Variance Estimates and Model Selection

Sıdıka Başçı® **Asad Zaman** **Arzdar Kiracı**

SESRIC International Islamic University of Islamabad Başkent University

ABSTRACT

The large majority of the criteria for model selection are functions of the usual variance estimate for a regression model. The validity of the usual variance estimate depends on some assumptions, most critically the validity of the model being estimated. This is often violated in model selection contexts, where model search takes place over invalid models. A cross validated variance estimate is more robust to specification errors (see, for example, Efron, 1983). We consider the effects of replacing the usual variance estimate by a cross validated variance estimate, namely, the Prediction Sum of Squares (*PRESS*) in the functions of several model selection criteria. Such replacements improve the probability of finding the true model, at least in large samples.

Key words: *Autoregressive Process, Lag Order Determination, Model Selection Criteria, Cross Validation*
JEL Classifications: C13, C15, C22, C52

1. INTRODUCTION

In applied work, model selection is a frequently occurring problem of great importance, as forecasts, conclusions, interpretations, etc. frequently depend critically on the particular model selected from the range of models examined. Most often, model selection is done by mechanical application of one or several of the criteria that have been developed for this purpose[1]. The large majority of these criteria assess regression models using a function of the usual estimate of error variance and the model dimension. Different criteria are based on different functions, but all use the usual variance estimate, $\hat{\sigma}^2$. The usual estimate is valid only if the model is correctly specified, and this assumption is especially dangerous in model search situations where we will inevitably search over incorrectly specified models. Efron

® Sıdıka Başçı, The Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC), Department of Statistics, Attar Sokak, No:4, 06700, Gaziosmanpasa, Ankara, Turkey, (email: sbasci@sesric.org), Tel. + 90312 4686172/308, Fax: + 90312 4673458.
Asad Zaman, International Institute of Islamic Economic, International Islamic University of Islamabad, (email: asadzaman@alum.mit.edu), Tel. +9251 9257939, Fax: + 9251 9258019.
Arzdar Kiracı, Başkent University, Department of Economics, 06533 Baglıca, Ankara, Turkey.
Earlier versions of this paper were presented at 1998 European Conferences of the Econometrics Community, Stockholm, 1998 International Symposium on forecasting, Edinburgh, 1997 Computing in Economics and Finance, Stanford and also at seminars given at Bilkent University, University of York and Ohio State University. We would like to thank to the participants. We also would like to thank to Karim Abadir, Michael P. Clements and Lutz Killian for their valuable comments.
[1] Some authors, such as Amemiya (1980), or Judge et al. (1985), have argued against such mechanical model selection, in favor of a theory-based approach. Hendry (1995) argues in favor of systematic model simplification starting form a model complicated enough to nest all possibilities. In situations where forecasts are of interest, it is also possible to use forecast combination and avoid selection; see Zaman (1984) and Diebold (1989) for discussion and further references. In this paper, we will ignore these alternatives.

(1983) finds that a cross validated (CV) variance estimate $\tilde{\sigma}^2$ is more robust to specification errors. Because of this, we may expect that the performance of model selection criteria can be improved by replacing the usual variance estimate by a CV variance estimate in their functions. Another motivation for trying $\tilde{\sigma}^2$ instead of $\hat{\sigma}^2$ comes from noting that the CV residual r_i' can be computed as $r_i' = r_i/(1 - h_{ii})$, where r_i is the OLS residual and h_{ii} is the i-th entry of the hat matrix $X(X'X)^{-1}X'$. Thus CV replaces the OLS residual by the 'almost unbiased' residual suggested by Horn, Horn, and Duncan (1975). MacKinnon and White (1985) found that this replacement substantially improves heteroskedasticity consistent covariance matrix estimates. Our results show that model selection criteria are similarly improved by this replacement.

In this paper we consider Autoregressive (AR) models so the problem of model selection becomes the problem of choosing the lag order. We compared model selection criteria having the form $f(\hat{\sigma}^2)$ with $f(\tilde{\sigma}^2)$, replacing $\hat{\sigma}^2$ by $\tilde{\sigma}^2$. In comparing the two forms, we consider the probability of selecting the true model. Since results depend on the sample size used and the value of the regression coefficients we present results for different sample sizes and different coefficient values. In this paper, we study replacing the usual variance estimate by a CV variance estimate in the functions of several popular model selection criteria. The criteria used for this aim are Akaike Information Criterion (*AIC*; Akaike, 1973; Akaike, 1974), Schwarz Criterion (*SC*; Schwarz, 1978; Rissanen 1978), Hannan-Quinn Criterion (*HQC*; Hannan and Quinn, 1979; Quinn, 1980) and a bias corrected version of *AIC* presented in Hurvich and Tsai (1989) which is denoted as AIC_C.

Our Monte Carlo results show that the probabilities of estimating the true model where CV variance estimate is used in the functions of criteria are better for large sample sizes. Also when a large value of coefficient is chosen for the highest order of the true model, using a CV variance estimate is better. The highest improvement from the replacement is obtained when it takes place in the function of AIC_C. When we consider the probabilities of overestimation, we see that criteria containing CV variance estimate rather than the usual variance estimate in their functions are more parsimonious. In section 2, we describe CV estimate of variance in detail. Section 3 presents the model that we base our Monte Carlo study. In section 4 we have the simulation results where the probability of estimating the lag order is under consideration. Finally in section 5, we have some concluding remarks.

2. CROSS-VALIDATED VARIANCE ESTIMATE

Efron (1983) shows that the error rate of a predictive rule is underestimated if the same data used to both construct and to evaluate the rule. The residual $r_i = y_i - \hat{y}_i$ underestimates the true error at i since the i-th observation has been used in fitting the equation[2]. One way to reduce the problem is to use $r_i' = y_i - \tilde{y}_i$, where $\tilde{y}_i$ is the forecast of y_i based on a regression which excludes the i-th observation (namely jackknifing). This procedure is described as the LOO (leave one out) CV method (Rao and Wu, 2001), as the *predictive residual* by Allen (1974) or simply as Cross-validation (Efron, 1983; Li, 1987). Allen names the sum of squares based on these residuals the Prediction Sum of Squares (*PRESS*) and suggests it as a basis for model selection. Allen's (1974) *PRESS* is equivalent to CV (Rao and Wu, 2001).

[2] One way to see this is to note that RSS($\hat{\beta}$) < RSS(β) - the residual sum of squares is minimized by $\hat{\beta}$ so that it must be smaller than the true residual sum squares based on the true parameter β.

In Arlot and Celisse (2010) it is stated that as T (sample size) tends to infinity the bias of LOO stays of order T^{-1} and is generally minimal compared with V-fold CV and bootstrap (Davison and Hall, 1992; Molinaro et al., 2005). Shao (1993) showed that minimizing the LOO CV estimate for multiple linear regression (MLR) lead to a statistically inconsistent choice of the true model. With large sample sizes, LOO CV identifies the variable subset belonging to the true model, but it also selects additional variables. This means that minimizing the LOO CV estimate results in overfitting and thus in a larger prediction error. However, Li (1987) showed that under some conditions, the LOO CV is consistent and is asymptotically optimal in some sense. According to Linhart and Zucchini (1986) CV provides a technique for developing an estimator of an expected discrepancy which need not be bias adjusted. Another argument for the small variance of LOO in regression was provided by Davies et al. (2005), with the log-likelihood contrast: assuming a well specified parametric model is available, the LOO estimator of the risk is the minimum variance unbiased estimator of its expectation.

