E C O N O M I C S   B U L L E T I N

Multicollinearity and maximum entropy estimators

Quirino Paris
University of California, Davis

Abstract

Multicollinearity hampers empirical econometrics. The remedies proposed to date suffer from pitfalls of their own. The ridge estimator is not generally accepted as a vital alternative to the ordinary least−squares (OLS) estimator because it depends upon unknown parameters. The generalized maximum entropy estimator depends upon subjective exogenous information. This paper presents a novel maximum entropy estimator that does not depend upon any additional information. Monte Carlo experiments show that it is not affected by any level of multicollinearity and dominates the OLS estimator uniformely. The same experiments provide evidence that it is asymptotically unbiased and its estimates are normally distributed.

Quirino Paris is a professor at the University of California, Davis and a member of the Giannini Foundation. I am indebted to Michael R. Caputo and Art Havenner for stimulating discussions that have improved an earlier version of this paper. All errors are mine.
Citation: Paris, Quirino, (2001) "Multicollinearity and maximum entropy estimators." *Economics Bulletin*, Vol. 3, No. 11 pp. 1−9
Submitted: July 29, 2001. **Accepted:** August 17, 2001.
URL: http://www.economicsbulletin.com/2001/volume3/EB−01C20002A.pdf

1. Introduction

Multicollinearity hampers empirical econometrics. The remedies proposed to date are not entirely ideal as they suffer from pitfalls of their own. This paper deals with multicollinearity in the classical linear statistical model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1}$$

where the dimensions of the various components are $\mathbf{y} \sim (T \times 1), \mathbf{u} \sim (T \times 1), \boldsymbol{\beta} \sim (K \times 1)$ and $\mathbf{X} \sim (T \times K)$. The vector $\mathbf{y}$ and the matrix $\mathbf{X}$ constitute sample information. The vector $\boldsymbol{\beta}$ represents parameters to estimate and the vector $\mathbf{u}$ contains random disturbances. It is well known that the ordinary least-squares (OLS) estimator of model (1) exhibits a large and increasing mean squared error (MSE) loss for an increasing level of multicollinearity as measured by Belsley's et al. condition number of the matrix $\mathbf{X}$.

In the seventies, the ridge estimator was proposed as a rival to the OLS estimator when sample data are affected by a high degree of multicollinearity. As Judge et al. (1980, p. 472) wrote: "Ridge regression was originally suggested as a procedure for investigating the sensitivity of least-squares estimates based on data exhibiting near-extreme multicollinearity, where small perturbations in the data may produce large changes in the magnitude of the estimated coefficients. Also discussed, however, is the hope that ridge regression produces 'improved estimates,' in the sense that they may have smaller risk than the conventional least squares estimator." The ridge estimator depends crucially upon an exogenous parameter, say κ, which can be interpreted as a noise to signal ratio. The ridge estimator received a thorough analysis and, by 1974, Theobald demonstrated in rather general terms that a sufficient condition for the ridge estimator to dominate the OLS estimator according to the MSE criterion is $\kappa < 2\sigma^2 / \boldsymbol{\beta}'\boldsymbol{\beta}$, where σ^2 is the population variance and $\boldsymbol{\beta}$ is the unknown vector of population parameters. This finding nullified the original hope that the ridge estimator could become a vital alternative for the OLS estimator in the presence of a high degree of multicollinearity because replacing the unknown population parameters by their sample estimates no longer guarantees the MSE gain.

