\boldsymbol\beta}_1 + (1-\lambda)\hat{\boldsymbol\beta}_0, \tag{6}$$

where $\lambda = 1$ if $|\hat\theta| > c$, and otherwise $\lambda = 0$.

More generally, any estimator of the form

$$\tilde{\boldsymbol\beta} = \lambda\hat{\boldsymbol\beta}_1 + (1-\lambda)\hat{\boldsymbol\beta}_0,\ 0 \le \lambda \le 1 \quad \text{and } \lambda = \lambda(\hat\theta) \tag{7}$$

will be called a weighted average least squares (WALS) estimator [11]. The pretest estimator is an example of such an estimator. Usually λ is a nondecreasing function of $|\hat\theta|$, so that the larger $|\hat\theta|$ the larger λ will be and hence more weight will be put on $\hat{\boldsymbol\beta}_1$ relative to $\hat{\boldsymbol\beta}_0$. It turns out that the pretest estimators (6) have poor properties [9] and better estimators can be found in the wider class of WALS estimators [10].

The following equivalence theorem, originally proved by Magnus and Durbin [11] and extended by Danilov and Magnus [3] turns out useful in the study of WALS estimators.

Theorem 1. *(Equivalence theorem) Let $\tilde{\boldsymbol\beta} = \lambda\hat{\boldsymbol\beta}_1 + (1-\lambda)\hat{\boldsymbol\beta}_0$ be a WALS estimator of $\boldsymbol\beta$ and $\tilde\theta = \lambda\hat\theta$, where λ is as in* (7). *Then*

$$\begin{aligned} E(\tilde{\boldsymbol\beta}) &= \boldsymbol\beta - E(\tilde\theta - \theta)\boldsymbol q, \\ \mathrm{Var}(\tilde{\boldsymbol\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathrm{Var}(\tilde\theta)\boldsymbol{qq}' \text{ and hence} \\ MSE(\tilde{\boldsymbol\beta}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + MSE(\tilde\theta)\boldsymbol{qq}'. \end{aligned}$$

The equivalence theorem expresses the expectation, variance and mean square error of a WALS estimator $\tilde{\boldsymbol\beta}$ of $\boldsymbol\beta$ as a function of the corresponding characteristics of the estimator $\tilde\theta$ of θ. Thus $\mathrm{Var}(\tilde{\boldsymbol\beta})$ and $MSE(\tilde{\boldsymbol\beta})$ are minimized if and only if $\mathrm{Var}(\tilde\theta)$ and $MSE(\tilde\theta)$, respectively, are minimized.

Theorem 1 is important because it shows that studying the WALS estimator (7) for the regression problem is equivalent to studying the estimator $\tilde\theta = \lambda(\hat\theta)\hat\theta$ of θ in the simple normal distribution $N(\theta, 1)$. Suppose that we have found an optimal estimator $\tilde\theta_0 = \lambda_0(\hat\theta)\hat\theta$ of θ. Then the equivalence theorem guarantees that this same λ-function λ_0 will provide the optimal WALS estimator (7) of $\boldsymbol\beta$. Thus the problem is to find an optimal λ-function. When the risk of an estimator $\tilde\theta$ is defined as its MSE,

$$R(\theta; \tilde\theta) = E_\theta(\tilde\theta - \theta)^2, \tag{8}$$

Magnus [9] showed that the traditional pretest estimator (6) has many undesirable risk properties and the WALS estimators (7) have advantages (cf. [10]) over the pretest estimators (6).

3. THE LAPLACE ESTIMATOR

Let $y \sim \mathrm{N}(\theta, 1)$ and let $T(y) = \lambda y$ be an estimator of θ, where $\lambda = \lambda(y)$ is a scalar function of y such that $0 \le \lambda(y) \le 1$. The ordinary LS estimator (and the ML estimator) of θ is $T(y) = y$ with $\lambda \equiv 1$ whereas the traditional pretest estimator is obtained when $\lambda(y) = 1$ if $|y| > c$ and $\lambda(y) = 0$ if $|y| \le c$ for some fixed threshold value $c > 0$. Note that $T(y) = \lambda y$ is a weighted average of the LS estimator y (corresponds to $\hat{\boldsymbol\beta}_1$) and the null estimator $T(y) \equiv 0$ (corresponds to $\hat{\boldsymbol\beta}_0$).