We will define the cross-validate variance estimate as $\tilde{\sigma}^2 = (T-K)^{-1}\sum_{t=1}^{T} r_i'^2 = PRESS/(T-K)$. Amemiya (1980) shows that $r_i' = r_i/(1-h_{ii})$, where r_i is the i-th OLS residual, and h_{ii} is the i-th diagonal entry of the hat matrix $X(X'X)^{-1}X'$. Thus the predictive residuals are equivalent to the nearly unbiased residuals of Horn, Horn, and Duncan (1975).

Hurvich and Tsai (1989) include *PRESS* in their Monte Carlo study where they compare finite sample properties of several different model selection criteria for regression models. AIC_C which is a bias corrected version of *AIC* suggested by the authors turns out to be the best criterion and performance of *PRESS* and other criteria are close to each other. Başçı (1998) shows that *PRESS* performs poorly because of its failure to penalize higher dimensional models. Magee and Veall (1990) have also considered and compared the use of *PRESS* and also White's *t*-statistics in model selection. Magee and Veall (1990) show that the *PRESS* and the White's *t*-statistic approximate each other. Li and Hui (2007) and Lang et al. (2007) use *PRESS* for selecting predictors using stepwise forward variable selection method to optimize the outcome prediction. In Billings and Wei (2008) a new adaptive orthogonal search algorithm is proposed for model subset selection and non-linear system identification, where the adjustable prediction error sum of squares (*APRESS*) is introduced and incorporated into a forward orthogonal search procedure. Christopher et al. (1998) or Peng and Wang (2007) add that the lower the difference between the *PRESS* value and the regression's sum square of error value, the more stable the model's predictive power. Özkale and Kaçıranlar (2007) propose and investigate *PRESS* statistic for selecting the biasing parameter d in Liu (1993) estimator. There are many examples that use *PRESS* in their analysis, for example, Xinjun (2010), Jabri et al. (2010), Xiongcai and Sowmya (2009), Nikolic and Agababa (2009), Neri (2009).

In Piepho and Gauch (2001), a simulation study is conducted to study the merits of *AIC*, AIC_C, AIC_U (McQuarrie and Tsai, 1998), *SC*, *HQC*, HQ_C (McQuarrie and Tsai, 1998), *FPE* (final prediction error; Akaike, 1973,), FPE_U (McQuarrie and Tsai, 1998), FPE_4 (Bhansali and Downham, 1977), R_P (Breiman and Freedman, 1983), C_P (Mallows, 1973), *GM* (Geweke and Meese, 1981) and *PRESS* for marker pair selection that uses model selection criteria for multiple linear regression. On the basis of their results, *PRESS* is not the best criterion for model selection. They note that that there exist several asymptotic equivalence relationships between *FPE* and *PRESS*. Wang and Schaalje (2009) note in their comparison that characteristics of the data, such as the covariance structure, parameter values, and sample size, greatly impacted performance of various model selection criteria. In their conclusion they state that none of *AIC*, *SC*, R^2, *PRESS* was consistently better than the Concordance

Correlation Coefficient (*CCC*) (Lin, 1989) criterion. Collett and Stepniewska (1999) some variable selection procedures used in conjunction with fitting logistic regression models are summarized and their performance investigated using a simulation study. They compare the performances of *AIC*, *SC*, modified C_P, *MD* (mean deviance), $M\chi^2$ (mean χ^2) and *PRESS*. In this simulation the PRESS statistics has values that is sometimes considerably below the values of these performance measures for the other criteria.

As an alternative to the CV variance estimate $\tilde{\sigma}^2$, we could consider the bootstrapped variance estimate, recommended by Efron (1983). Efron shows that bootstrapping generally gives better results than CV. However, he shows that for smooth functions, CV behaves like the bootstrap. Since the sum of squared residuals is a very smooth function, and substantially simpler to compute, we prefer the use of CV to bootstrapping in the present example.

3. MONTE CARLO DESIGN

We describe first the Monte-Carlo design used for our comparison of the *PRESS* criterion with other methods of model selection. We restrict ourselves to the context of selection of the true order in an autoregressive model. Assume that the $T \times 1$ vector of observations Y is generated from an AR(p) process (allowing for nonzero mean a_0):

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + u_t$$

where u_t have a common distribution of F. We assume that there is a maximum possible lag order M. The econometrician wants to estimate lag order p, where p must be between $1,\ldots, M$. In this paper, we concentrate on the model where error terms are generated from a normal distribution. Results for the case where error terms are generated from a skewed distribution can be found in Başçı (1998). There, the criterion *AIC* is under consideration and it is shown that for a skewed distribution case there still exists improvements over *AIC* from substituting the CV estimates of variance into the function of *AIC* but these improvements are less than the improvements that we obtain for normal distribution case presented in this paper. See also Başçı and Zaman (1998) for a study of effects of skewness and kurtosis on model selection criteria.

4. COMPARISONS WITH TRADITIONAL CRITERIA

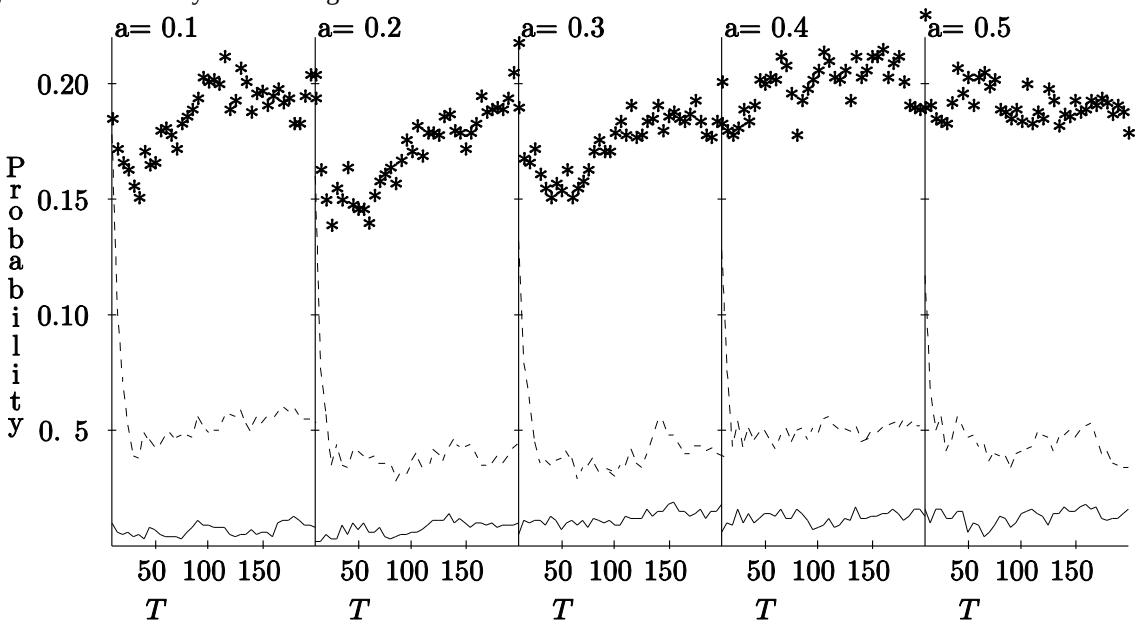
A standard method for model selection is to start with the largest model and drop the highest order insignificant lag (where significance is measured by the t-statistic). The process is repeated until the last lag is significant, and this model is chosen. It is agreed upon that the best model is the one with the smallest residual variance, and for this reason it would be sensible to use $\hat{\sigma}^2$ as a measure for model performance. This one of the elements of Hendry's methodology (see, for example Hendry, 1995), and the standard error of the regression $\hat{\sigma}$ is frequently used as a measure to assess the performance of model.

