OPTIMAL CRITICAL REGIONS FOR PRE-TEST ESTIMATORS USING A BAYES RISK CRITERION*

C.S. ROEHRIG

Vector Research, Inc., and University of Michigan, Ann Arbor, MI 48106, USA

Optimal critical regions for pre-test estimators are developed in the context of the normal linear regression model with a conjugate prior, where the criterion is Bayes risk and where the pre-testing involves a single coefficient. When prior information is neutral regarding the sign of the coefficient, it is shown that the optimal include–exclude decision involves no pre-testing. When prior information is weighted toward a particular sign, pre-testing is appropriate but the optimal critical regions are quite different from those associated with the traditional pre-test estimator.

1. Introduction

Consider the standard linear model

$$y = W\alpha + \varepsilon,$$

representing T observations on the dependent variable y and the K explanatory variables W, where the random error term ε is assumed to be $N(0, \sigma^2 I)$ conditional upon W. Partition W into X and Z, where X consists of K_1 variables whose coefficients are of particular interest to the researcher, and Z consists of K_2 variables whose coefficients, in and of themselves, are not necessarily of interest. The model can then be written

$$y = X\beta + Z\gamma + \varepsilon.$$

In addition to the estimation of β, the researcher is also concerned with making predictions using the estimated model.

Although the ordinary least squares (OLS) regression of y on X and Z produces minimum variance unbiased estimators and predictors, there are conditions under which the mean square error of estimation and prediction is reduced by excluding Z from the regression in spite of $\gamma \neq (0)$ [null vectors and matrices are denoted by (0)]. Specifically, let

$$\hat{\alpha} = \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = (W'W)^{-1}W'y, \qquad \check{\alpha} = \begin{bmatrix} b \\ (0) \end{bmatrix} = \begin{bmatrix} (X'X)^{-1}X'y \\ (0) \end{bmatrix}$$

be the estimators of α corresponding to the inclusion and exclusion of Z. Letting MSE denote matrix mean square error, it can be shown that $\text{MSE}(\hat{\alpha})$ − $\text{MSE}(\check{\alpha})$ is positive semidefinite if and only if $\text{MSE}(\hat{\gamma}) - \gamma\gamma'$ is positive semidefinite, where $\text{MSE}(\hat{\gamma}) = \sigma^2 (Z'M_x Z)^{-1}$ and $\gamma\gamma'$ is the MSE of (0) as an estimator of γ, and where M_x is defined as $I - X(X'X)^{-1}X'$.[1] Thus, whether the objective is to estimate β or to predict y conditional on X and Z, the key issue is whether $\hat{\gamma}$ or (0) is the better estimator of γ in terms of MSE.

The problem, of course, is that the conditions under which it is best to exclude Z depend upon the unknown parameters γ and σ^2. Therefore these conditions alone do not constitute a practical decision rule for the inclusion or exclusion of Z. One possible solution to this problem is to replace the unknown parameters with the unbiased OLS estimators of those parameters. Noting that $\sigma^2 (Z'M_x Z)^{-1} - \gamma\gamma'$ is positive semidefinite if and only if $\gamma'Z'M_x Z\gamma/(K_2\sigma^2) \le 1/K_2$ [see Toro-Vizcarrondo and Wallace (1968, p. 561)], this approach leads to excluding Z from the regression when $\hat{F} \le 1/K_2$, where $\hat{F}$ is the usual F-statistic used to test for $\gamma = (0)$ ($\hat{\sigma}^2$ is the usual unbiased estimator of σ^2 using the residuals from the regression of y on W),

$$\hat{F} = \frac{\hat{\gamma}'Z'M_x Z\hat{\gamma}}{K_2\hat{\sigma}^2}. \tag{1}$$

The practice of including or excluding Z in the final regression based upon whether or not $\hat{F}$ exceeds some critical value δ^2 defines a 'pre-test' estimator of α given by

$$\tilde{\alpha} = \begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} \quad \text{if} \quad \hat{F} \ge \delta^2,$$

$$= \begin{bmatrix} b \\ (0) \end{bmatrix} \quad \text{if} \quad \hat{F} < \delta^2, \tag{2}$$

where $b = (X'X)^{-1}X'y$. The preceding analysis provides a motivation for using pre-test estimators and also suggests a critical value. However, studies of pre-test estimators reveal that there is no single critical value that is 'best' in terms of commonly used MSE criteria. For example, see Wallace and Ashar (1972), Feldstein (1973), Judge and Bock (1978), Leamer (1978), and Judge et al. (1980) [Sawa and Hiromatsu (1973) and Toyoda and Wallace (1976) derive critical values that are 'optimal' in terms of 'minimax regret']. The mean square error associated with various critical values depends upon the unknown parameters γ and σ^2, and no critical value dominates over the entire parameter space. Thus, it has been suggested that the proper choice of a critical value