Our aim is to find a good WALS estimator of the regression parameter $\boldsymbol\beta$ with respect to the MSE criterion. The equivalence theorem tells us that a good estimator $T(y) = \lambda y$ of θ under the model $\mathrm{N}(\theta, 1)$ will allow us to make a good estimator of $\boldsymbol\beta$. Typically $0 \le \lambda \le 1$ is a nondecreasing function of $|y|$ so that $T(y) = \lambda y$ shrinks the LS estimator y towards zero. Magnus [10] considered a wide range of possible estimators of θ and finally he preferred the Laplace estimator which is admissible, has bounded risk, has good properties around $|\theta| = 1$, and is near optimal in terms of minimax regret. The value $|\theta| = 1$ is an important pivot since the risk of the null estimator of θ is less than the risk of the LS estimator y if and only if $|\theta| < 1$ ([9], [10]).

Assuming a Laplace prior density $\frac{1}{2}c\exp(-c|\theta|)$, $-\infty < \theta < \infty$, the posterior mean and variance of θ given y can be written as [12]

$$\begin{aligned} E(\theta|y) &= \frac{1+h(y)}{2}(y-c) + \frac{1-h(y)}{2}(y+c) \\ &= y - h(y)c, \end{aligned} \tag{9}$$

$$\mathrm{Var}(\theta|y) = 1 + c^2[1-h^2(y)] - \frac{c[1+h(y)]\phi(y-c)}{\Phi(y-c)},$$

where

$$h(y) = \frac{1 - e^{2cy}\Psi(y)}{1 + e^{2cy}\Psi(y)}, \quad \Psi(y) = \frac{\Phi(-y-c)}{\Phi(y-c)}$$

and $\phi(\cdot)$ denotes the density and $\Phi(\cdot)$ the distribution function of the standard normal distribution, respectively. The hyperparameter c is chosen $c = \log 2$ which implies that $\mathrm{median}(\theta) = 0$ and $\mathrm{median}(\theta^2) = 1$. The posterior mean (9) is the Laplace estimator

$$L(y) = y - h(y)c = \lambda(y)y \tag{10}$$

with $\lambda(y) = (1 - \frac{h(y)c}{y})$. The function $h(\cdot)$ is monotonically increasing with $h(-\infty) = -1,\ h(0) = 0,\ h(\infty) = 1$ and $h(-y) = -h(y)$, and hence $\lambda(y) = \lambda(-y) \to 1$ as $y \to \infty$. It can be shown that $h(y) \to 0.58956$ as $y \to 0$. Magnus [10] and Danilov [2] have studied the properties of the Laplace estimator in detail.

4. WALS ESTIMATION

4.1. Restricted LS estimators

We can always find an orthogonal $m \times m$ matrix $\mathbf{P}$ such that $\mathbf{P}'\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{P} = \mathbf{\Lambda}$ is diagonal and define new auxiliary regressors $\mathbf{Z}^* = \mathbf{Z}\mathbf{P}\mathbf{\Lambda}^{-1/2}$ and new auxiliary parameters $\boldsymbol{\gamma}^* = \mathbf{\Lambda}^{1/2}\mathbf{P}'\boldsymbol{\gamma}$ as noted by Magnus et al. [12]. Hence there is no loss of generality to posit

$$\mathbf{Z}'\mathbf{M}\mathbf{Z} = \mathbf{I}_m. \tag{11}$$

We assume in the sequel that (11) holds. Then it follows from (3) that

$$\hat{\boldsymbol{\gamma}} = \mathbf{Z}'\mathbf{M}\boldsymbol{y} \quad \text{and} \quad \hat{\boldsymbol{\gamma}} \sim \mathrm{N}(\boldsymbol{\gamma}, \sigma^2\mathbf{I}_m).$$

In general, given the assumption (11), the restricted LS estimator for $\boldsymbol{\beta}$ under the model $\mathcal{M}_i, 1 \le i \le M$, is

$$\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W}_i\hat{\boldsymbol{\gamma}}. \tag{12}$$

4.2. The WALS estimator of $\boldsymbol{\beta}$

The WALS estimator of $\boldsymbol{\beta}$ is defined as

$$\tilde{\boldsymbol{\beta}} = \sum_{i=1}^{M} v_i\hat{\boldsymbol{\beta}}_i, \tag{13}$$

where the weight functions v_i satisfy the conditions

$$v_i \ge 0, \quad \sum_i v_i = 1, \quad v_i = v_i(\hat{\boldsymbol{\gamma}}). \tag{14}$$