4.1. Dimension Six

In our first Monte Carlo, we assess the performance of these two traditional criteria and compare them with $\tilde{\sigma}^2$. For the two variance estimators we pick the model yielding the smallest variance. For the t-statistic we pick the largest model for which the highest order lag is significant - assessing significance by the criterion that $t > 2$ is significant. Aside from simplicity, Magee and Veall (1995) show that if the t is based on a heteroskedasticity adjusted covariance estimate, this rule should be asymptotically equivalent to the use of the *PRESS* criterion for model selection. We set $a_1 = a_2 = 0.5$ and vary a_3 from 0.1 to 0.5 in steps of 0.1;

$a_4 = a_5 = a_6 = 0$ and $M = 6$. Figure 4.1 below gives the probability that a model of dimension 6 is selected - in some sense, the probability of the biggest mistake, for the three criteria under study. In each graph, the *y*-axis gives the probability of selecting the model of dimension 6, while the *x*-axis is the sample size, which varies from 10 to 200 in steps of 5.

Figure 4.1 Probability of Selecting Dimension 6



The results from this Monte Carlo were quite surprising to the authors. The OLS variance appears to have an asymptotic probability of about 20%, and small sample probabilities are roughly around this number as well. This is a huge probability of selecting a model which is quite far from and quite easily distinguishable from the true model for a sample size of 200. This clearly explains why, despite its intuitive plausibility, practitioners do not rely on $\hat{\sigma}^2$ for model selection. It has a *huge* probability of overestimating the size of the lag, and is generally known to favor large models. The behavior of the usual *t*-statistic is quite predicable and in accordance with theory. In large samples the event $t > 2$ occurs with probability 4.6% under the null hypothesis. Since the true coefficient of the sixth lag is exactly 0, we expect that the $t > 2$ method will choose the six dimensional model around 4.6% of the time. The observed probabilities are closer to 5% because the dynamic model, and variation in the degrees of freedom with sample size, makes the *t* an approximate rather than exact distribution.

The major surprise was the behavior of the *PRESS*, or $\tilde{\sigma}^2$. Its probabilities of selecting a model of dimension 6 stay comfortably under those of the *t*-statistic (averaging 2.8% over the cases studied), and are radically different from those of $\hat{\sigma}^2$. Since $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ are both convergent to the true σ^2 for the true model asymptotically, we did not expect such a huge difference in performance relative to model selection. Based on these results, we would recommend the routine use of $\tilde{\sigma}^2$ to replace $\hat{\sigma}^2$ in conventional regression statistics. It would be well worth exploring the higher order asymptotics to account for the differences between $\tilde{\sigma}^2$ and $\hat{\sigma}^2$. Based on the Magee and Veall (1995) paper, we expected roughly equivalent performance for the *PRESS* and *t*-statistic, but were surprised to see that *PRESS* handily outperforms the *t*-statistic as well.

4.2. Overestimation Probabilities

We graph below the probabilities of overestimating the dimension of the model. The case of dimension 6 has already been discussed, and the graphs below give the probabilities of selection for dimensions 5 and 4.

Figure 4.2 Probability of Selecting Dimension 5

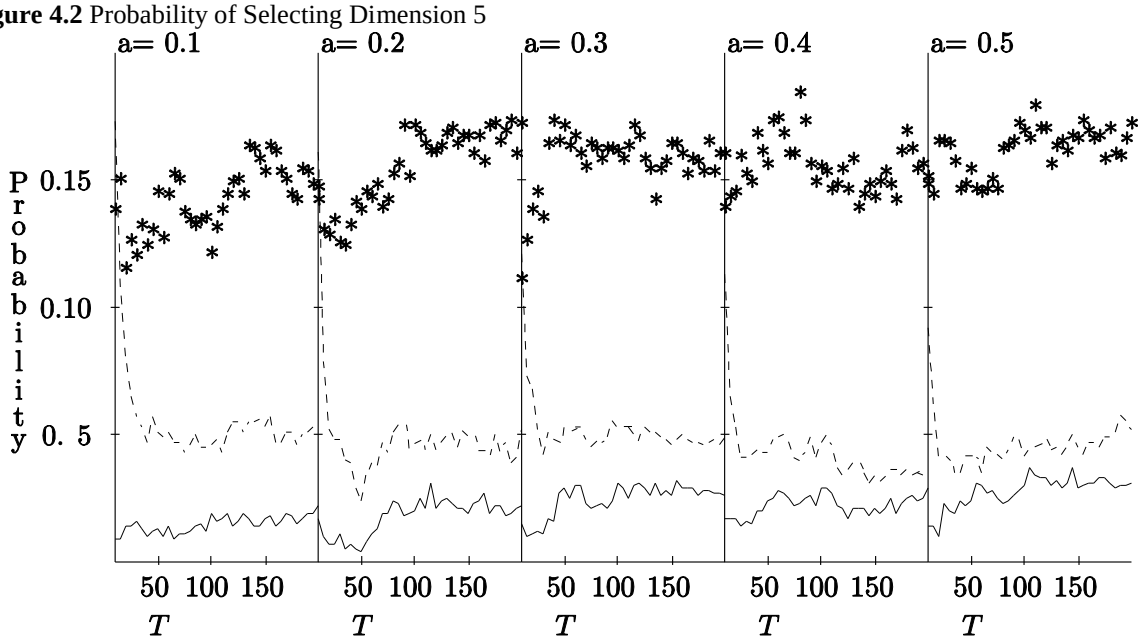
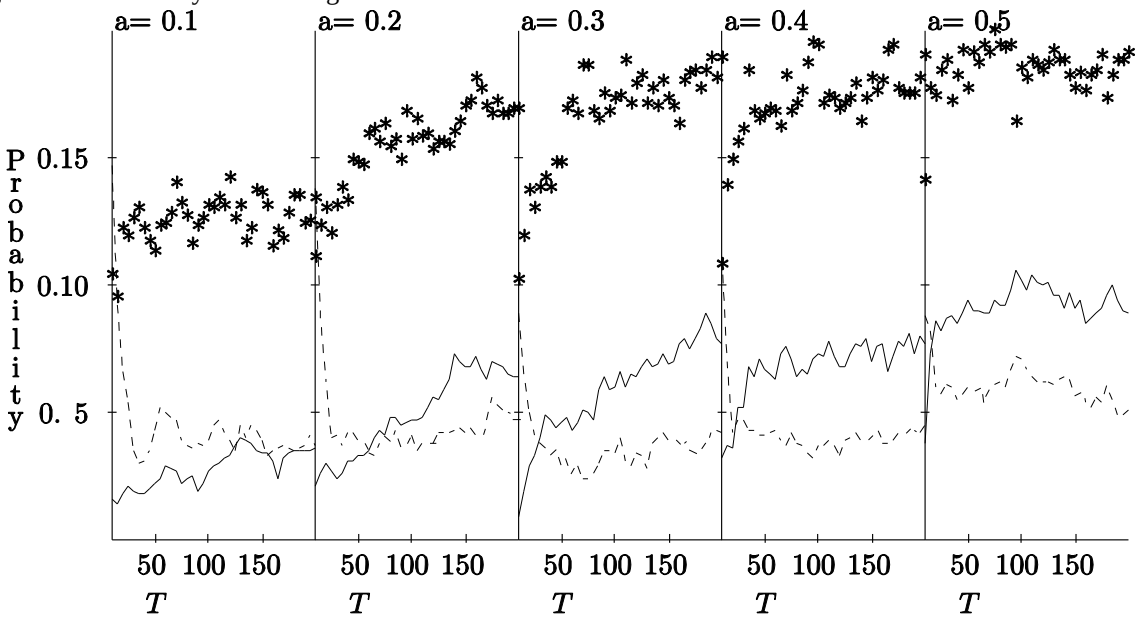