In 1996, Golan et al. proposed an estimator based on the maximum entropy formalism of Jaynes which they called the generalized maximum entropy (GME) estimator and proceeded to show, by means of Monte Carlo analyses, that this estimator is unaffected for some levels of Belsley's condition number[1]. The GME estimator is consistent and asymptotically normal under some regularity conditions. The idea underlying the GME estimator consists in viewing the components of the parameter vector $\boldsymbol{\beta}$ as convex linear combinations of some known and discrete support values and unknown proportions to be interpreted as probabilities. To be specific, a parameter β_k of

---

[1] Golan, Judge and Miller in their 1996 book (Chapter 8) analyze the behavior of the GME estimator against the OLS estimator using the wrong notion of condition number. Although they quote Belsley, their condition number is simply the ratio of the maximum to the minimum eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrix (not the square root of this ratio, as indicated by Belsley et al.). In their Monte Carlo analysis, they selected values of the condition number that varied from 1 to 100 which correspond to values of Belsley's condition number from 1 to 10. Because multicollinearity begins to signal its deleterious effects when Belsley's condition number is around 30, the discussion of Golan et al. does not involve problems that are ill-conditioned.

the linear statistical model is regarded as the mathematical expectation of some discrete support values Z_{km} such that

$$\beta_k = \sum_m Z_{km} p_{km} \tag{2}$$

where $p_{km} \geq 0$ are probabilities and, of course, $\sum_m p_{km} = 1$ for $k = 1,...,K$. The element Z_{km} constitutes a-priori information provided by the researcher, while p_{km} is an unknown probability whose value must be determined by solving a maximum entropy problem. In matrix notation, let $\boldsymbol{\beta} = \mathbf{Zp}$, with $p_{km} \geq 0$ and $\sum_m p_{km} = 1$ for $k = 1,...,K$. Also, let $\mathbf{u} = \mathbf{Vw}$, with $w_{tg} \geq 0$ and $\sum_g w_{tg} = 1$ for $t = 1,...,T$. Then, the GME estimator can be stated as

$$\max H(\mathbf{p}, \mathbf{w}) = -\sum_{k,m} p_{km} \log(p_{km}) - \sum_{t,g} w_{tg} \log(w_{tg}) \tag{3}$$

subject to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{XZp} + \mathbf{Vw}$$

$$\sum_m p_{km} = 1 \qquad k = 1,...,K$$

$$\sum_g w_{tg} = 1 \qquad t = 1,...,T.$$

The $H(\mathbf{p}, \mathbf{w})$ function measures the entropy in the model. The GME estimator has no closed-form solution and some of its properties must be analyzed via the associated KKT conditions. The GME estimator is not sensitive to multicollinearity because the matrix $\mathbf{X}$ appears off the main diagonal of the appropriate KKT conditions.

The GME estimator, however, has important weaknesses: The estimates of the parameter β_k and residual u_t are sensitive, in an unpredictable way, to changes in the support intervals. A concomitant but distinct weakness is that the GME estimates and their variances are affected by the number of support values. Also, while knowledge of the bounds for some parameters may be available and, therefore, ought to be used, it is unlikely that this knowledge can cover all the parameters of a model. In other words, the GME estimator depends crucially upon the subjective and exogenous information supplied by the researcher: The same sample data in the hands of different researchers willing to apply the GME estimator will produce different estimates of the parameters and, likely, different policy recommendations.

This paper proposes a novel maximum entropy estimator inspired by the theory of light that does not require any subjective and exogenous information. It is called the Maximum Entropy Leuven (MEL) estimator[2] and it is not affected by any degree of multicollinearity. The MEL estimator is scale-invariant in the sense of the OLS estimator. Preliminary Monte Carlo experiments indicate that the MEL estimator is asymptotically unbiased. Furthermore, the same experiments have failed to reject the normality hypothesis for the parameter estimates.

[2] This paper is written in honor and loving memory of my wife, Carlene Paris, who died of leiomyosarcoma – a very rare cancer – on May 5, 2001. The MEL estimator is named after the town of Leuven, Belgium, where she spent the last months of her life while undergoing a difficult cancer treatment.

2. The Maximum Entropy Leuven Estimator

The MEL estimator is inspired by the theory of light. The analogy with economic analysis runs as follows. Light carries information about the physical environment. When light reaches the eyes (photomultipliers) of a person, the perceived image may be out-of-focus. That person will squint and adjust his eyes in order to improve the reproduction of the image in his brain. Economic data carry information about economic environments and the decision processes that generated those data. As with any picture, the information reaching a researcher may correspond to an image that is out-of-focus. The goal of econometric analysis, then, is to reconstruct the best possible image of an economic decision process as the way to better understand the economic agent's environment.