It follows from (12) and (13) that the WALS estimator can be written as

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\mathbf{W}\hat{\boldsymbol{\gamma}}, \tag{15}$$

where $\mathbf{W} = \sum_i v_i\mathbf{W}_i$ is a diagonal random matrix such that

$$\hat{\boldsymbol{\gamma}}'\mathbf{W} = (\lambda_1\hat{\gamma}_1, \ldots, \lambda_m\hat{\gamma}_m) \quad \text{with} \quad \hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \sigma^2).$$

We choose $\lambda_j = \lambda_j(\hat{\gamma}_j), 1 \le j \le m$. Since $\hat{\gamma}_1, \ldots, \hat{\gamma}_m$ are independent, it follows that $\lambda_1, \ldots, \lambda_m$ are independent. Hence we have m identical one-dimensional problems to estimate the elements λ_j.

4.3. LAPLACE weights

If we denote $\boldsymbol{\theta} = \boldsymbol{\gamma}/\sigma$ and compute $\hat{\boldsymbol{\theta}}$, then the elements $\hat{\theta}_1, \ldots, \hat{\theta}_m$ of $\hat{\boldsymbol{\theta}}$ are independent and $\hat{\theta}_j \sim \mathrm{N}(\theta_j, 1)$, $1 \le j \le m$. Then we have identical estimation problems in the models $\mathrm{N}(\theta_j, 1)$. By (10) and (9) we can compute the Laplace estimates $\tilde{\theta}_j = L(\hat{\theta}_j)$ and their variances $\phi_j^2 = \mathrm{Var}(\theta_j|\hat{\theta}_j), 1 \le j \le m$. We define $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_m)$, $\mathbf{\Phi} = \mathrm{diag}(\phi_1^2, \ldots, \phi_m^2)$ and note that $\boldsymbol{\gamma} = \sigma\boldsymbol{\theta}$. Then the WALS estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with Laplace weights can be computed as

$$\tilde{\boldsymbol{\gamma}} = \sigma\tilde{\boldsymbol{\theta}} \quad \text{and} \quad \tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 - \mathbf{Q}\tilde{\boldsymbol{\gamma}}.$$

The variance of $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\tilde{\boldsymbol{\gamma}}) = \sigma^2\mathbf{\Phi} \quad \text{and}$$
$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\,\mathrm{Var}(\tilde{\boldsymbol{\gamma}})\mathbf{Q}'.$$

The results above are based on the assumption that σ^2 is known, but in practice σ^2 is unknown and it must be estimated from data. This problem is solved by replacing σ^2 by its usual unbiased s^2 obtained from the unrestricted model. Danilov [2] showed that this approximation is very accurate and the main properties of the Laplace estimator change only marginally.

4.4. The equivalence theorem

The equivalence theorem proved by Danilov and Magnus ([3], Theorem 1) states that if the assumption (11) holds and the conditions (14) on weight functions are satisfied, then

$$\begin{aligned}\mathrm{Var}(\tilde{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\,\mathrm{Var}(\mathbf{W}\hat{\boldsymbol{\gamma}})\mathbf{Q}' \\ MSE(\tilde{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{Q}\,MSE(\mathbf{W}\hat{\boldsymbol{\gamma}})\mathbf{Q}'.\end{aligned}$$

Now $\mathrm{Var}(\tilde{\boldsymbol{\beta}})$ depends only on $\mathrm{Var}(\mathbf{W}\hat{\boldsymbol{\gamma}})$ and $MSE(\tilde{\boldsymbol{\beta}})$ only on $MSE(\mathbf{W}\hat{\boldsymbol{\gamma}})$.

5. MEDICAL CARE COSTS OF HIP FRACTURE TREATMENTS

Hip fracture is a common and important cause leading to lowered mobility or ultimately to death among the elderly population. In Finland, the number of hip fractures in people aged 50 or over was on average 5564 between the years 1998-2002 ([17]). Not only patients suffer from hip fractures, but they also cause remarkable costs to society. The costs of treating a hip fracture patient are about threefold compared to the caring for a patient without a fracture ([4]).

