

Eliminating Extraneous Parameters in Linear Models when the Data is Ill-conditioned

Robert E. Wheeler
ECHIP, Inc.
7460 Lancaster Pike, Suite 6
Hockessin, DE 19707

It has been noted that small F-values indicate a problem with the assumed model. MSE is used as a criterion to construct tables of critical probabilities that may be used to eliminate extraneous parameters from a linear model.

KEY WORDS : Preliminary test: Mean square error: Regret.

1. INTRODUCTION

Högfeldt (1979) writes, "... it is a fairly well established rule of thumb to question the correctness of the assumed model for data, if, in testing the hypothesis of a restricted model, one obtains an F-test quotient value which is significantly low." He derives a test for model inadequacy and shows that the critical values should be chosen from the lower tail of the F distribution. The present paper studies the problem of choosing critical probabilities for this test so that mean square errors (MSE) will be minimized.

It is well known that the elimination of parameters with small population values can produce estimates with MSE's smaller than least squares. The straightforward elimination of parameters corresponding to estimates with magnitudes smaller than some constant will not work because estimates can be small for two reasons: (1) the population values can be small, (2) a chance fluctuation can made the estimates small. There is no way to decide between these two possibilities in any particular case. However, one can adopt a strategy with respect to small estimates that will prove beneficial in the long run.

The strategy to be discussed here consists in using a preliminary test of significance to decide whether or not to eliminate parameters. This has been studied by Sawa and Hiromatsu (1973), [see also Gun (1967)] who provide a table of critical t-test values in the case of perfectly conditioned data. One way to use their table is to classify parameters into three classes as follows. Let t_i be the observed test statistic for the i th parameter, λ_1 and λ_2 be two percentage points, λ_1 from the usual tables for, say, the 5% level and λ_2 from the Sawa and Hiromatsu table. The classification is:

$$\begin{aligned}
 \text{(S)} \quad & \lambda_1 \leq t_i, \\
 \text{(U)} \quad & \lambda_2 \leq t_i < \lambda_1, \\
 \text{(E)} \quad & t_i < \lambda_2.
 \end{aligned} \tag{1}$$

Class S comprises the significant parameters; class E the extraneous parameters that can well be eliminated; and class U the uncertain class for which no decision can be made. When this classification is used in practice, one finds that eliminating class E parameters has negligible effect on the other parameter estimates. This is not the case for class U parameters.

As the number of error degrees of freedom increases, λ_2 increases up to an asymptotic value corresponding to a two tailed probability of about 0.17. Thus with the probability for λ_1 in the usual range (less than 10%), class U will never be empty. As the information increases with increasing error degrees of freedom, the certainty increases for classifying extraneous parameters in E.

It is possible to consider more than one parameter, in which case the test statistic will follow an F distribution. The present paper extends the Sawa and Hiromatsu table to sets of parameters and allows for ill-conditioning. Table 3 gives upper tail probabilities for the central F distribution corresponding to λ_2 , as a function of the number of parameters in the set v_1 , the error degrees of freedom, v_2 , and the condition number, $\kappa^{-2} = 1/(\Theta\theta)$, where θ is the smallest eigenvalue of the appropriate $v_1 \times v_1$ submatrix of $X'X$, and Θ , the largest eigenvalue of the corre-

sponding submatrix of $(X'X)^{-1}$. For $v_1 = 1$, θ and Θ are scalars readily available from standard computations. The Sawa Hiromatsu table is a subset of Table 3.

Bock et. al. (1973), Brook (1976), and Brook and Fletcher (1981) treat similar problems using quadratic risk defined as the sum of all individual parameter MSE's. Toyoda and Wallace (1976) use an average risk criterion. The risk criteria used give equal weight to the MSE's for large and small population parameters, which may not always be appropriate. CASE 5 of Example 1 illustrates the difficulty. The present paper differs from these by emphasizing the control of MSE for large population parameters and in treating only the v_1 interesting parameters.