---

[1]See Toro-Vizcarrondo and Wallace (1968, p. 561) or Judge and Bock (1978, p. 29) for a derivation of this result.

depends upon prior information regarding the unknown parameters. For example, Wallace and Ashar (1972, p. 177) suggest that 'the stronger the belief that $\gamma \neq 0$', the larger should be the critical value, and the critical value should be smaller 'the stronger the prior doubts about the inclusion of Z'.

This paper is concerned with the relationship between prior information and the appropriate critical value for pre-test estimation. It proceeds by deriving critical regions that minimize the Bayes risk of the pre-test estimator under particular prior distributions on the unknown regression coefficients. It is important to note that even when these 'optimal' critical regions are used, the resulting pre-test estimator is only optimal within the class of all pre-test estimators. Compared with other estimators it may be far from optimal. Frequentists can complain that it is admissible, and Bayesians can point out that the posterior mean will have lower Bayes risk under a quadratic loss function. While this estimator is thus guaranteed to displease purists of either persuasion, it should prove interesting to those who are not in the habit of using Bayesian procedures in applied work, but who are willing to accept the basic correctness of the Bayesian viewpoint. The incorporation of prior information is much simpler than in the full Bayes treatment (it turns out that only prior information regarding the Z variable must be specified) and the results can be presented in a familiar non-Bayesian fashion. It is hoped that even pure Bayesians might view these results as a nudge in the right direction to those who find the full Bayesian treatment to be too much trouble.

The paper is organized as follows. Section 2 develops the relationship between the mean square error of the pre-test estimator of the complete coefficient vector α, and the mean square error of the pre-test estimator of γ, the subset of α associated with the variables to be included or excluded. It shows that the 'best' pre-test estimator of γ is also the 'best' pre-test estimator of α in a mean square error sense, so that attention can be restricted to γ. Section 3 derives a similar result for the Bayes risk of the pre-test estimator and develops a decision rule that defines the minimum Bayes risk pre-test estimator when Z is a single variable. In section 4 this rule is applied using conjugate priors. Section 5 provides a summary and conclusions.

2. Some basic mean square error results

The pre-test estimator to be analyzed here is less restrictive than that defined in (2) in terms of the form of the critical region. It is defined here as

$$\tilde{\alpha} = \begin{bmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} \quad \text{if} \quad \hat{\gamma} \in A(\hat{\sigma}^2, W),$$

$$= \begin{bmatrix} b \\ (0) \end{bmatrix} \quad \text{otherwise,}$$

(3)

where $A(\hat{\sigma}^2, W)$ is a critical region that depends upon the OLS estimator of σ^2 and the explanatory variables W. The reason for this more general representation of a critical region will become apparent in a later section covering the analysis of priors in which there is information regarding the sign of γ.

The following theorem relates $\mathrm{MSE}(\tilde{\alpha})$ to $\mathrm{MSE}(\tilde{\gamma})$:

Theorem 1.

$$\mathrm{MSE}(\tilde{\alpha}) = \left[\begin{array}{c|c} \sigma^2(X'X)^{-1} + c\,\mathrm{MSE}(\tilde{\gamma})c' & -c\,\mathrm{MSE}(\tilde{\gamma}) \\ \hline -\mathrm{MSE}(\tilde{\gamma})c' & \mathrm{MSE}(\tilde{\gamma}) \end{array}\right],$$

where $c = (X'X)^{-1}X'Z$.

The proof of this theorem is given in the appendix.