Comparison of treatments and outcomes between medical centres treating hip fractures can yield information for the development of treatment and serve as a quality assessment of care. Profiling medical care providers on the basis of quality of care and utilization of resources has become a widely used analysis in health care policy and research. Risk-adjustment is desirable when comparing hospitals or hospital districts with respect to a performance indicator such as treatment cost. Adjustment is intended to account for possible differences in patient case mix

This paper presents a model for hip fracture treatment costs in Finland using linear regression. Data were obtained by combining from several national registries [16] and consisted of 36492 patients aged 50 or over from the years 1999-2005. There are 21 hospital districts in Finland but here we report only results of the seven largest hospital districts. We concentrate only on patients who have not been institutionalised before the fracture and were not institutionalised after the fracture but were able to return home after the treatment. Patients who died within a year after the fracture were removed from the data. After all the exclusions, the data set used in this paper contained 11961 patients.

The dependent variable in the model is the treatment cost. The model contains seven focus regressors and 38 auxiliary regressors. The seven largest hospital districts are chosen as the focus regressors because we wish to test whether there is difference in treatment costs between the hospital districts. The hospital district of Helsinki and Uusimaa is taken as baseline. The set of auxiliary regressors contains a number of important comorbidities like congestive heart failure, diabetes and cancer. The auxiliary regressors are intended to reflect the mix of patients treated by a hospital or unit. The focus regressors are given in Table 1 and the auxiliary regressors are explained in Table 3.

6. ESTIMATION RESULTS

We estimate the model using WALS as discussed in Section 4. We also estimated the model applying a backward elimination (BE) technique. In BE we start using the full model with all $p+m$ variables and we eliminate auxiliary variables having the smallest F statistic but we never add regressors (Matlab's stepwise fit routine). The selected model thus contains all focus regressors and a subset of the auxiliary regressors. The reported LS estimates and standard errors are thus conditional on the model selected. BE selection procedure is considered because stepwise selection procedures are commonly used in practice.

The guidelines for hip fracture treatment are the same in the whole country and therefore we assume that all hospital districts treat the patients according to the same standards. In Table 1 all four estimation techniques indicate differences in treatment costs between hospital districts. The results of WALS seem to be in agreement with those of BE and the unrestricted LS whereas the restricted LS estimates show some discrepancy. The restricted LS indicates that costs in Central Finland are significantly higher than in Helsinki and Uusimaa but the other methods do not support the difference. According to the restricted LS the costs in Central Finland are significantly higher than than in Helsinki and Uusimaa but the other methods do not indicate such a difference. Satakunta hospital district have the highest and Northern Savo the lowest treatment costs, and these costs differ significantly from the costs in Helsinki and Uusimaa. By cost the most significant auxiliary variables are age, waiting for operation over 2 days, Parkinson disease, alcohol abuse, hypertension and diabetes.

The limited practical experience in the use of WALS seems to show that BE has a tendency to give larger absolute t-values on the average than WALS [12]. This tendency is mildly visible also in our results (Table 1 and Table 2). The variances of the BE estimates are conditional on the set of variables selected and the conditional estimates may be spuriously precise resulting in misleadingly high t-values. In our data the auxiliary variables are not correlated or only weakly, therefore WALS, BE and the unrestricted LS does not differ much.

We emphasize that even though we have found statistically significant differences in the treatment costs between the hospital districts, inferences concerning these results should be made with caution. Before drawing conclusions on the performance of the hospital districts a more careful and extensive study and interpretation of the results should be conducted with experts from the medical field participating actively in the research. These methods will be applied in this wider study and the work is in progress. The present application is a pilot study.

Table 1. Estimates $\hat{\boldsymbol{\beta}}$, standard errors (in parentheses) and t-values of the focus regressors.