2. DETAILS

Let $y = X\beta + \epsilon$, where y is $n \times 1$, X is $n \times k$ and of full rank, β is $k \times 1$, and ϵ is $N(0, \sigma^2 I)$. Define β_0 , without loss of generality, as the first v_1 elements of β , and b_0 as the first v_1 elements of b , the least squares estimate of β . Then with C the leading $v_1 \times v_1$ submatrix of $(X'X)^{-1}$, the MSE is $E(b_0 - \beta_0)(b_0 - \beta_0)' = C\sigma^2$, and if the first v_1 terms are eliminated, it is $E(0 - \beta_0)(0 - \beta_0)' = \beta_0\beta_0'$. When a preliminary test of significance is used, Larson and Bancroft (1963) have shown that the resulting MSE is a linear combination of the above MSE's with weights given by the noncentral F distribution. To be precise the MSE is

$$G = \{1 - q_1(\lambda)\}C\sigma^2 + \{2q_1(\lambda) - q_2(\lambda)\}\beta_0\beta_0'$$

where

$$q_j = P\{F'(v_1 + 2j, v_2; \delta^2) \leq v_1\lambda / (v_1 + 2j)\},$$

with $v_2 = n - k$, and $\delta^2 = \beta_0'C^{-1}\beta_0/\sigma^2$. For the class of linear functionals $u'\beta_0$ with $u'u = 1$, the supremum of $u'Gu$ is

$$[\{1 - q_1(\lambda)\} + \{2q_1(\lambda) - q_2(\lambda)\}\delta^2]\Theta\sigma^2, \tag{2}$$

where Θ is the largest eigenvalue of C .

Let θ be the smallest eigenvalue of the $v_1 \times v_1$ leading submatrix of $X'X$. The *condition index* will be taken as $\kappa = \sqrt{\Theta/\theta}$ [see Stewart(1987)], which ranges upward from unity, with unity the condition index for perfectly conditioned data. Dividing X by $\sqrt{\theta}$ redefines (2) as

$$g(\lambda) = \{[1 - q_1(\lambda)] + [2q_1(\lambda) - q_2(\lambda)]\delta^2\}(\kappa\sigma)^2,$$

which may be described as the maximum risk for any linear functional in the specified class.

Substituting $\lambda = 0$ in $g(\lambda)$, one obtains $g_0 = (\kappa\sigma)^2$, and substituting $\lambda = \infty$, one obtains $g_\infty = (\kappa\sigma\delta)^2$: g_0 is the risk for least squares and g_∞ the risk when all v_1 parameters are eliminated. These are shown in Figure 1. The maximum of $g(\lambda)$ occurs in the range $\delta^2 > 1$, and this maximum decreases as λ decreases. The shaded region denotes regret. For $\delta^2 > 1$ the maximum regret occurs at the maximum of $g(\lambda)$, and for $\delta^2 < 1$, the maximum regret $g(\lambda) - g_\infty$ occurs for $\delta^2 = 0$. It is in fact $\{1 - q_1(\lambda)\}(\kappa\sigma)^2$ and it increases as λ decreases. As λ decreases, $g(\lambda)$ converges to g_0 , and regret for $\delta^2 > 1$ goes to zero, while regret for $\delta^2 < 1$ goes to $(\kappa\sigma)^2$.

Sawa and Hiromatsu minimized the maximum regret, which is achieved by balancing the maximum regrets in the two regions. The minimax regret λ satisfies $\{1 - q_1(\lambda)\}(\kappa\sigma)^2 = \max_{\delta^2 > 1}(g(\lambda) - g_0)$. Since $(\kappa\sigma)^2$ cancels from both sides of this equation, the minimax regret λ does not depend on the conditioning. Minimax regret values are included in Table 3 as a function of v_1 and v_2 .

If λ_m is the minimax regret λ for a given v_1 and v_2 , then the minimax regret is $\{1 - q_1(\lambda_m)\}(\kappa\sigma)^2$. This increases as κ increases, which means for badly conditioned data, the minimax regret can be large. There is nothing that can be done about this, but one can control the maximum regret for $\delta^2 > 1$ by choosing a different λ . Table 3 was constructed by doing this. In particular, λ_2 was chosen to satisfy $\{1 - q_1(\lambda_m)\}(\kappa\sigma)^2 = \max_{\delta^2 > 1}(g(\lambda) - g_0)$, which means that no matter how badly conditioned the data, the maximum regret for $\delta^2 > 1$ will never be larger

than that for perfectly conditioned data. One could of course attempt to control the risk for $\delta^2 < 1$, but this would result in a gross MSE's for large β_0 's.

3. EXAMPLES

EXAMPLE 1.

Table 1 shows a constructed regression example. There are 7 observations, 5 parameters including a constant. The example was constructed to allow the ill-conditioning to be adjusted: the average value of κ^{-1} over the 4 nonconstant parameters is shown below each set of estimated parameters. The population parameters are (1.0, 0.5, 0.0, 0.0). The classification (1) is shown beside each α , where α is an upper tail central F probability.