Figure 4.3 Probability of Selecting Dimension 4



The case of dimension 5 is quite similar to dimension 6. The usual variance estimate $\hat{\sigma}^2$ selects this model with probabilities around 16%, substantially worse than the 4% selection probabilities for the $t > 2$ rule. However the best performance is put in by the *PRESS* variance $\tilde{\sigma}^2$, with probabilities near 3% .

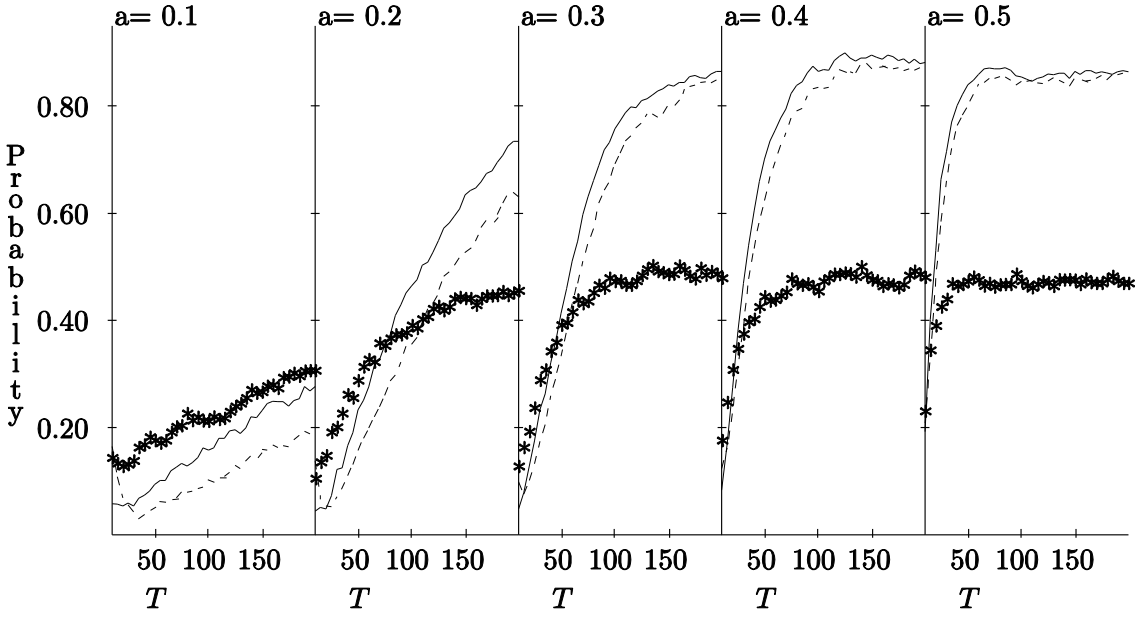
The case of dimension 4, which is one more than the true dimension, leads to a deterioration in the performance of *PRESS*. It now selects this model with probabilities nearing 8% , more than the 4.5% achieved by the $t > 2$ rule. There is an exception to this when a_3 is small,

reduces the probability selection dimension 4 below that of *t*. Heuristically, we could say that for models of dimensions two or more over the true dimension, $\tilde{\sigma}^2$ is better than $t > 2$. When a_3 is small, the 4 dimensional model comes close to being two over the true dimension and hence $\tilde{\sigma}^2$ has better performance. Note that the value of a_2 should have little or no effect on the performance of the $t > 2$ rule since the *t* rule tests the significance of a_4 which is exactly 0. As before, the standard variance is hopeless in comparison to these two, having probabilities around 16%.

4.3. Correct Dimension Estimation Probabilities

Figure 4.4 gives the probabilities of selecting the correct dimension for the three criteria under study. Generally the performance of $\tilde{\sigma}^2$ and the $t > 2$ rules are similar, with the former being slightly superior, over the range of situations studied. Generally, in larger samples and for larger values of a_3, $\hat{\sigma}^2$ is substantially inferior. However, in small samples and with small values of a_3, $\hat{\sigma}^2$ can be superior to the other rules. This does not recommend $\hat{\sigma}^2$ to us, since in such situations the probability of finding the true model is low anyway.

Figure 4.4 Probability of Selecting Dimension 3



4.4. Underestimation Probabilities

For models of dimensions 1 and 2, both less than the true model, the probabilities of selection go to zero for all three criteria under study. For $T > 75$ the probabilities were close enough to zero for all three that the graphs in Figures 4.5 and 4.6 have been truncated at $T = 75$. Generally the probabilities of underestimation decline to zero rapidly for all three, and the performance of the *PRESS* variance $\tilde{\sigma}^2$ is similar to the $t > 2$ rule. The usual variance estimate $\hat{\sigma}^2$ compensated for it tendency to overestimate by have smaller probabilities of underestimation than the other two rules.

4.5. Large Sample Probabilities

While $\tilde{\sigma}^2$ and the $t > 2$ rule have similar performance, they do not appear to be asymptotically equivalent. To test whether the Magee and Veall (1995) equivalence holds, we tried replacing

the standard *t* statistics with the (asymptotically equivalent) White *t*-statistics, but found only trivial differences in their performances.

