THE FALSTAFF ESTIMATOR

ROGER KOENKER AND JOSÉ A. F. MACHADO

Abstract. Correcting for heteroscedasticity in GMM estimation of the linear model can improve upon the Gauss-Markov estimator *even when there is no heteroscedasticity to correct.*

1. Introduction

We will show that a heteroscedasticity corrected GMM estimator can improve upon the performance of the least squares estimator in certain, iid error, classical linear regression models, *even though there is no heteroscedasticity to correct.* Consider the classical linear regression model

$$y_i = \sum_{j=1}^{n} x_{ij}\beta_j + u_i \qquad i = 1, ..., n.$$

Throughout, we will assume that the error sequence $\{u_i\}$ is independent and identically distributed with common distribution function F. Under plausible conditions on F we find an unbiased estimator, $\tilde{\beta}_n$, of β with strictly smaller covariance matrix than the classical Gauss-Markov estimator $\hat{\beta}_n = (X'X)^{-1}X'y$. Our estimator, which we call the Falstaff estimator for reasons which will become gradually apparent, is a variant on generalized method of moments estimators which have attracted considerable recent interest in the literature.

2. The Falstaff Estimator

Stein's (1956) celebrated shrinkage results imply that one can improve upon $\hat{\beta}_n$ in Gaussian regression problems with parametric dimension $p \geq 3$ by shrinking $\hat{\beta}_n$ toward some fixed point, thereby trading bias for variance reduction. Judge and Bock (1978) treat this subject in some detail from an econometric standpoint. One might characterize this as statistical stoicism – through restraint, self-discipline and temperance we achieve the noble purpose of reduced mean square error. For others, it may be interpreted as a form of Bayesian parsimony.

We have no quarrel with such philosophies. They are fine for those who, like La Fontaine's ant, prefer to toil all summer to prepare for the hardships of winter. But who speaks out for the profligacy of the grasshopper, for gluttony and reckless abandon? Can such ideas have a place in the dismal annals of econometrics? We beg

Date: March 1, 1996.

the gentle reader's momentary indulgence to consider the following foolishly profligate estimator:

> Augment the p-columns of the initial design matrix X, by q *randomly generated* columns, D. Let $Z = [X\ D]$, and consider estimating the augmented model by ordinary least squares,
>
> $$\hat{\delta}_n = (Z'Z)^{-1}Z'y$$
>
> and denote the familiar Eicker-White covariance matrix estimator for $\hat{\delta}_n$ by
>
> $$\hat{V}_n = (Z'Z)^{-1}Z'RZ(Z'Z)^{-1}$$
>
> where $R = diag(r_i^2)$ and $r = y - Z\hat{\delta}$. Finally, let $G = [I_p\, 0]'$, so $\delta = G\beta$, and define the GMM estimator of β, by solving
>
> $$\min(\hat{\delta} - G\beta)'\hat{V}_n^{-1}(\hat{\delta} - G\beta),$$
>
> yielding,
>
> $$\tilde{\beta}_n = (G'\hat{V}_n^{-1}G)^{-1}G'\hat{V}_n^{-1}\hat{\delta}_n.$$

This may appear to be the recipe of some demented sack-guzzler, but there is method in the madness. Our first result shows that we are not trading bias for variance reduction as in Stein estimation.

Proposition 1. *If F is symmetric about zero, then $\tilde{\beta}_n$ is symmetrically distributed about β, and if it exists, $E\tilde{\beta}_n = \beta$.*

Proof: The argument is essentially that of Kakwani(1967); see also the treatment in Schmidt(1976). Fix Z and write

$$\tilde{\beta}_n = \beta + (G'\hat{V}_n^{-1}G)^{-1}G'\hat{V}_n^{-1}u.$$

Observe that $\hat{V}_n$ is an even function of u, that is u and $-u$ yield the same $\hat{V}_n$. Since by assumption u and $-u$ have the same distribution, it follows that $\tilde{\beta}_n - \beta$ and $-(\tilde{\beta}_n - \beta)$ have the same distribution. And the result then follows by unconditioning on Z. ■

We can conclude from this result that any improvement in mean squared error acheived by $\tilde{\beta}_n$ must be purely a matter of variance reduction. Since for Gaussian F it is well known that $\hat{\beta}_n$ is minimum variance unbiased we obviously must narrow the class of F's to exclude this case. Our next result which is an immediate corollary of Theorem 2.2 of Koenker, Machado, Skeels, and Welsh(1994), henceforth KMSW, specifies the class of distributions for which we may expect an improvement.