Next, let $\tilde{\alpha}_I$ and $\tilde{\alpha}_{II}$ be pre-test estimators as defined in (3) corresponding to critical regions $A_I(\hat{\sigma}^2, W)$ and $A_{II}(\hat{\sigma}^2, W)$, respectively. Then:

Theorem 2. $\mathrm{MSE}(\tilde{\alpha}_I) - \mathrm{MSE}(\tilde{\alpha}_{II})$ *is positive semidefinite (p.s.d.) if and only if* $\mathrm{MSE}(\tilde{\gamma}_I) - \mathrm{MSE}(\tilde{\gamma}_{II})$ *is p.s.d.*

Proof. Let $D = \mathrm{MSE}(\tilde{\gamma}_I) - \mathrm{MSE}(\tilde{\gamma}_{II})$. Then, by Theorem 1,

$$\mathrm{MSE}(\tilde{\alpha}_I) - \mathrm{MSE}(\tilde{\alpha}_{II}) = \left[\begin{array}{c|c} cDc' & -cD \\ \hline -Dc' & D \end{array}\right].$$

Let a_1 and a_2 be arbitrary vectors with K_1 and K_2 elements, respectively, and let $a' = (a_1', a_2')$. Then

$$a'\left[\begin{array}{c|c} cDc' & -cD \\ \hline -Dc' & D \end{array}\right]a = \left(a_1'c - a_2'\right)D\left(a_1'c - a_2'\right)',$$

which is non-negative for all (a_1', a_2') if and only if D is p.s.d.

Theorem 2 shows that if one critical region is better than another in terms of the MSE of the resulting estimator of γ, it is also better in terms of the MSE of the resulting estimator of the full coefficient vector α. Thus under the MSE criterion it is possible to restrict attention to $\tilde{\gamma}$.

This result simplifies the problem of deriving the optimal critical value but a number of difficulties remain. First, if Z consists of more than one variable, $\mathrm{MSE}(\tilde{\gamma})$ will not be a scalar, and choosing between competing critical values requires the determination as to whether the difference between the associated

MSE matrices is p.s.d.[2] Second, even if Z is a single variable, MSE($\tilde{\gamma}$) depends upon the unknown parameters γ and σ^2 and no single critical region is optimal for all γ and σ^2.

As a simple illustration of this last point, suppose that Z is a single variable and σ^2 is known so that $V(\hat{\gamma})$ (the variance of the OLS estimator of γ) is also known. Define the pre-test estimator of γ as

$$\tilde{\gamma} = \hat{\gamma} \quad \text{if} \quad |\gamma / \sqrt{V(\hat{\gamma})}| \geq \delta,$$

$$= (0) \quad \text{otherwise,}$$

and define the 'true t' [as in Feldstein (1973)] as $|\gamma / \sqrt{V(\hat{\gamma})}|$. It has been shown (e.g., by Feldstein) that when the true t is less then one, the optimal critical value is infinite, while the optimal critical value is zero if the true t is greater than one. No critical value dominates over the entire range of the true t and therefore prior information regarding the true t is crucial to the determination of an appropriate critical value. A sensible approach is to specify a prior density for the true t, and then evaluate the Bayes risk associated with various critical values. The optimal critical value is defined as that which has the smallest Bayes risk among all possible critical values. The next section implements a somewhat more general version of this approach.

3. The Bayes risk of pre-test estimators

Let the prior density of β, γ, and σ^2 be denoted by $h^*(\beta, \gamma, \sigma^2)$. Then the Bayes risk matrix for the pre-test estimator $\tilde{\alpha}$ defined in (3) is $R(\tilde{\alpha}) = E_{h^*}(\text{MSE}(\tilde{\alpha}))$, where E_{h^*} is the expectation with respect to the density h^*. The objective is to determine the critical region that 'minimizes' $R(\tilde{\alpha})$ given the prior density h^*.

As an initial simplification, note from Theorem 1 that MSE($\tilde{\alpha}$) depends only upon σ^2, W and MSE($\tilde{\gamma}$). But MSE($\tilde{\gamma}$) is determined completely by the joint distribution of $\hat{\gamma}$ and $\hat{\sigma}^2$ which does not depend upon β. Therefore $E_{h^*}(\text{MSE}(\tilde{\alpha})) = E_h(\text{MSE}(\tilde{\alpha}))$, where h is the marginal prior density for γ and σ^2 found by integrating h^* with respect to β. This means that in evaluating the Bayes risk of pre-test estimators, only the marginal prior density for γ and σ^2 needs to be specified.

[2] The pre-test estimation literature typically deals with this problem by defining an alternative MSE criterion that is a scalar regardless of the number of variables represented in Z. Two such criteria are $E((\tilde{\alpha} - \alpha)'(W'W)(\tilde{\alpha} - \alpha))$ and $E((\tilde{\alpha} - \alpha)'(\tilde{\alpha} - \alpha))$, where the objective is to find a critical value that minimizes the given expectation. Judge and Bock (1978) and Judge et al. (1980) provide analyses of pre-test estimators under various MSE criteria including those mentioned above.