In the theory of light (Feynman), the probability that a photomultiplier is hit by a photon reflected from a sheet of glass is equal to the square of its amplitude. The amplitude of a photon is an arrow (a vector) that summarizes all the possible ways in which a photon could have reached a given photomultiplier.

In an econometric model with noise, it is impossible to measure exactly the parameters involved in the generation of the sample data. Each parameter depends on every other parameter specified in the model and its measured dimensionality is affected by the available sample information as well as by the measuring procedure. Following the theory of light, it is possible to estimate the probability of such parameters using their revealed image. The revealed image of a parameter can be thought of as the estimable dimensionality that depends on the sample information available for the analysis. Hence, in the MEL estimator we postulate that the probability of a parameter β_k (which carries economic information) is equal to the square of its "amplitude" where by amplitude we intend its normalized dimensionality. Thus, the MEL estimator is specified as follows:

$$\min H(\mathbf{p}_\beta, L_\beta, \mathbf{u}) = \mathbf{p}_\beta' \log(\mathbf{p}_\beta) + L_\beta \log(L_\beta) + \mathbf{u}'\mathbf{u} \qquad (4)$$

subject to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$
$$L_\beta = \boldsymbol{\beta}'\boldsymbol{\beta}$$
$$\mathbf{p}_\beta = \boldsymbol{\beta}\Theta\boldsymbol{\beta} / L_\beta$$

where $\mathbf{p}_\beta \geq \mathbf{0}$, the symbol Θ indicates the element-by-element Hadamard product and $\mathbf{u}'\mathbf{u}$ is the sum of squared residuals. The same probability statement could be applied to the error term but in this paper the entropy formulation is kept at its minimum. Also a signal-to-noise parameter could be introduced into the objective function but, without any compelling justification for this specification, its value is chosen as unity.

The MEL estimator does not possess a closed form representation. Its solution requires the use of a computer code for nonlinear programming problems such as GAMS by Brooke et al. This feature is in common with the GME estimator. In spite of its apparent complexity, the solution of numerous test problems was swift and efficient on the same level of rapidity of the OLS estimator.

3. Distributional Properties of the MEL Estimator

The nonlinearity of the MEL estimator makes it difficult to derive its distributional properties. This goal will be the subject of further research. To shed some light on these properties, several Monte Carlo experiments were performed. In particular, asymptotic unbiasedness and normality of the estimated parameters were considered. Furthermore, the behavior of the estimator under increasing levels of multicollinearity was analyzed.

Asymptotic unbiasedness was measured by the magnitude of squared bias in a risk (or average loss) function $\rho(\boldsymbol{\beta},\hat{\boldsymbol{\beta}})$, also called MSEL, as suggested by Judge et al. (1982, p. 558), where

$$\begin{aligned}\rho(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) &= trMSE(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) = trE[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})] \\ &= trCOV(\hat{\boldsymbol{\beta}}) + tr[BIAS(\hat{\boldsymbol{\beta}}) * BIAS(\hat{\boldsymbol{\beta}}')].\end{aligned} \quad (5)$$

Tables 1 and 2 present the results of a non-trivial Monte Carlo experiment that deals with a true model exhibiting the following specification. There are ten parameters β_k, $k = 1,...K$, to estimate. Each parameter β_k was drawn from a uniform distribution $U[-1.7, 2.0]$. Each element of the matrix of regressors $\mathbf{X}$ was drawn from a uniform distribution $U[1,5]$. Finally, each component of the disturbance vector $\mathbf{u}$ was drawn from a normal distribution $N[0,2]$. One hundred samples of increasing size, from 50 to 2000 observations, were replicated. The GME estimator was implemented with three discrete support values for the parameters and the error terms selected as [-5,0,5] and [-10,0,10], respectively. The condition number (CN) of the $\mathbf{X}$ matrix is given for each sample size.