Figure 4.5 Probability of Selecting Dimension 2

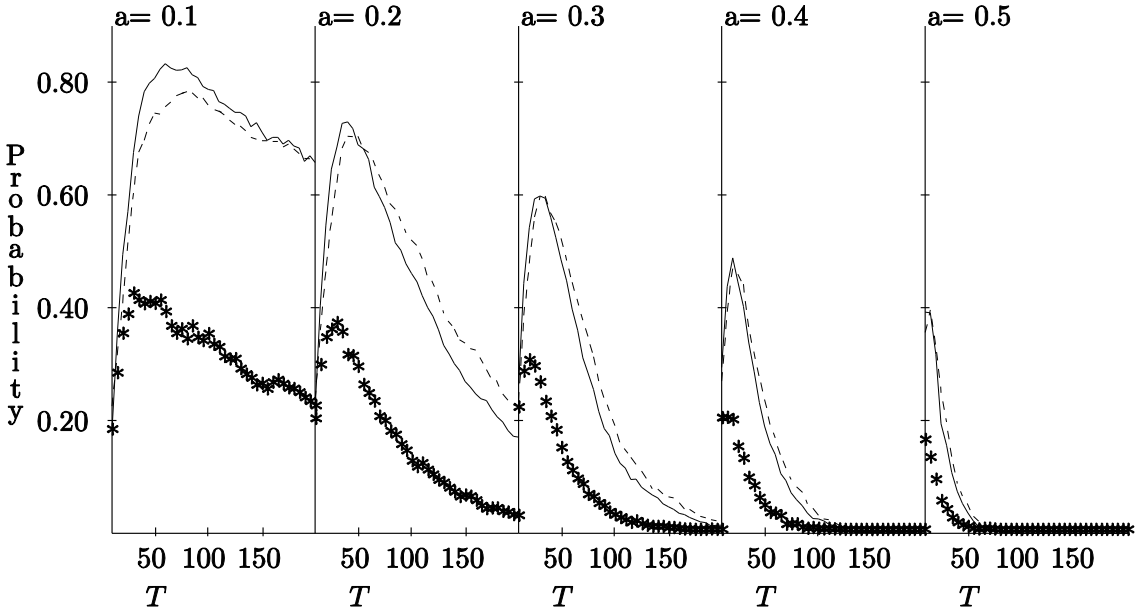
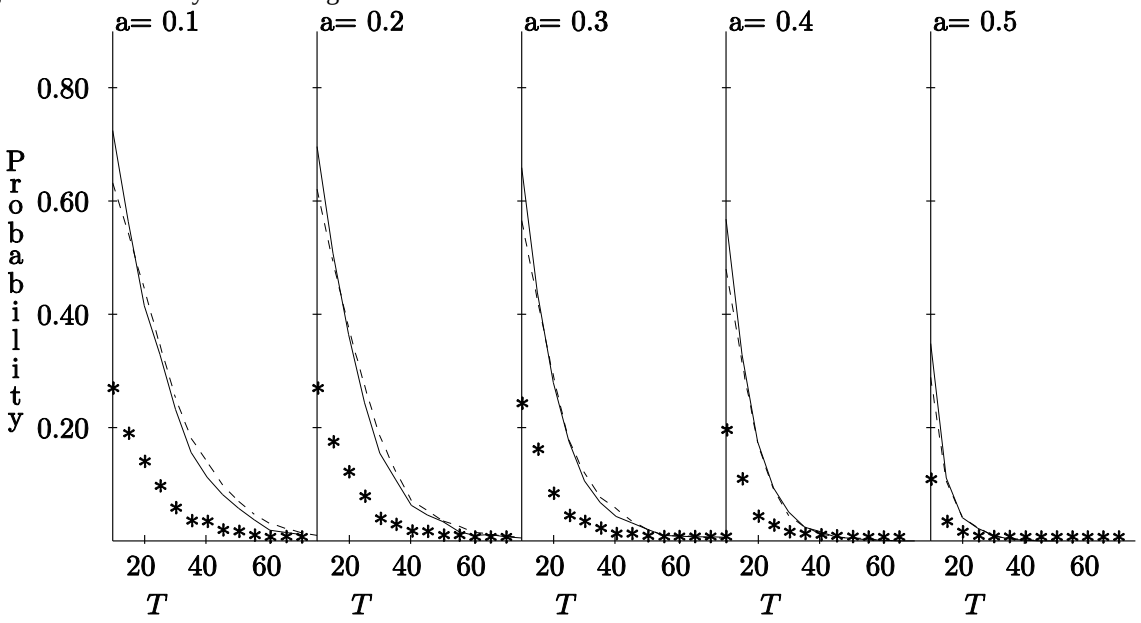


Figure 4.6 Probability of Selecting Dimension 1



The standard *t*-statistics are constructions from the covariance estimate $\hat{\sigma}^2(X'X)^{-1}$, the White *t*-statistics use the covariance estimate $(X'X)(X'DX)^{-1}(X'X)$, where D is a diagonal matrix of squared OLS residuals - better results are obtained by replacing OLS residuals e_t by the Horn, Horn and Duncan almost unbiased residuals $e_t' = e_t/(1 - h_t)$. We also tested the use of $\tilde{\sigma}^2(X'X)^{-1}$ as a basis for the *t*-statistics. However, all variant forms of the *t*-statistics gave essentially the same result, with around 4.6% probability of selecting the model of dimension 6. The cross-validated variance estimate performs significantly better than all variants of *t*-statistics which we tried.

To get a better picture of asymptotics, we did a few large samples, ending up with the following probabilities displayed in Table 4.1. In large samples, *none* of the three criteria underestimates the model. The conventional variance estimate $\hat{\sigma}^2$ is hopeless, with only a 47% probability of estimating the true model, and a 21% probability of estimating a model of dimension 6. While the rule of choosing $t > 2$ and the *PRESS* variance estimate $\tilde{\sigma}^2$ have similar probabilities of choosing the right model, there is an important and interesting difference in overestimation probabilities. Since the null holds for models higher than the true dimension, the probability of rejecting the null in *each* of the higher dimensions is about the same, around 4.6%. However, use of $\tilde{\sigma}^2$ leads to a rule which appears to be consistent - the probability appears to decline to zero for dimension 6, and is headed that way for dimension 5. To compensate, dimension 4 (one more than the true dimension) is estimated to be the true model more often by $\tilde{\sigma}^2$ relative to the $t > 2$ rule. It is worth noting that the pattern of overestimation probabilities of $\tilde{\sigma}^2$ is much preferable to that of the $t > 2$ rule - when the wrong model is selected it is helpful if it is only slightly bigger than the true model. The $t > 2$ rule picks all three larger models with roughly equal probabilities.

Dim=	1	2	3*	4	5	6
$\hat{\sigma}^2$	0.00	0.00	0.47	0.16	0.16	0.21
$\tilde{\sigma}^2$	0.00	0.00	0.89	0.08	0.03	0.01
$t > 2$	0.00	0.00	0.88	0.04	0.04	0.05

Table 4.1 $T = 400$ Model Selection Probabilities.

Since it appears clearly desirable to use a consistent rule, the Monte Carlo study leads us to prefer $\tilde{\sigma}^2$ to the $t > 2$ rule. If *t*-statistics are to be used, we should avoid a mechanical fixed significance level, as it leads to an inconsistent rule for model selection. It is possible to devise schemes for changing significance levels with sample size so as to achieve consistency in large samples. It is clear that the Magee and Veall (1995) asymptotics do not hold in this model. Indeed it is easy to establish that there local-to-zero assumption is not valid for the Monte Carlo setup we describe. If a_2 is sent to zero in a suitable way we could recover the Magee and Veall asymptotics.