Now let $\tilde{\alpha}_I$ and $\tilde{\alpha}_{II}$ be two pre-test estimators corresponding to critical regions $A_I(\hat{\sigma}^2, W)$ and $A_{II}(\hat{\sigma}^2, W)$, respectively. Then it can be shown:

Theorem 3. $R(\tilde{\alpha}_I) - R(\tilde{\alpha}_{II})$ *is p.s.d. if and only if* $R(\tilde{\gamma}_I) - R(\tilde{\gamma}_{II})$ *is p.s.d., where* $R(\tilde{\gamma}) = E_h(MSE(\tilde{\gamma}))$.

The proof is given in the appendix.

This theorem is equivalent to Theorem 2 except that Bayes risk is substituted for MSE. It is important because it implies that the critical region with the smallest associated Bayes risk in terms of γ also has the smallest Bayes risk in terms of α. Thus consideration can be limited to evaluating the Bayes risk of $\tilde{\gamma}$.

To simplify the problem of deriving the minimum Bayes risk critical region (i.e., the critical region with the smallest Bayes risk of all critical regions) only the case where Z is a single variable will be considered. The minimum Bayes risk critical region in this case is simply that region that minimizes $R(\tilde{\gamma})$. The pre-test estimator associated with this critical region will be the minimum Bayes risk estimator and predictor, among all possible pre-test estimators defined in (3), by virtue of Theorem 3.

While it is difficult to actually evaluate the Bayes risk associated with a given prior and a given critical region, there is a rather simple method for determining the critical region that minimizes the Bayes pre-test estimator risk $R(\tilde{\gamma})$. This method is based on the following theorem:

Theorem 4.   The minimum Bayes risk pre-test estimator critical region $A^*(\hat{\sigma}^2, W)$ includes $\hat{\gamma}$ if and only if $E((\hat{\gamma} - \gamma)^2; \hat{\gamma}, \hat{\sigma}^2, W) < E((\gamma)^2; \hat{\gamma}, \hat{\sigma}^2, W)$.

The proof is given in the appendix.

This theorem states that to minimize Bayes risk within the class of pre-test estimators, Z should be included or excluded according to which decision minimizes the posterior expected loss. It is a special case of a standard theorem in Bayesian analysis [see Berger (1980, p. 109)].

Let γ_E and $(\gamma^2)_E$ be the posterior expectations of γ and γ^2, respectively. Then Theorem 4 states that Z should be included if and only if $\hat{\gamma}^2 - 2\hat{\gamma}\gamma_E + (\gamma^2)_E < (\gamma^2)_E$. Adding $(\gamma_E)^2 - (\gamma^2)_E$ to both sides of this inequality shows that this condition is equivalent to $(\hat{\gamma} - \gamma_E)^2 < (\gamma_E)^2$ or $|\hat{\gamma} - \gamma_E| < |\gamma_E|$. Thus, this decision rule chooses $\hat{\gamma}$ or zero as an estimator of γ depending upon which is closer to γ_E. This is a sensible result because γ_E minimizes Bayes risk (of course γ_E would be preferable, in terms of Bayes risk, to either $\hat{\gamma}$ or zero, but it is not one of the choices associated with pre-test estimators). The implications of this decision rule are developed in the next section.

4. An application using conjugate priors

Conditional upon W, γ, and σ^2, $\hat{\gamma}$ has the normal distribution with mean γ and variance $V(\hat{\gamma}) = \sigma^2 (Z'M_x Z)^{-1}$, where $M_x = I - X'(X'X)^{-1}X$. Assume a normal-gamma prior for γ and $V(\hat{\gamma})^{-1}$ such that the prior distribution on γ, conditional upon W and σ^2, is normal with mean γ_p and variance $V(\hat{\gamma})/n^*$. Then the posterior expectation of γ is

$$\gamma_E = \lambda_1 \gamma_p + \lambda_2 \hat{\gamma},$$

where $\lambda_1 = n^*/(n^* + 1)$ and $\lambda_2 = 1/(n^* + 1)$ [see Leamer (1978, p. 60)].

The decision rule derived from Theorem 4 states that Z should be included if and only if $(\hat{\gamma} - \gamma_E)^2 < (\gamma_E)^2$. If $\hat{\gamma} \neq 0$, this condition is equivalent to[3]

$$\gamma_p/\hat{\gamma} > (n^* - 1)/(2n^*). \tag{4}$$

If $\hat{\gamma} = 0$, the include–exclude decision is irrelevant.