The GME estimator implemented with the optimization program GAMS failed to reach an optimal solution with a sample size of T>400. This event might be due to the large number of probabilities that must be estimated for an increasing number of error terms. The GME estimator produces results that approximate very closely uniform probabilities and this characteristic of the GME estimator may make it difficult with large samples to locate a maximum value of the objective function. Invariably, the GAMS program terminated with a feasible but non-optimal solution when T>400.

The levels reported in Table 1 represent the sum of the squared bias over ten parameters. It would appear that the MEL estimator performs as well as the OLS estimator in small samples. This result is confirmed in Table 2 that presents the levels of MSEL for the same experiment and sample sizes.

The hypothesis that the parameter estimates are distributed according to a normal distribution was tested by the Bera-Jarque (1980) statistic involving the coefficients of skewness and kurtosis that the authors show to be distributed as a χ^2 variable with two degrees of freedom. In all the runs associated with Tables 1 and 2, the normality hypothesis was not rejected with ample margins of safety.

The above results provide evidence that the MEL estimator performs as well as the OLS estimator, under a well-conditioned $\mathbf{X'X}$ matrix. The MEL estimator out-performs the OLS estimator under a condition of increasing multicollinearity. Following Belsley et al. (1980), multicollinearity can be detected in a meaningful way by means of a condition number computed as the square root of the ratio between the maximum and the minimum eigenvalues of a matrix $\mathbf{X'X}$ (not a moment matrix) whose columns have been normalized to a unit length. Equivalently, the same condition number can be obtained by computing the singular value decomposition of a normalized matrix, $norm\mathbf{X}$, such that $norm\mathbf{X} = \mathbf{UDV'}$, where $\mathbf{D}$ is a diagonal matrix of singular values while $\mathbf{U}$ and $\mathbf{V}$ are matrices such that $\mathbf{U'U} = \mathbf{I}$ and $\mathbf{V'V} = \mathbf{I}$. The condition number of the $norm\mathbf{X}$ matrix, then, is the ratio between the maximum and the minimum singular values measured as absolute values. Because $norm\mathbf{X'X} = \mathbf{VDU'UDV'} = \mathbf{VD^2V'} = \mathbf{VLV'}$, with eigenvalues $\mathbf{L} = \mathbf{D}^2$, the two definitions of condition number are consistent. Belsley et al. found that the negative effects of multicollinearity begin to surface when the condition number is around 30. A Monte Carlo experiment was conducted to examine the behavior of the MEL, GME and OLS estimators using the MSEL criterion under increasing values of the condition number with a sample size of T=50. The experiment's structure is identical to that one associated with Tables 1 and 2. The results are presented in Table 3.

The MEL estimator reveals low levels of MSEL and a remarkable stability as the condition number increases. On the contrary, and as expected, the OLS estimator shows a dramatic increase in the MSEL levels for values of the condition number that can be easily encountered in empirical econometric analyses. The GME estimator was implemented in two versions with two different support intervals of the parameters. The first version of GME, with narrow support intervals, reveals a stability comparable to that of the MEL estimator. The second version of GME, with wider support intervals, exhibits a significant increase in MSEL values for certain levels of the condition number. When the number of repeated samples was increased to 300, the results were very similar to those given in tables 1, 2 and 3.

4. Scale Invariance of the MEL Estimator

The MEL estimator is "invariant" to an arbitrary change of measurement units of the sample information in the same sense that the OLS estimator is "invariant" to a change of scale of either the dependent variable or the regressors or both. In reality, a more proper characterization of the OLS and MEL estimators under different scaling is that their estimates change in a known way due to a known (but arbitrary) choice of measurement units of either the dependent variable or regressors or both. Because of this knowledge, it is always possible to recover the original estimates obtained prior to the scale change and, in this sense, both the OLS and the MEL estimators are said to be scale invariant. We now scale the dependent variable $\mathbf{y}$ of the linear statistical model in equation (1) by an arbitrary but known scalar parameter R and the matrix of regressors $\mathbf{X}$ by an arbitrary but known linear operator $\mathbf{S}$ regarded as a non-singular matrix of dimensions $(K \times K)$.