CASE 1 This case has the best conditioning, $\kappa^{-1} = 0.6604$, and the parameters are well estimated. Parameters 3 and 4 are class E, and when they are eliminated, the estimates can be seen to improve slightly. The magnitude of the α 's for the extraneous parameters 3 and 4 supports Högfeldt (1979).

CASE 2 The ill-conditioning has increased, $\kappa^{-1} = 0.3955$, and parameter 4 has moved into class U.

Note how the increase in ill-conditioning has changed the α 's for parameters 1 and 2.. When parameter 3 is eliminated, parameter 4 remains in class U. Sometimes a class U parameter will move into another class when parameters are eliminated because σ is better estimated by an increase in the number of error degrees of freedom. It has been found useful in such cases to eliminate any new class E parameters.

CASE 3 The ill-conditioning is now $\kappa^{-1} = 0.2058$, and both parameters 3 and 4 are in class U.

CASE 4 As ill-conditioning increases, α 's for significant parameters increase until at some point the estimates become nonsignificant. This has happened here for parameter 1 at the 5% level. The right hand set of columns show the consequence of blindly eliminating nonsignificant parameters. Note especially, the substantial increase in $\hat{\sigma}$, and note also that the elimination of class E parameters in other cases has negligible influence on $\hat{\sigma}$.

CASE 5 The ill-conditioning has now grown so bad that all parameters are class U. Brook and Fletcher (1981) provide tables for two risk functions. The critical probabilities from their tables for CASE 5 are 0.135 and 0.0066 , thus they would recommend estimates of zero for all four parameters. The resulting MSE's would be very large.

EXAMPLE 2.

Hocking (1976) has surveyed most of the techniques for variable selection in regression. He compares several popular techniques with a set of data giving the gasoline mileages for 1973-1974 automobiles. There are 10 explanatory variables. There is a general consensus among the techniques that variables 3, 9, and 10 are the most influential and that 2, 5, and 6 may be helpful. Table 2 shows the results with the present methodology.

At first there are no significant variables, however the elimination of the class E variables causes variables 3, 9, and 10 to become significant. Variables 5 and 6 remain in class U, and their elimination can effect both $\hat{\sigma}$ and the estimates of the other parameters.

REFERENCES

- Bock, M. E., Yancy, T. A., and Judge, G. G. (1973). "The Statistical Consequences of Preliminary Test Estimators in Regression," *Jour. Amer. Statist. Assoc.* 68, 109-116.
- Brook, R. J. (1976), "On the Use of a Regret Function to Set Significance Points in Prior Tests of Estimation," *Jour. Amer. Statist. Assoc.* 71, 126-131.
- Brook, R. J., and Fletcher, R. H. (1981), "Optimal Significance Levels of Prior Tests in the Presence of Multicollinearity," *Commun. Statist. Theor. Meth.* A10(14), 1401-1413.
- Höglfeldt, P. (1979), "On Low F-test Values in Linear Models," *Scand J. Statist.* 6, 275-178.
- Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.
- Larson, H. J., and Bancroft, T. A. (1963), "Biases in Prediction by Regression for Certain Incompletely Specified Models," *Biometrika*, 50, 391-402.
- Sawa, T., and Hiromatsu, T. (1973), "Minimax Regret Significance Points for a Preliminary Test in Regression Analysis," *Econometrica*, 41, 1093-1101.
- Stewart, G. W. (1987), "Collinearity and Least Squares Regression," *Statistical Science*, 2, 68-84.
- Toyoda, T., and Wallace, T. D. (1976), "Optimal Critical Values for Pre-testing in Regression," *Econometrica*, 44, 365-375.