5. IMPROVING ON *PRESS*

We have demonstrated that the *PRESS* variance $\tilde{\sigma}^2$ is substantially superior to the conventional variance estimate $\hat{\sigma}^2$ and somewhat superior to the convential $t > 2$ rule for model selections. On this basis, it would clearly be worthwhile to include $\tilde{\sigma}^2$ (indeed, even replace $\hat{\sigma}^2$) in statistics on a conventional regression printout. Nonetheless, the performance of the *PRESS* variance estimate is not satisfactory from an absolute point-of-view. Achieving only 89% probabilities of selecting the correct model in large samples, where this probability should be close to 100% is not quite satisfactory. Since the underestimation probabilities converge to zero, we conclude that lack of consistency of *PRESS* is cause by overestimation - the probability of selecting a model of dimension larger than the true model is too large. To reduce this problem, we should penalize the choice of higher dimensional models more heavily than is done by *PRESS*.

How should we select a penalty factor to improve the performance of *PRESS*? Nearly all model selection criteria (*AIC*, BIC, Schwartz, etc.) are based on adding a dimension penalty to the usual estimate of the variance. Since *PRESS* is also a variance estimate, it seems logical to try these various penalties in the hope of improving the performance of the *PRESS*. Lütkepohl (1985) compares the finite sample performances of 12 different identification approaches for

AR models in a Monte-Carlo study. In his study, *SC* and *HQC* emerge as being the best among the compared criteria. *AIC*, also performs well. For this reason we consider these criteria in our study. Also these criteria are very popular among the practitioners. *SC*, *HQC* and *AIC* have similar functional forms. They all include the logarithm of the usual variance estimate but they add to it different linear penalty factors. AIC_C suggested by Hurvich and Tsai (1989) also contains logarithm of the usual variance estimate but it contains a nonlinear penalty factor. We also consider AIC_C in our study to see the effect of a nonlinear penalty factor. In Başçı and Zaman (1998) the performance of this criterion in terms of probability of estimating the true model is studied for normal variables. The results there show that AIC_C is the best criterion for small samples. Given a model selection criterion of the form $MSC(\hat{\sigma}^2)$, we define the cross-validated form *MSCCV* as $MSCCV=MSC(\tilde{\sigma}^2)$. This substitution yields the following four new model selection criteria:

$$AICCV(k) = \ln \tilde{\sigma}_k^2 + (2k)/T$$

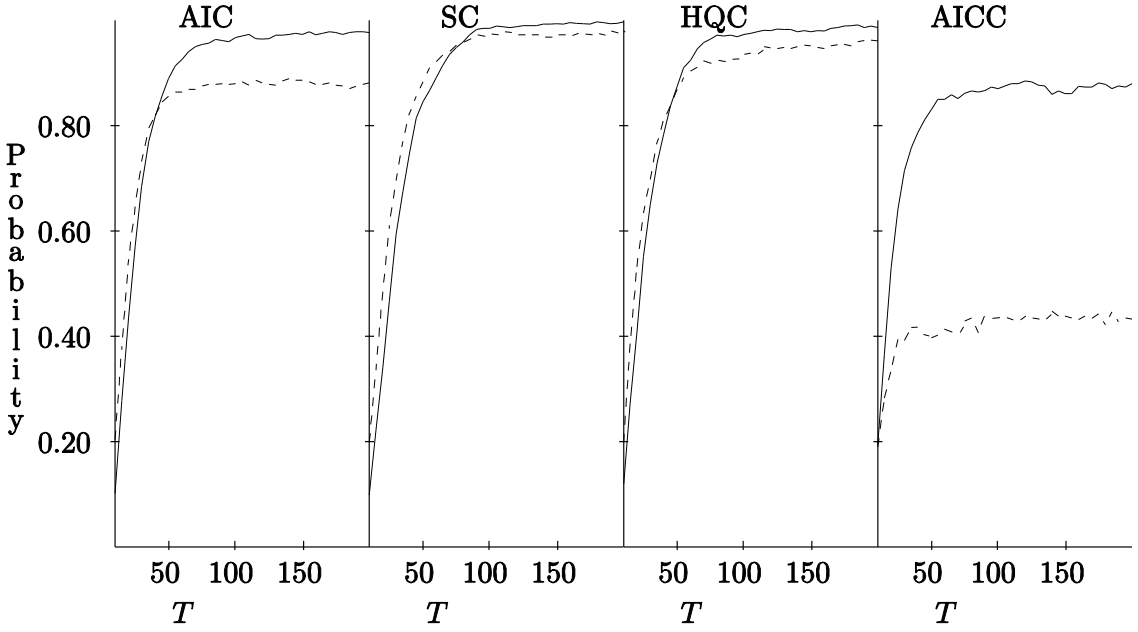
$$SCCV(k) = \ln \tilde{\sigma}_k^2 + k \ln(T)/T$$

$$HQCCV(k) = \ln \tilde{\sigma}_k^2 + 2k \ln(\ln(T))/T$$

$$AIC_CCV(k) = T \ln \tilde{\sigma}_k^2 + T\,[1+(k/T)]/[1-(k+2)/T]$$

Our hope is to eliminate or reduce the overestimation problem of $\tilde{\sigma}^2$ by incorporating these penalties. In addition, we hope to check whether the model selection criteria can be improved by replace the conventional variance estimator with the *PRESS* estimator. Comparisons are made of the probability of selecting the true model and also of the relative forecasting performance of the models selected by the different criteria.

Figure 5.7 Probability of Selecting Correct Dimension



5.1. Correct Dimension Probabilities

Figure 5.7 shows the effects of modifying the four model selection criteria on the probability of estimating the correct dimension (namely 3) in the setup already described earlier. For all four criteria, the probability of selecting the correct dimension are improved in large samples. It is well-known that the *AIC* is inconsistent -- the graph shows that the *AIC* probability is converging to about 89% . It is quite surprising that replacing $\hat{\sigma}^2$ by $\tilde{\sigma}^2$ changes the *AIC* into

a consistent criterion. Large sample gains from the substitution are around 10%, and the probability of selecting the correct model appears convergent to unity. The Schwartz Criterion is consistent to start with. It is also improved by the substitution in large samples but the difference is quite minor - around 1% typically. For the *HQC*, which is also consistent, the gains are around 3% in large samples. The bias-corrected *AIC*, labelled AIC_C, improves the most by the substitution. Its probabilities of selecting the correct model almost double, from 44% to 88%. This is mainly because the AIC_C appears to be quite poor compared to the other criteria. Nonetheless, it is quite surprising the simply replacing $\hat{\sigma}^2$ by the asymptotically equivalent *PRESS* variance estimate improves all the model selection criteria, sometimes substantially.

It is important to note that these improvements are only available in large samples, for sufficiently large values of a_3. Furthermore, the improvements are minor for the best of the four criteria, namely the SC. In small samples, and for small values of a_3, there can be substantial loss (up to 20% in some cases) in the probability of correctly estimating the dimension. We would conclude from this that one should use the unmodified Schwartz criterion for model selection if the probability of correctly selecting the true model is the major goal. We can never tell if the coefficient of the highest lag is exactly zero, or merely close to zero. In the latter case, the modified Schwartz (using $\tilde{\sigma}^2$ instead of $\hat{\sigma}^2$) can be substantially inferior to the unmodified form, while in the former case, the modified version yields only trivial gains over unmodified form.