4.1. Case 1: $\gamma_p = 0$

Consider first the case of $\gamma_p = 0$ which implies that prior information is neutral with respect to the sign of γ (positive and negative values are equally likely). Then $\gamma_E = \hat{\gamma}/(n^* + 1)$ and by the decision rule presented in (4), the minimum Bayes risk pre-test estimator includes Z if and only if $n^* < 1$. This result is noteworthy because n^* is a parameter of the normal-gamma prior. The include–exclude decision is therefore unrelated to $\hat{\gamma}$ and $\hat{\sigma}^2$ as it is based completely on prior information. The optimal critical region $A(\hat{\sigma}^2, W)$ is either $(-\infty, +\infty)$ or null depending upon whether or not $n^* < 1$.

This result has an intuitively plausible interpretation. The include–exclude decision can be viewed as choosing between two sources of information regarding γ. Prior information centers γ on zero with precision $n^*/V(\hat{\gamma})$ conditional upon σ^2 and W. Sample information centers γ on $\hat{\gamma}$ with precision $1/V(\hat{\gamma})$, also conditional upon σ^2 and W. The decision rule developed above requires that the more precise information be used, because if $n^* < 1$ the sample information is more precise than the prior information.

The neatness of this result is due to the normal-gamma prior in which the precision of the prior information regarding γ is dependent upon σ^2. This dependency means that the precision of both the prior and the sample information are subject to uncertainty (due to uncertainty about σ^2). But it also means that the precision of one relative to the other is known with certainty because the uncertain parameter σ^2 cancels out.

---

[3] Thanks to Ron Mittelhammer for suggesting this simple representation of the basic decision rule. The derivation is in the appendix.

Leamer (1978, p. 80) expresses some reservations regarding this dependency of prior precision on σ^2 although he does point out a possible supporting argument. The position taken here is that Leamer's reservations are well founded. It seems unlikely that the relative precision of prior and sample information regarding γ would be part of prior knowledge when σ^2 is unknown. The problem is that if the prior precision is assumed to be independent of σ^2, the relative precision then depends upon σ^2, and this complicates the analysis of decision rules unless σ^2 is known.

If σ^2 is assumed known, and the prior distribution for γ is normal with mean zero and variance $V_p(\gamma)$, it is easy to show that $\gamma_E = \hat{\gamma}/((V(\hat{\gamma})/V_p(\gamma)) + 1)$ which implies a decision rule that includes Z if and only if $V(\hat{\gamma}) < V_p(\gamma)$. Thus when σ^2 is known, the decision is again based on the relative precision of the prior and sample information. This relative precision is known, conditional upon W, and therefore there is no pre-testing involved in the include–exclude decision.

4.2. Case 2: $\gamma_p \neq 0$

Now consider the case of $\gamma_p \neq 0$. In this case prior information suggests that one sign is more likely than the other. Before examining the include–exclude decision in this case, it is interesting to consider a modification of the pre-test estimator that chooses between $\hat{\gamma}$ and γ_p rather than between $\hat{\gamma}$ and zero. In other words, the alternative to including Z in the regression is to constrain its coefficient to be equal to the prior expectation of γ. It is easy to show that the minimum Bayes risk pre-test estimator decision rule is identical to that derived in the previous section. If $n^* < 1$, Z is included in the regression. If $n^* > 1$, the Z coefficient is set equal to γ_p. The plausibility of this result is obvious in view of the preceding analysis. One chooses prior information or sample information according to which is more precise.

When the allowable decisions are either to include or exclude Z, and the prior expectation of γ is not zero, the choice is no longer between prior and sample information because the prior expectation is not one of the choices. In this case the decision rule associated with the normal-gamma prior is given in (4). The resulting conditions for the inclusion of Z are somewhat complex but have a logical interpretation.