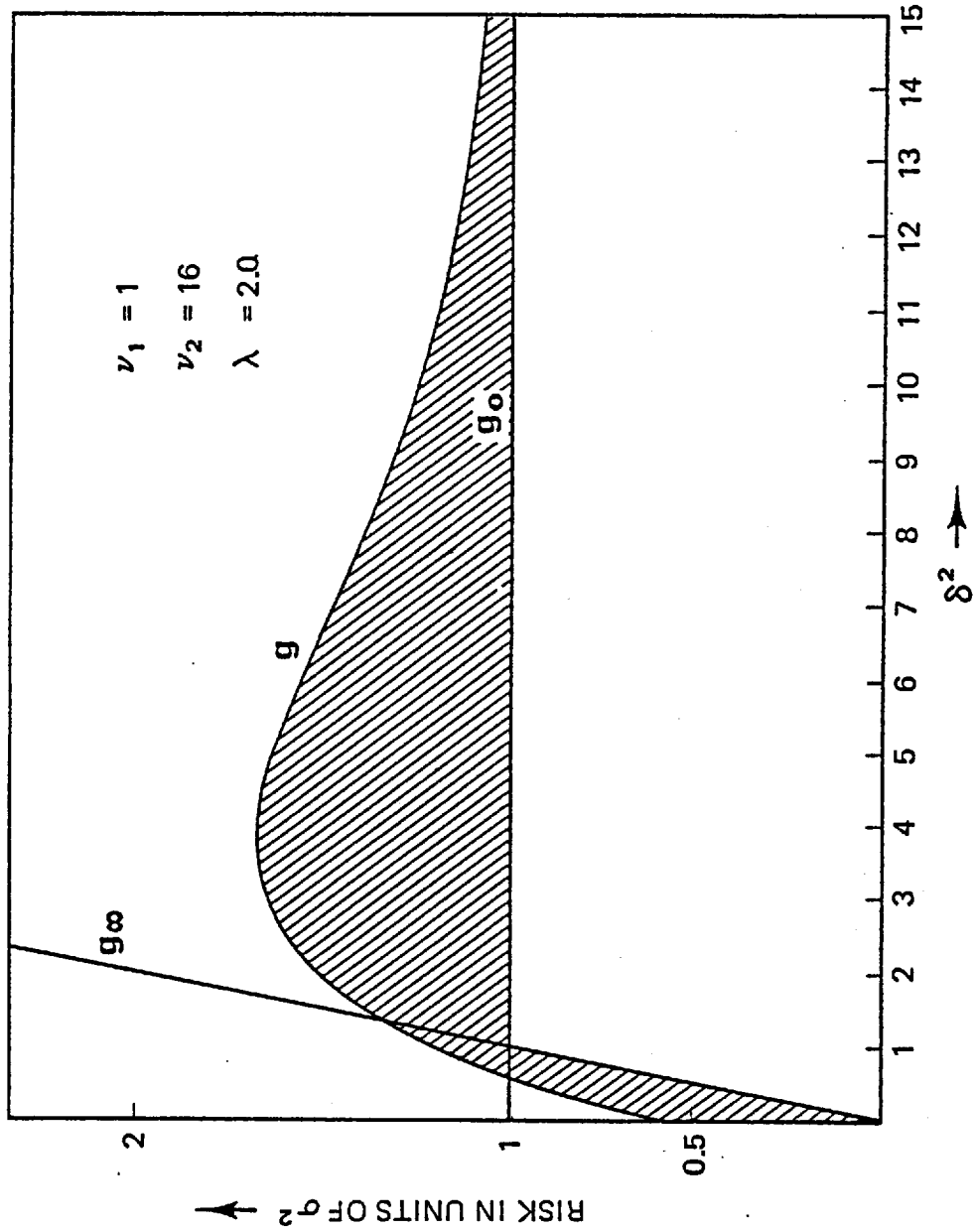


Figure 1. Risk curves for three estimators.

Table 1. Constructed Example

Parameters	CASE 1				CASE 2				CASE 3		CASE 4				CASE 5	
	b	α	b	α	b	α	b	α	b	α	b	α	b	α	b	α
1	0.9933	0.0007S	1.0000	0.0000S	0.9833	0.0029S	0.9630	0.0001S	0.9633	0.0126S	0.9233	0.0532U			0.9033	0.0833U
2	0.5033	0.0007S	0.5010	0.0000S	0.5067	0.0028S	0.5100	0.0002S	0.5133	0.0106S	0.5267	0.0387S	1.5073	0.0000S	0.5333	0.0572U
3	-0.0067	0.6667E			-0.0133	0.6667E			-0.0267	0.6667U	-0.0533	0.6667U			-0.0667	0.6667U
4	0.0133	0.4227E			0.0267	0.4227U	0.0300	0.2722U	0.0533	0.4227U	0.1067	0.4227U			0.1333	0.422U
$\hat{\sigma}$	0.0116		0.0110		0.0116		0.0100		0.0116		0.116		0.05192		0.0116	
Average κ^{-1}	0.6604		0.7638		0.3955		0.4167		0.2058		0.1029		1.0000		0.0822	

NOTE: b's are estimated regression coefficients, and α 's are upper tail central F probabilities.