The specification of the optimization model that will produce scale-invariant estimates of the MEL estimator can then be stated as

$$\min H(\mathbf{p}_\beta, L_\beta, \mathbf{u}) = \mathbf{p}'_\beta \log(\mathbf{p}_\beta) + L_\beta \log(L_\beta) + R^2 \mathbf{u}^{*\prime} \mathbf{u}^* \tag{6}$$

subject to

$$\frac{\mathbf{y}}{R} = \left(\frac{\mathbf{X}}{R}\mathbf{S}^{-1}\right)\boldsymbol{\beta}^* + \mathbf{u}^*$$

$$L_\beta = \boldsymbol{\beta}^{*\prime}\mathbf{S}^{-1\prime}\mathbf{S}^{-1}\boldsymbol{\beta}^*$$

$$\mathbf{p}_\beta = \mathbf{S}^{-1}\boldsymbol{\beta}^* \Theta \mathbf{S}^{-1}\boldsymbol{\beta}^* / L_\beta$$

where $\boldsymbol{\beta}^* \equiv \mathbf{S}\boldsymbol{\beta}$, $\mathbf{u}^* \equiv \mathbf{u} / R$ and where the vectors $\boldsymbol{\beta}$ and $\mathbf{u}$ refer to the originally unscaled model. If the scalar R is equal to one and the matrix $\mathbf{S}$ is taken as the identity matrix, the model specified in equation (6) is identical to the unscaled model exhibited in equation (4).

5. Conclusion

The MEL estimator is inspired by the theory of light and rivals the GME estimator of Golan et al. by performing very well under the MSEL risk function while avoiding the requirement of subjective exogenous information that is a necessary component of the GME estimator. The MEL estimator is invariant to a change of scale in the sense of the OLS estimator, and appears to be asymptotically unbiased based on Monte Carlo experiments. Furthermore, the same experiments failed to reject the hypothesis that the distribution of the parameter estimates is normal.

In comparison to the GME estimator, the MEL estimator is parsimonious with respect to the number of parameters to be estimated. For example, the solution of the MEL estimator has (*2K+T*) components (*K* parameters β_k, *K* probabilities p_{β_k}, and *T* error terms u_t). The solution of the GME estimator for a similar model has (*MK+GT*) components, where *M* is the number of discrete supports for the parameter β_k and *G* is the number of discrete supports for the error term u_t. The empirical GME literature indicates that, in general, *M*=5 and *G*=3.

Another distinctive feature of the MEL estimator regards the parameter probabilities that, in general, do not approach the uniform distribution as do the corresponding probabilities of the GME estimator. In order to illustrate this proposition, the parameter and probability estimates for one data sample of the Monte Carlo experiment described above are reported in Table 4. The condition number for the **X** matrix of this sample is equal to 1018. There are ten parameters with true values as reported in the first column. The GME estimator was implemented with three support values for the parameters and a support interval of [-20,0,20]. The three support values for each error term were selected as [-10,0,10]. As anticipated, the probabilities of the GME estimator tend toward the uniform distribution with all values very near to 1/3. On the contrary, the probability values of the MEL estimator are far from the uniform distribution. Because of the presence of a high level of multicollinearity, some of the OLS parameter estimates are very far from the true values.

The MEL estimator appears to succeed where the ridge estimator failed: Under any levels of multicollinearity, the MEL estimator uniformely dominates the OLS estimator according to the mean squared error criterion. It rivals also the GME estimator without requiring any subjective additional information.

6. References

Belsley, D.A., E. Kuh, and R.E. Welsch. *Regression Diagnostics*. John Wiley, New York, 1980.

Bera, A. and C. Jarque. "Efficient Test for Normality, Heteroscedasticity and Serial Independence of Regression residuals: Monte Carlo Evidence." Economics Letters 7,1981, pp.313-318.