First consider the case in which $\hat{\gamma}$ has the expected sign (i.e., $\gamma_p/\hat{\gamma} > 0$) and prior information is precise relative to sample information ($n^* > 1$). Then the decision rule is to include Z if and only if $|\hat{\gamma}| < k|\gamma_p|$, where $k = 2n^*/(n^* - 1)$. This result is striking because, viewed conditional upon $\hat{\sigma}^2$, it is equivalent to including Z only when the t-statistic is *smaller* (in absolute value) than some critical value. This is of course directly opposed to the usual rule of including Z when the t-statistic is large but the common sense of the rule is easily demonstrated. Suppose that both $\hat{\gamma}$ and γ_p are positive. If $\hat{\gamma}$ is small (relative to γ_p) it will lie between 0 and γ_E and will therefore be preferable to 0 as an

estimator of γ. As $\hat{\gamma}$ becomes larger and larger relative to γ_p, the 'strong' prior will eventually result in $\hat{\gamma}$ moving farther from γ_E than γ_E is from 0.

If the sign of $\hat{\gamma}$ is as expected but the prior information is relatively imprecise, the decision rule is to always include Z. On the other hand, if the sign of $\hat{\gamma}$ is unexpected with a relatively precise prior, the rule is to always exclude Z. Only in the case where the sign is unexpected and the prior is relatively imprecise is Z included for large $|\hat{\gamma}|$.

The results of this section have been derived under the assumption that the prior density for γ can be precisely specified. Given the difficulty in perfectly specifying one's prior information [see Hill (1975)], it would be interesting to examine the robustness of this pre-test estimator with respect to errors in the specification of the prior. Berger (1982) demonstrates that Bayesian robustness is generally improved through the use of more fat-tailed priors such as the Cauchy, rather than the commonly used normal priors. The advantage of the fat-tailed prior is that when prior and sample information come into conflict, the fat-tailed prior leads to relatively more weight being given to the sample information. This in turn provides robustness with respect to misspecification of the prior. A complete analysis of this issue is beyond the scope of this paper. However, to achieve a degree of robustness in practice, it is only necessary to be aware of situations in which prior and sample information are in conflict, and to reexamine both prior and sample information in such situations.

5. Summary and conclusions

This paper has been concerned with the normal linear regression model $y = W\alpha + \varepsilon = X\beta + Z\gamma + \varepsilon$, where the inclusion of Z in the final regression depends upon whether or not the OLS estimator $\hat{\gamma}$ falls in some critical region defined in terms of $\hat{\sigma}^2$ and W. The objective was to determine what critical regions minimize Bayes risk under normal-gamma priors for $(\beta, \gamma, \sigma^2)$, where the Bayes risk is $R(\tilde{\alpha}) = E_{h*}((\tilde{\alpha} - \alpha)(\tilde{\alpha} - \alpha)')$, $\tilde{\alpha}$ is the pre-test estimator of α, and h^* is the normal-gamma prior for $(\beta, \gamma, \sigma^2)$. The key results are the following:

(1) $R(\tilde{\alpha})$ depends only upon the marginal prior for (γ, σ^2). Thus to evaluate $R(\tilde{\alpha})$ it is not necessary to specify the prior for β.
(2) When Z is a single variable, the critical region that minimizes $R(\tilde{\gamma})$ also minimizes $R(\tilde{\alpha})$ in the sense that for any other critical region, $R(\tilde{\alpha})$ exceeds the 'minimum' by a positive semidefinite matrix.
(3) The critical region that minimizes $R(\tilde{\gamma})$ is defined to include $\hat{\gamma}$ if and only if $|\hat{\gamma} - \gamma_E| < |\gamma_E|$ where γ_E is the posterior expectation of $\tilde{\gamma}$.
(4) When prior information is neutral with respect to the sign of γ, the include–exclude decision should be based entirely upon prior information (and knowledge of W) and thus no pre-testing is required.
(5) When prior information is not neutral with respect to the sign of γ, pre-testing is appropriate but the optimal critical region is quite different

from that implicit in the common rule of including Z when the absolute value of the t-statistic exceeds some critical value.

A central parameter in the determination of these optimal critical regions is the relative precision of the prior and sample information. Under a normal-gamma prior for (γ, σ^2), the precision of the prior information is $n^*/V(\hat{\gamma})$. The precision of the sample information is simply $1/V(\hat{\gamma})$. Thus the relative precision is n^*. For those who are not happy with a prior whose precision depends upon $V(\hat{\gamma})$, it is useful to consider a normal prior with expectation γ_p and precision $1/V_p$. If σ^2 is known, the relative precision is also known, and is simply $V(\gamma)/V_p = \sigma^2/((Z'M_x Z) \cdot V_p)$. If there is relatively little prior information regarding σ^2, it seems sensible to use $\hat{V}(\hat{\gamma})/V_p = \hat{\sigma}^2/((Z'M_x Z) \cdot V_p)$ as an approximate to the relative precision. Then, for example, if there is no prior information regarding the sign of γ, one would regress y on X and Z, for purposes of observing $\hat{V}(\hat{\gamma})$, and then include Z in the final regression only if $\hat{V}(\hat{\gamma}) < V_p$.