Figure 5.8 Probability of Overestimating Dimension

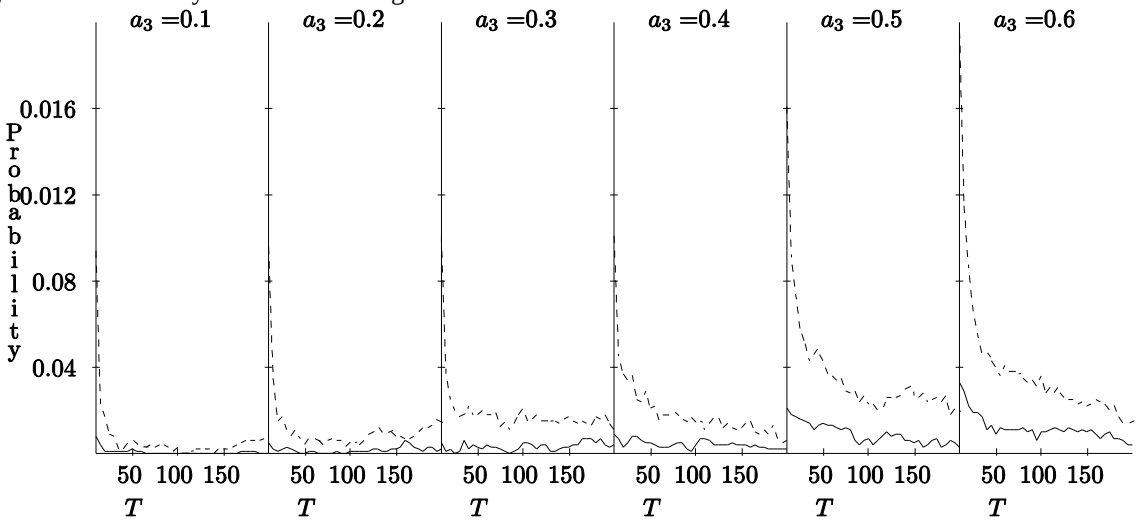


Figure 5.8 gives the 'overestimation' probabilities for both the Schwartz criterion and the modified Schwartz criterion based on the *PRESS* variance. It reveals that the overestimation probabilities are substantially smaller for the modified Schwartz in all cases considered - small and large a_3 as well as small and large sample sizes. It is well-known that forecasting quality does not correlate very well with probabilities of correct estimation; see Başçı (1998) for some discussion and references. There is an interesting tension between model selection required in finding the true model and model selection for forecasting; see for example Diebold (1989) for discussion and references. For the purposes of model selection as required by Hendry's methodology, it is better to choose a model which is too big (and hence nests the true model) rather than too small (and hence misspecified). For forecasting, the reverse holds. Extraneous regressors reduce precision of estimates and lead to poor forecasts - the bias introduced by dropping regressors with small coefficients tends to be small in comparison.

This suggests that the modified Schwartz criterion may be superior to the original one on forecast quality. This issue can be examined in another work.

6. CONCLUSION

In this paper, we studied using CV estimate of variance, $\tilde{\sigma}^2$, instead of the usual estimate of variance, $\hat{\sigma}^2$, in the context of model selection problem. Also a comparison with $t > 2$ criteria takes place. Specifically, we used an Autoregressive (AR) model so the problem of model selection became the problem of lag order determination. In our simulation study we investigated the probabilities of selecting higher dimension, true dimension and lower dimension. The results obtained can be summarized as follows:

1. For dimension six, the usual estimate of variance, $\hat{\sigma}^2$, appears to have an asymptotic probability of choosing this dimension about 20%, and small sample probabilities are roughly around this number as well. It has a *huge* probability of overestimating the size of the lag. The behavior of the usual *t*-statistic is quite predicable and in accordance with theory. In large samples the event $t > 2$ occurs with probability 4.6% under the null hypothesis. The major surprise was the behavior of the *PRESS*, or $\tilde{\sigma}^2$. Its probabilities of selecting a model of dimension 6 stay comfortably under those of the *t*-statistic (averaging 2.8% over the cases studied), and are radically different from those of $\hat{\sigma}^2$.

2. The case of dimension 5 is quite similar to dimension 6. The usual variance estimate $\hat{\sigma}^2$ selects this model with probabilities around 16%, substantially worse than the 4% selection probabilities for the $t > 2$ rule. However the best performance is put in by the *PRESS* variance $\tilde{\sigma}^2$, with probabilities near 3%.

3. The case of dimension 4, which is one more than the true dimension, leads to deterioration in the performance of *PRESS*. It now selects this model with probabilities nearing 8%, more than the 4.5% achieved by the $t > 2$ rule. There is an exception to this when a_3 is small, reduces the probability selection dimension 4 below that of *t.* Heuristically, we could say that for models of dimensions two or more over the true dimension, $\tilde{\sigma}^2$ is better than $t > 2$. As before, the standard variance is hopeless in comparison to these two, having probabilities around 16%.

4. For the case of selecting the correct dimension for the three criteria under study, generally the performance of $\tilde{\sigma}^2$ and the $t > 2$ rules are similar, with the former being slightly superior, over the range of situations studied. Generally, in larger samples and for larger values of a_3, $\hat{\sigma}^2$ is substantially inferior. However, in small samples and with small values of a_3, $\hat{\sigma}^2$ can be superior to the other rules. This does not recommend $\hat{\sigma}^2$ to us, since in such situations the probability of finding the true model is low anyway.

5. For models of dimensions 1 and 2, both less than the true model, the probabilities of selection go to zero for all three criteria under study.

6. In large samples, *none* of the three criteria underestimates the model. The conventional variance estimate $\hat{\sigma}^2$ is hopeless, with only a 47% probability of estimating the true model, and a 21% probability of estimating a model of dimension 6. While the rule of choosing $t > 2$ and the *PRESS* variance estimate $\tilde{\sigma}^2$ have similar

probabilities of choosing the right model, there is an important and interesting deference in overestimation probabilities. Since the null holds for models higher than the true dimension, the probability of rejecting the null in *each* of the higher dimensions is about the same, around 4.6%. However, use of $\tilde{\sigma}^2$ leads to a rule which appears to be consistent – the probability appears to decline to zero for dimension 6, and is headed that way for dimension 5. Since it appears clearly desirable to use a consistent rule, the Monte Carlo study leads us to prefer $\tilde{\sigma}^2$ to the $t > 2$ rule.

Still, the performance of the *PRESS* variance estimate is not satisfactory from an absolute point-of-view. Achieving only 89% probabilities of selecting the correct model in large samples, where this probability should be close to 100%, is not quite satisfactory. Since the underestimation probabilities converge to zero, we conclude that lack of consistency of *PRESS* is cause by overestimation – the probability of selecting a model of dimension larger than the true model is too large. To reduce this problem, we penalized the choice of higher dimensional models more heavily than is done by *PRESS*. As penalized factors, we used the ones involved in the functions of model selection criteria *AIC*, *SC*, *HQC* and AIC_C. The results obtained from modifying the four model selection criteria showed that for all four criteria, the probability of selecting the correct dimension improved in large samples.