Brooke, A., D. Kendrick, and A. Meeraus. *GAMS: A User's Guide*. The Scientific Press, 1988.

Feynman, Richard, *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton, New Jersey, 1985.

Golan, Amos, George Judge and Douglas Miller, *Maximum Entropy Econometrics*, Wiley and Sons, Chichester, UK, 1996.

Jaynes, E. T., "Information Theory and Statistical Mechanics," *Physics Review* 106(1957):620-630, Part II, 108(1957):171-190.

Judge, G.G., W.E. Griffiths, R.C. Hill, T.C. Lee. *The Theory and Practice of Econometrics*. John Wiley, New York, 1980.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl, T.C. Lee. *Introduction to the Theory and Practice of Econometrics*. John Wiley, New York, 1982.

Theobald, C. M., "Generalizations of Mean Square Error Applied to Ridge Regression," *Journal of the Royal Statistical Society, Series B,* 36(1974):103-106.

Table 1. Asymptotic unbiasedness measured by the sum of squared biases. 100 samples.

Estimator	*T*=50 CN=11.5	*T*=200 CN=10.3	*T*=400 CN=9.5	*T*=1000 CN=9.1	*T*= 2000 CN=8.8
MEL	0.03630	0.00371	0.00043	0.00078	0.00013
GME	0.04986	0.00451	0.00057	---------	---------
OLS	0.00893	0.00170	0.00012	0.00056	0.00013

Table 2. MSEL for MEL, GME and OLS estimators. 100 samples

Estimator	*T*=50 CN=11.5	*T*=200 CN=10.3	*T*=400 CN=9.5	*T*=1000 CN=9.1	*T*=2000 CN=8.8
MEL	0.5448	0.1334	0.0709	0.0295	0.0132
GME	0.5469	0.1341	0.0715	--------	--------
OLS	0.5882	0.1351	0.0715	0.0294	0.0132

Table 3. MSEL of rival estimators for increasing condition number. *T*=50, 100 samples

Condition Number	Estimators			
	MEL	GME(-5,5)	GME(-20,20)	OLS
11	0.545	0.547	0.584	0.588
30	0.800	0.773	1.059	1.092
60	0.922	0.858	2.195	2.561
101	0.876	0.818	3.890	6.120
203	0.792	0.758	5.316	23.009
304	0.768	0.742	4.424	50.908
508	0.754	0.733	2.694	117.601
1,018	0.749	0.730	1.346	219.350
4,478	1.120	1.103	1.155	560.338
42,187	1.126	1.108	1.109	601.108

Table 4. Estimates of parameters and probabilities in rival estimators

		MEL		GME[-20,20]				OLS
	True Beta	Beta	Prob(Beta)	Beta	Prob1	Prob2	Prob3	Beta
β_1	-0.0258	-0.0301	0.0001	-0.0208	0.3339	0.3333	0.3328	-0.0048
β_2	-1.0752	-0.8594	0.0957	-0.9108	0.3564	0.3328	0.3108	-0.9552
β_3	0.4149	0.8509	0.0937	0.8469	0.3124	0.3329	0.3547	0.8855
β_4	1.4772	1.1832	0.1815	1.2400	0.3028	0.3324	0.3648	1.2203
β_5	-1.5673	-1.1938	0.1847	-0.9819	0.3582	0.3327	0.3091	7.5817
β_6	-0.3852	-0.5096	0.0337	-0.7394	0.3520	0.3330	0.3150	-9.2844
β_7	-0.4499	-0.8148	0.0861	-1.0607	0.3602	0.3326	0.3072	-9.5962
β_8	0.1000	-0.1732	0.0039	-0.1822	0.3379	0.3333	0.3288	-0.2112
β_9	-0.7403	-0.7781	0.0785	-0.7923	0.3535	0.3329	0.3136	-0.7938
β_{10}	1.5974	1.3662	0.2420	1.4385	0.2980	0.3320	0.3699	1.4385