Table 2 Gas Mileage Example

Variables	b	α	b	α
1	0.318	0.8814E		
2	-0.111	0.9161E		
3	2.520	0.2340U	3.470	0.0275S
4	0.655	0.6652E		
5	0.013	0.4635U	0.011	0.2990U
6	-0.021	0.3350U	-0.021	0.1564U
7	-0.199	0.8122E		
8	0.787	0.6353E		
9	-0.0041	0.0633U	-0.0041	0.0021S
10	0.821	0.2739U	1.007	0.0439S
$\hat{\sigma}$	2.65		2.42	
Average κ^{-1}	0.3655		0.4493	

Table 3. Upper Tail Central F Critical Probabilities

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.650	0.677	0.717	0.767	0.822	0.872
κ^{-2}						
1.000	0.295	0.432	0.569	0.688	0.784	0.856
0.400	0.439	0.594	0.722	0.815	0.871	0.923
0.100	0.628	0.776	0.871	0.925	0.955	0.973
0.040	0.722	0.854	0.926	0.960	0.977	0.987
0.010	0.823	0.925	0.969	0.986	0.992	0.996
0.004	0.869	0.952	0.983	0.993	0.996	0.998
0.001	0.917	0.976	0.993	0.997	0.999	0.999

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.628	0.651	0.687	0.736	0.792	0.848
κ^{-2}						
1.000	0.238	0.363	0.500	0.628	0.738	0.824
0.400	0.391	0.542	0.677	0.781	0.855	0.907
0.100	0.594	0.748	0.851	0.912	0.947	0.968
0.040	0.696	0.835	0.914	0.954	0.974	0.985
0.010	0.807	0.915	0.964	0.983	0.991	0.995
0.004	0.857	0.946	0.980	0.992	0.996	0.998
0.001	0.910	0.973	0.992	0.997	0.999	0.999

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.614	0.635	0.666	0.710	0.764	0.822
κ^{-2}						
1.000	0.206	0.318	0.447	0.576	0.692	0.788
0.400	0.363	0.508	0.642	0.749	0.830	0.889
0.100	0.575	0.728	0.834	0.899	0.938	0.962
0.040	0.682	0.822	0.904	0.947	0.969	0.982
0.010	0.797	0.902	0.960	0.981	0.990	0.994
0.004	0.850	0.941	0.978	0.990	0.995	0.997
0.001	0.905	0.970	0.991	0.997	0.999	0.999

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.607	0.625	0.653	0.692	0.741	0.797
κ^{-2}						
1.000	0.189	0.292	0.412	0.535	0.650	0.750
0.400	0.348	0.488	0.617	0.724	0.807	0.870
0.100	0.564	0.716	0.822	0.889	0.929	0.955
0.040	0.674	0.814	0.897	0.942	0.965	0.979
0.010	0.792	0.904	0.957	0.979	0.989	0.993
0.004	0.846	0.939	0.976	0.989	0.995	0.997
0.001	0.903	0.969	0.990	0.996	0.998	0.999

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.603	0.619	0.645	0.680	0.724	0.776
κ^{-2}						
1.000	0.180	0.178	0.390	0.506	0.617	0.716
0.400	0.340	0.476	0.603	0.706	0.788	0.851
0.100	0.559	0.709	0.815	0.881	0.922	0.949
0.040	0.669	0.810	0.893	0.938	0.962	0.976
0.010	0.789	0.902	0.955	0.977	0.987	0.993
0.004	0.844	0.937	0.975	0.989	0.994	0.997
0.001	0.902	0.968	0.990	0.996	0.998	0.999

v_1	1	2	4	8	16	32
$(\text{Minimax Regret}) / (\text{cc})^2$	0.601	0.617	0.640	0.673	0.713	0.760
κ^{-2}						
1.000	0.175	0.270	0.379	0.489	0.594	0.689
0.400	0.336	0.470	0.594	0.696	0.775	0.837
0.100	0.556	0.706	0.811	0.877	0.917	0.944
0.040	0.667	0.807	0.891	0.935	0.959	0.974
0.010	0.788	0.901	0.954	0.971	0.987	0.992
0.004	0.843	0.937	0.974	0.988	0.994	0.996
0.001	0.901	0.968	0.990	0.996	0.998	0.999

NOTE: Logarithmic interpolation in v_1 , v_2 or κ^{-2} is accurate to two places.