Variable	$\hat{\boldsymbol{\beta}}_W$	t_W	$\hat{\boldsymbol{\beta}}_{BE}$	t_{BE}
Helsinki and Uusimaa	2331.04(307.08)	7.59	2004.82(307.55)	6.52
SW Finland	−65.70(117.75)	−0.56	−78.26(117.59)	−0.67
Satakunta	484.97(147.07)	3.30	451.30(146.90)	3.07
Pirkanmaa	−128.44(115.09)	−1.12	−148.18(114.93)	−1.29
N Savo	−826.07(143.74)	−5.75	−841.63(143.67)	−5.86
C Finland	142.92(142.79)	1.00	118.55(142.64)	0.83
N Ostrobothnia	−348.58(136.21)	−2.56	−373.42(136.26)	−2.74
Variable	**Unrestricted**	$\mathbf{t_u}$	**Restricted**	$\mathbf{t_r}$
Helsinki and Uusimaa	1927.96(310.16)	6.22	9052.56(63.50)	142.56
SW Finland	−69.33(117.82)	−0.59	45.24(122.16)	0.37
Satakunta	454.98(147.19)	3.09	825.78(151.24)	5.46
Pirkanmaa	−140.60(115.16)	−1.22	118.31(119.09)	0.99
N Savo	−831.68(143.80)	−5.78	−662.68(149.20)	−4.44
C Finland	128.79(142.90)	0.90	395.95(148.04)	2.67
N Ostrobothnia	−368.96(136.42)	−2.70	21.40(139.62)	0.15

SW-Southwest, N-Northern, C-Central

Table 2. Estimates and t-values of the auxiliary regressors using WALS, the unrestricted LS and BE.

Regressor	$\hat{\boldsymbol{\beta}}_W$	t_W	$\boldsymbol{\beta}_M$	t_M	$\hat{\boldsymbol{\beta}}_{BE}$	t_{BE}
HOSP90	8.35	2.16	7, 20	1, 82	7.93	2.01
AGE	81.41	19.82	86, 02	20, 77	85.79	20.87
HOSPDUM	284.53	2.77	351, 88	3, 17	380.24	3.46
OPWAIT	1570.97	12.73	1688, 85	13, 55	1689.15	13.57
FEMALE	−230.18	−2.60	−261, 29	−2, 90	−274.75	−3.09
CHF	313.91	3.00	280, 90	2, 41	318.79	2.77
Arr	188.00	1.78	267, 79	2, 33	288.89	2.54
Val	73.11	0.36	105, 69	0, 41	0	0
PCD	191.37	0.65	324, 99	0, 89	0	0
PVD	223.16	1.28	222, 19	1, 19	0	0
Par	922.27	2.18	1231, 32	2, 85	1251.85	2.90
PaD	1078.49	5.41	1198, 46	5, 56	1186.85	5.52
Dem	333.46	2.32	384, 28	2, 49	377.36	2.45
OND	490.59	2.74	617, 27	3, 19	630.53	3.27
CPD	193.30	1.74	229, 24	1, 88	275.30	2.29
Hyp	112.95	0.86	121, 86	0, 73	0	0
Ren	513.71	1.43	745, 64	1, 93	798.21	2.08
LiD	95.38	0.26	168, 56	0, 36	0	0
PUD	190.99	0.65	307, 69	0, 88	0	0
Lym	82.58	0.17	41, 17	0, 06	0	0
MCa	171.16	0.33	184, 85	0, 28	0	0
STu	−28.76	−0.21	−9, 20	−0, 06	0	0
Rhe	313.92	2.82	326, 22	2, 39	352.58	2.59
Coa	613.41	0.71	996, 15	0, 95	0	0
Obe	1253.28	1.01	1948, 63	1, 35	0	0
WeL	61.93	0.12	32, 80	0, 05	0	0
FED	354.99	1.83	496, 26	2, 09	521.40	2.21
BLA	1846.60	3.24	2270, 21	3, 98	2281.51	4.03
DeA	−61.26	−0.30	−21, 97	−0, 09	0	0
Alc	1023.23	5.57	1218, 61	5, 78	1297.83	6.27
Dru	568.81	0.98	897, 39	1, 33	0	0
Psy	619.72	3.90	730, 95	4, 23	747.53	4.39
Dep	34.08	0.21	56, 95	0, 32	0	0
Pne	178.14	1.40	230, 52	1, 69	0	0
UTI	−110.77	−0.78	−145, 45	−0, 90	0	0
Inj	142.79	1.63	110, 72	1, 25	0	0
Hyt	369.91	4.58	396, 62	4, 87	405.76	5.00
Dia	407.89	3.74	393, 08	3, 26	407.41	3.40

Table 3. Explanation of the auxiliary regressors.