Proposition 2. *Under the conditions of Theorem 2.2 of KMSW(1994) with $\{u_i\}$ iid and $k(F) = Eu_1^4/\sigma^4$, we have the variance expansion,*

$$Var(\sqrt{n}(\tilde{\beta}_n - \beta) = \sigma^2\Omega_n + n^{-1}\sigma^2(5 - k(F))\Omega_{2n} + o(n^{-1}),$$

where $\Omega_n = (G'H_nG)^{-1} = (X'X/n)^{-1}$, $H_n = Z'Z/n$, $\Omega_{2n} = \Omega_n G'M_nG\Omega_n$, and $M_n = n^{-1}\sum z_iz_i'[H_n - G(G'H_nG)^{-1}G']z_iz_i'$.

The first term in this variance expansion is familiar, it is the variance that would result had we used the true $V = \sigma^2 I$. The second term which is of order $O(n^{-1})$ may be attributed to the "heteroscedasticity correction" of the GMM estimator and is probably less so. It is easy to see that the matrix Ω_{2n} is positive definite and consequently for distributions with kurtosis greater than 5, the Falstaff estimator, $\tilde{\beta}_n$, has strictly smaller asymptotic covariance matrix, to order $O(n^{-1})$ than the Gauss-Markov estimator $\hat{\beta}_n$.

Of course, for Gaussian F and other distributions with modest kurtosis the second term contributes a positive component and consequently the $\tilde{\beta}_n$ "correction for heteroscedasticity" is counter-productive. This degradation in performance at the Gaussian model is hardly surprising since classical sufficiency arguments as in Rothenberg(1984) imply such a loss is inevitable.

Intuitively, we would expect that ignoring the fact that our observations are homoscedastic couldn't help us. We *should* be punished for ignoring relevant information. Shouldn't we? How then do we gain from the profligate behavior of the Falstaff estimator? How can estimating an artificially expanded model and then correcting for heteroscedasticity *in an iid error model* conceivably increase the precision of our estimates? To explore these questions we begin by considering the particularly simple special case of estimating a scalar location parameter. Since in this case $X = 1$ an n-vector of ones, the form of Ω_{2n} is especially simple: $G = e_1$, $Gz_i = 1$, and therefore $\Omega_{2n} = q$, the number of augmented columns in D. Thus our expansion reduces in this case to

$$Var(\sqrt{n}(\tilde{\beta}_n - \beta) = \sigma^2[1 + (5 - k(F))q/n] + o(n^{-1}).$$

In the next section we report on a small monte-carlo experiment designed to evaluate the accuracy of this expansion for moderate sample sizes. What is the Falstaff estimator *doing* in this simple location context? The Falstaff estimator of location may be expressed as

$$\tilde{\beta}_n = e_1' \hat{V}_n^{-1} \hat{\delta} / e_1' \hat{V}_n^{-1} e_1.$$