Appendix

A.1. Proof of Theorem 1

Note that

$$\text{MSE}(\tilde{\alpha}) = \left[\begin{array}{c|c} \text{MSE}(\tilde{\beta}) & \text{E}\big((\tilde{\beta} - \beta)(\tilde{\gamma} - \gamma)'\big) \\ \hline \text{E}\big((\tilde{\gamma} - \gamma)(\tilde{\beta} - \beta)'\big) & \text{MSE}(\tilde{\gamma}) \end{array} \right]. \tag{A.1}$$

It will prove useful to note that $\tilde{\beta} = b - c\tilde{\gamma}$, where $b = (X'X)^{-1}X'y$ and $c = (X'X)^{-1}X'Z$. Therefore,

$$\tilde{\beta} - \beta = (b - \beta - c\gamma) - c(\tilde{\gamma} - \gamma). \tag{A.2}$$

The expectation and variance of b are

$$\text{E}(b; W) = \beta + c\gamma, \qquad \text{V}(b; W) = \sigma^2 (X'X)^{-1},$$

which combine with (A.2) to yield

$$\begin{aligned} \text{MSE}(\tilde{\beta}) &= \text{E}\big((\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'\big) \\ &= \sigma^2 (X'X)^{-1} + c\,\text{MSE}(\tilde{\gamma})\,c' \\ &\quad - \text{E}\big((b - \beta - c\gamma)(\tilde{\gamma} - \gamma)'c'; W\big) \\ &\quad - c\,\text{E}\big((\tilde{\gamma} - \gamma)(b - \beta - c\gamma)'; W\big), \end{aligned} \tag{A.3}$$

and

$$E\big((\tilde{\beta} - \beta)(\tilde{\gamma} - \gamma)'; W\big) = E\big((b - \beta - c\gamma)(\tilde{\gamma} - \gamma)'; W\big)$$
$$- c\,E\big((\tilde{\gamma} - \gamma)(\tilde{\gamma} - \gamma)'; W\big). \qquad (A.4)$$

The following lemmas help simplify (A.3) and (A.4):

Lemma A.1. b, $\hat{\gamma}$, and $\hat{\sigma}^2$ *are mutually independent.*

Proof. Note that $Z = Xc + M_x Z$ so that $y = X\beta + Z\gamma + \varepsilon = X(\beta + c\gamma) + M_x Z\gamma + \varepsilon = X\beta^* + Z^*\gamma + \varepsilon$. The OLS estimators of β^* and γ using this form of the equation are b and $\hat{\gamma}$, respectively, and therefore b and $\hat{\gamma}$ are independent of $\hat{\sigma}^2$ by a well-known theorem. Furthermore, b and $\hat{\gamma}$ are themselves independent because they are normally distributed and uncorrelated (by virtue of the orthogonality of X and Z^*).

Lemma A.2. $E((b - \beta - c\gamma)(\tilde{\gamma} - \gamma)'; W) = (0)$.

Proof. Note that $\tilde{\gamma}$ is fixed [at either $\hat{\gamma}$ or (0)] conditional upon W, $\hat{\gamma}$, and $\hat{\sigma}^2$. Therefore, $E((b - \beta - c\gamma)(\tilde{\gamma} - \gamma)'; W, \hat{\gamma}, \hat{\sigma}^2) = E((b - \beta - c\gamma); W, \hat{\gamma}, \hat{\sigma}^2) \times (\tilde{\gamma} - \gamma)' = (0)$ by Lemma A.1.

Combining (A.1), (A.3), (A.4), and Lemma A.2 yields

$$\text{MSE}(\tilde{\alpha}) = \left[\begin{array}{c|c} \sigma^2(X'X)^{-1} + c\,\text{MSE}(\tilde{\gamma})c' & -c\,\text{MSE}(\tilde{\gamma}) \\ \hline -\text{MSE}(\tilde{\gamma})c' & \text{MSE}(\tilde{\gamma}) \end{array}\right],$$

which proves Theorem 1.