Variable	Explanation	Variable	Explanation
HOSP90	days spent in hospital before the fracture	Lym	Lymphoma*
AGE	age of the patient	MCa	Metastatic cancer*
HOSPDUM	hospitalized during 90 days before the fracture*	STu	Solid tumor without metastasis*
OPWAIT	waited for operation over 2 days*	Rhe	Rheumatoid arthritis*
FEMALE	patient is a female*	Coa	Coagulopathy*
CHF	Congestive heart failure*	Obe	Obesity*
Arr	Cardiac arrhythmias*	WeL	Weigth loss*
Val	Valvular disease*	FED	Fluid and electrolyte disorders*
PCD	Pulmonary circulation disorders*	BLA	Blood loss anemia*
PVD	Peripheral vascular disorders*	DeA	Deficiency anemia*
Par	Paralysis*	Alc	Alcohol abuse*
PaD	Parkinson disease*	Dru	Drug abuse*
Dem	Dementia*	Psy	Psychoses*
OND	Other neurological disorders*	Dep	Depression*
CPD	Chronic pulmonary disorders*	Pne	Pneumonia*
Hyp	Hypothyroidism*	UTI	Urinary tract infection*
Ren	Renal failure*	Inj	Injuries*
LiD	Liver disease*	Hyt	Hypertension*
PUD	Peptic ulcer disease*	Dib	Diabetes*

dummy variables are marked with *

7. ACKNOWLEDGMENTS

The work of Reijo Sund was financially supported by the Yrjö Jahnsson Foundation (grant 5978).

8. REFERENCES

[1] Buckland, S. T. Burnham, K. P. and Augustin, N. H. (1999). Model Selection: An Integral Part of Inference. *Biometrics*, 53, 603–618.

[2] Danilov, D. (2005). Estimation of the mean of a univariate normal distribution when the variance is not known. *Econometrics Journal* 8, 277–291.

[3] Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27–46.

[4] Haentjens, P., Autier, P., Barette, M., Boonen, S. and Belgian Hip Fracture Study Group (2001) The economic cost of hip fractures among elderly women. A one-year, prospective, observational cohort study with matched-pair analysis. *J Bone Joint Surg Am*, 83-A, 493–500.

[5] Judge, G. G., and Bock, M. E. Bock *The statistical implications of pre-test and Stein-rule estimators in econometrics*, Amsterdam, North-Holland.

[6] Leeb, H. and Pötcher, B. M. (2005) Model selection and inference: facts and fiction. *Econometric Theory*, 21, 21–59.

[7] Liski, E. P. and Trenkler, G. (1993). MSE-Improvement of the Least Squares Estimator by Dropping Variables. *Metrika* 40, 263–269.

[8] Liski, E. P. and Liski, A. (2008). MDL Model Averaging for Linear Regression. In: Grñwald, P., Myllymäki, P., Tabus, I., Weinberger, M., and Yu, B. (Eds.). *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, 145-154. Tampere, Tampere International Center for Signal Processing.

[9] Magnus, J. R. (1999). The traditional pretest estimator. *Theory of Probability and Its Applications*, 44, 293–308.

[10] Magnus, J. R. (2002). Estimation of the mean of a univariate normal distribution with a known variance. *Econometrics Journal*, 5, 225–236.

[11] Magnus, J. R. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, 67, 639–643.

[12] Magnus, J. R., Powell, O. and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154, 139–153.

[13] Miller, A. (2002) *Subset selection in regression*, 2rd ed. London, Chapman & Hall/CRC.

[14] Rao, C. R., Toutenburg, H., Shalabh and Heuman, C. (2008). *Linear Models and Generalizations. Least squares and Alternatives*, 3rd ed. Berlin, Springer.

[15] Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis*, 2nd ed. New York, Wiley.

[16] Sund, R., Juntunen, M., Lüthje, P., Huusko, T., Mäkelä, M., Linna, M., Liski, A., Häkkinen, U.(2006). *PERFECT - Hip Fracture, Performance, Effectiveness and Cost of Hip Fracture Treatment Episodes* (In Finnish), National Research and Development Centre for Welfare and Health, Helsinki.

[17] Sund, R. (2007) Utilization of routinely collected administrative data in monitoring the incidence of aging dependent hip fracture. *Epidemiologic Perspectives & Innovations*, 2007, 4:2. http://www.epi-perspectives.com/content/4/1/2

[18] Toro-Vizcarrondo, C. and Wallace, W. D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association*, 63, pp. 558–572.