It is well-known that the *AIC* is inconsistent. Replacing $\hat{\sigma}^2$ by $\tilde{\sigma}^2$ changes the *AIC* into a consistent criterion. Large sample gains from the substitution are around 10%, and the probability of selecting the correct model appears convergent to unity. The Schwartz (*SC*) and Hannan Quinn (*HQC*) Criteria are consistent. The improvement over these are only 1% and 3%, respectively. For the case of AIC_C probability of selecting the correct model almost double. This is mainly because the AIC_C appears to be quite poor compared to the other criteria.

As a result, we can say that it is quite surprising that simply replacing $\hat{\sigma}^2$ by the asymptotically equivalent *PRESS* variance estimate improves all the model selection criteria, sometimes substantially.

REFERENCES

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In 2*nd International Symposium on Information Theory*, ed. B.N. Petrov and F. Csàki. Budapest: Akadèmiai Kiadò, 267-281

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.

Allen, D.M. (1974). The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics*, 16, 125-7.

Amemiya, T. (1980). Selection of Regressors. *International Economic Review*, 21, 331-345.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79.

Başçı, S. (1998). *Computer Intensive Techniques for Model Selection*. Ph. D. Dissertation, Bilkent University.

Başçı, S. and A. Zaman (1998). Effects of Skewness and Kurtosis on Model Selection Criteria. *Economics Letters*, 59, 17-22.

Başçı, S., M. Orhan, and A. Zaman (1998). Model Selection by Cross Validation: Computational Aspects. *Working Paper*, Bilkent University.

Bhansali, R.J. and D.Y. Downham (1977). Some properties of the order of an autoregressive model selected by a generalized Akaike's EPF criterion. *Biometrika*, 64, 547-551.

Billings, S.A., H.L. Wei (2008). An adaptive orthogonal search algorithm for model subset selection and non-linear system identification. *International Journal of Control*, 81(5), 714-724.

Breiman, L. and D. Freedman (1983) How many variables should be entered in a regression equation? *Journal of the American Statistical Association*. 78, 131-136.

Chatterjee, S. and A.S. Hadi (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, Inc., New York.

Christopher, T.B.S., A.M. Mokhtaruddin, M.H.A. Husni and M.Y. Abdullah (1998). A simple equation to determine the breakdown of individual aggregate size fractions in the wet-sieving method. *Soil & Tillage Research*, 45, 287-297

Collett, D., K. Stepniewska (1999). Some practical issues in binary data analysis. *Statistics in Medicine*, 18(17-18), 2209-2221.

Davies, S.L., A.A. Neath and J.E. Cavanaugh (2005). Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. *Statistical Methodology*, 2(4), 249–266.

Diebold, F.X. (1989). Forecast Combination and Encompassing: Reconciling Two Divergent Literatures. *International Journal of Forecasting*, 5(4), 589-92.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78(382), 316-331.

Geweke, J. and R. Meese (1981). Estimating regression models of finite but unknown order. *International Economic Review*, 22, 55-70.

Hannan E.J. and B.G. Quinn, (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society B*, 41, 190-195.

Hendry, David F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

Horn, S.D., R.A. Horn, and D.B. Duncan (1975). Estimating Heteroskedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380-385.

Hurvich, C.M. and C.L. Tsai (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2), 297-307.

Jabri, M. El, S. Abouelkaram, J.L. Damez and P. Berge (2010). Image analysis study of the perimysial connective network, and its relationship with tenderness and composition of bovine meat. *Journal of Food Engineering*, 96(2), 316-322.

Judge, G.G., W.E. Griffiths, R.C. Hill and T. Lee (1985). *The Theory and Practice of Econometrics 2nd Edition*. John Wiley & Sons, Inc., New York: Wiley Series in Probability and Mathematical Statistics.

Lang, L., S. Hui, G. Pennello, Z. Desta, S. Todd, A. Nguyen, D. Flockhart (2007). Estimating a Positive False Discovery Rate for Variable Selection in Pharmacogenetic Studies. *Journal of Biopharmaceutical Statistics*, 17(5), 883-902.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics, 15(3), 958-975.

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

Li, L. and S. Hui (2007). Positive False Discovery Rate Estimate in Step-Wise Variable Selection. *Communications in Statistics - Simulation and Computation*, 36(6), 1217-1231.

Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics - Simulation and Computation*, 22(2), 393-402.

Linhart, H. and W. Zucchini (1986). *Model Selection*. Wiley, New York.

Lütkepohl, H. (1985). Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process. *Journal of Time Series Analysis*, 6(1), 35-52.

Magee, L. and M.R. Veall (1991) Selecting Regressors for Prediction Using *PRESS* and White T Statistics. *Journal of Business and Economic Statistics*, 9, 91-96.

Mallows, C.L. (1973). Some comments on C_P. *Technometrics*, 15, 661-675.

McQuarrie, A., R. Shumway, C.-L. Tsai (1997). The model selection criterion AICu. *Statistics & Probability Letters*, 34, 285-292.

McQuarrie, A.D.R. and C.-L. Tsai (1998). *Regression and Time Series Model Selection*. World Scientific Publishers, Singapore.

Molinaro, A.M., R. Simon and R.M. Pfeiffer (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15), 3301-3307.

Neri, P. (2009). Nonlinear characterization of a simple process in human vision. *Journal of Vision*, 9(12), 1, 1-29.

Nikolic, K. and D. Agababa (2009). Prediction of hepatic microsomal intrinsic clearance and human clearance values for drugs. *Journal of Molecular Graphics and Modelling*, 28(3), 245-252.

Özkale, M.R. and S. Kaçıranlar (2007). A Prediction-Oriented Criterion for Choosing the Biasing Parameter in Liu Estimation. *Communications in Statistics - Theory and Methods*, 36(10), 1889-1903

Peng, X. and Y. Wang (2007). A normal least squares support vector machine (NLS-SVM) and its learning algorithm. *Neurocomputing*, 72(16-18), 3734-3741.

Piepho, H.-P. and H.G. Jr. Gauch (2001). Marker Pair Selection for Mapping Quantitative Trait Loci, *Genetics*, 157, 433-444.

Rao, C. R, Y. Wu (2001). On model selection. With discussion by Sadanori Konishi and Rahul Mukerjee and a rejoinder by the authors. IMS Lecture Notes - Monograph Series, Model selectionh h (38), , 1-64..

Quinn, B.G. (1980). Order Determination for a Multivariate Autoregression. *Journal of the Royal Statistical Society B*, 42, 182-185.

Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, 14, 465-471.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6, 461-464.

Shao J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88, 486-494.

Wang, J. and G.B. Schaalje (2009). Model Selection for Linear Mixed Models Using Predictive Criteria. *Communications in Statistics - Simulation and Computation*, 38(4) 788- 801.

Xinjun, P. (2010). TSVR: An efficient Twin Support Vector Machine for regression. *Neural Networks*, 23, 365-372

Xiongcai, C. and A. Sowmya (2009). Learning to tune level set methods. In *Image and Vision Computing New Zealand, 2009. IVCNZ '09. 24th International Conference*, 310-315, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5378391&isnumber=5378349 (accessed May 16, 2010).

Zaman, A. (1984). Avoiding Model Selection by the use of Shrinkage Techniques. *Journal of Econometrics*, 25, 239-246.