A.2. Proof of Theorem 3

In the proof of Theorem 2 it was shown that

$$\text{MSE}(\tilde{\alpha}_I) - \text{MSE}(\tilde{\alpha}_{II}) = \left[\begin{array}{c|c} cDc' & -cD \\ \hline -Dc' & D \end{array}\right],$$

where

$$D = \text{MSE}(\tilde{\gamma}_I) - \text{MSE}(\tilde{\gamma}_{II}).$$

Therefore,

$$R(\tilde{\alpha}_I) - R(\tilde{\alpha}_{II}) = E_h(\text{MSE}(\tilde{\alpha}_I)) - E_h(\text{MSE}(\tilde{\alpha}_{II}))$$
$$= \left[\begin{array}{c|c} cR(D)c' & -cR(D) \\ \hline -R(D)c' & R(D) \end{array}\right],$$

where $R(D) = E_h(\tilde{\gamma}_I) - E_h(\tilde{\gamma}_{II}) = R(\tilde{\gamma}_I) - R(\tilde{\gamma}_{II})$. Following the steps found in the proof of Theorem 2 it is easily demonstrated that $R(\tilde{\alpha}_I) - R(\tilde{\gamma}_{II})$ is p.s.d. if and only if $R(D)$ is p.s.d. and the theorem is proved.

A.3. Proof of Theorem 4

Define

$g_1(\hat{\gamma}, \hat{\sigma}^2; W, \gamma, \sigma^2) =$ joint density of $\hat{\gamma}$ and $\hat{\sigma}^2$ conditional upon W, γ, and σ^2,
$g_2(\hat{\gamma}, \hat{\sigma}^2, \gamma, \sigma^2; W) =$ joint density of $\hat{\gamma}$, $\hat{\sigma}^2$, γ, and σ^2 conditional upon W,
$g_3(\hat{\gamma}, \hat{\sigma}^2; W) \qquad =$ marginal joint density of $\hat{\gamma}$ and $\hat{\sigma}^2$ conditional upon W,
$g_4(\gamma, \sigma^2; \hat{\gamma}, \hat{\sigma}^2, W) = g_2/g_3 =$ joint posterior density of γ and σ^2 conditional
upon $\hat{\gamma}$, $\hat{\sigma}^2$, and W.

Then

$$R(\tilde{\gamma}) = E_{g_2}(\tilde{\gamma} - \gamma)^2 = E_{g_3}\Big(E_{g_4}(\tilde{\gamma} - \gamma)^2\Big).$$

Note that in $E_{g_4}(\tilde{\gamma} - \gamma)^2$, $\tilde{\gamma}$ is fixed and is either $\hat{\gamma}$ or (0) depending upon whether or not $\hat{\gamma} \in A(\hat{\sigma}^2, W)$. Therefore $R(\tilde{\gamma})$ will be minimized by choosing $A(\sigma^2, W)$ such that it includes $\hat{\gamma}$ if $E_{g_4}(\hat{\gamma} - \gamma)^2 < E_{g_4}(\gamma_p - \gamma)^2$, and the theorem is proved.

References

Berger, James O., 1982, Bayesian robustness and the Stein effect, Journal of the American Statistical Association 77, 358–368.

Berger, James O., 1980, Statistical decision theory: Foundations, concepts, methods (Springer-Verlag, New York).

Hill, Bruce, 1974, On coherence, inadmissibility, and inference about many parameters in the theory of least squares, in: S. Fienberg and A. Zellner, eds., Studies in Bayesian econometrics and statistics (North-Holland, Amsterdam).

Feldstein, Martin, 1973, Multicollinearity and the mean square error criterion, Econometrica 41, 337–346.

Leamer, Edward E., 1978, Specification searches: Ad hoc inference with non-experimental data (Wiley, New York).

Judge, George C. and M.E. Bock, 1978, The statistical implications of pre-test and Stein-rule estimators in econometrics (North-Holland, Amsterdam).

Sawa, Takamitsu and Takeshi Hiromatsu, 1973, Minimax regret significance points for a preliminary test in regression analysis, Econometrica 41, 1093–1101.

Toro-Vizcarrondo, G. and T.D. Wallace, 1968, A test of the mean square error criterion for restrictions in linear regressions, Journal of the American Statistical Association 63, 558–572.

Toyoda T. and T.D. Wallace, 1976, Optimal critical values for pre-testing in regression, Econometrica 44, 365–375.

Wallace, T.D. and V.G. Ashar, 1972, Sequential methods of model construction, Review of Economics and Statistics 54, 172–178.