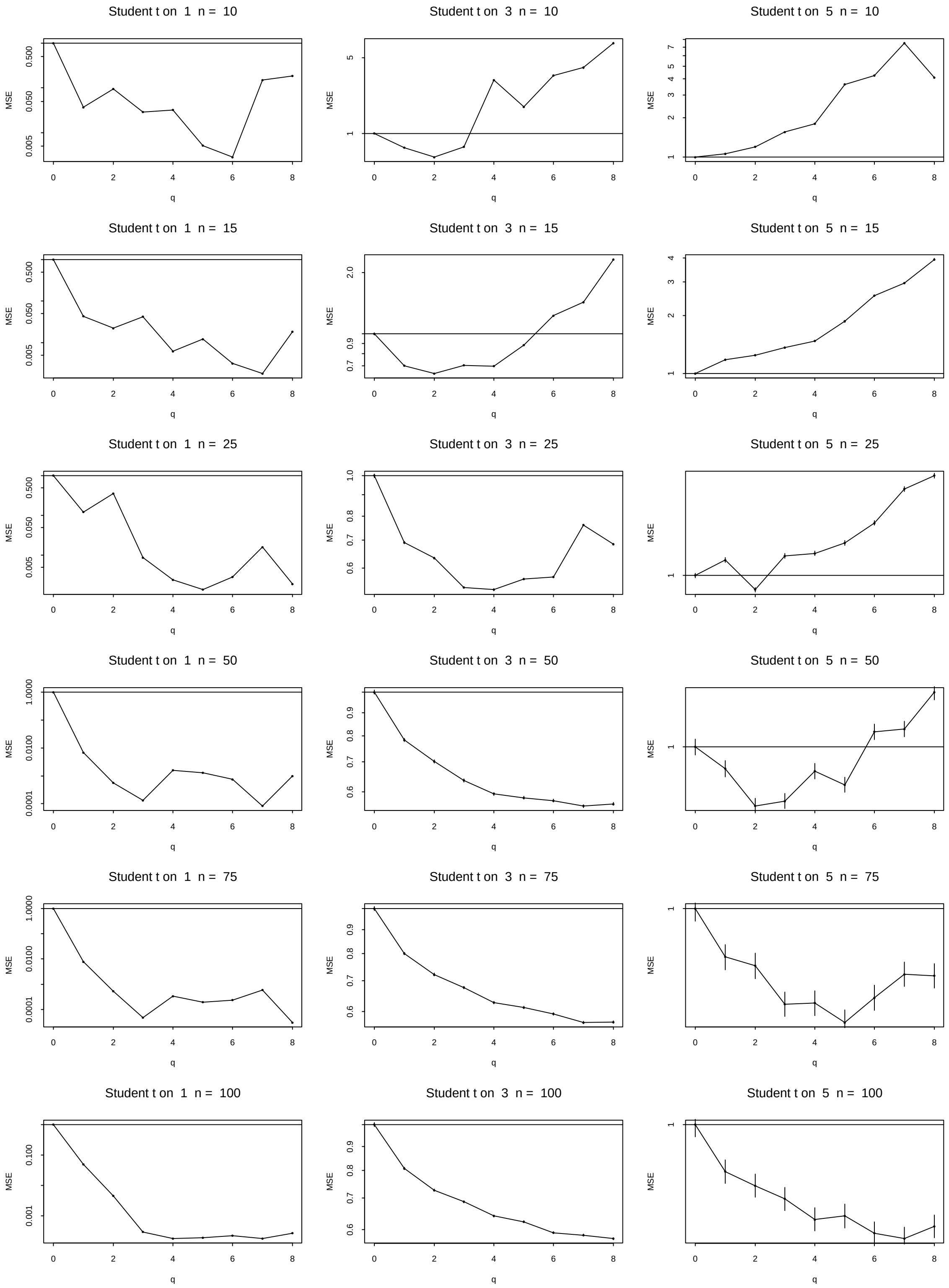
Obviously, if $\hat{V}_n$ is proportional to the identity matrix and D is orthogonal to X then $\tilde{\beta}_n = \hat{\beta}_n = \bar{y}$. However, generally, all the coordinates of $\hat{\delta}$ contribute to $\tilde{\beta}_n$. If $\hat{V}_n$ converges in probability to a nonstochastic matrix, the iid error assumption ensures that the limit *is* proportional to the identity. This simply restates the obvious fact that if $\hat{V}_n$ is consistent as it would be in the present circumstances if q were fixed, then the Falstaff improvement vanishes as $n \to \infty$. Whether there may be some scope for asymptotic improvement if the sequence of D_n matrices could be chosen to preserve a stochastic contribution from $\hat{V}_n$ remains an open question.

3. Monte Carlo

In this Section we report on a brief Monte Carlo experiment designed to evaluate the performance of the Falstaff estimator of location. We limit the choice of error densities to the Student t family in order to exploit a simple normal/independent variance reduction technique. Instead of simply simply generating t-variates and directly computing mean squared errors of the estimators we generate observations from the model $y_i = z_i/v_i$ for $z_i \sim \mathcal{N}(0,1)$ and $v_i \sim \sqrt{\chi^2_\nu/\nu}$. This enables us to compute the optimal (weighted least squares) estimator *conditional of the v_i's*. Since this estimator as a known distribution we can remove this source of variability from the Monte Carlo and focus on the departure of our estimators from this idealized, but obviously unattainable estimator. This is most easily accomplished by replacing the realized n-vector y by the standardized vector $\tilde{y} = (y - \hat{\mu})/\hat{\sigma}$ where $\hat{\mu} = (v'v)^{-1}v'z$ and $\hat{\sigma}^2 = z'(I - v(v'v)^{-1}v')z/(n-1)$. See, e.g. Simon (1976) for further details.

We consider three choices of the degrees of freedom parameter of the t-distribution: $\nu \in \{1, 3, 5\}$ and six choices of the sample size $\{10, 15, 25, 50, 75, 100\}$. The coefficient of kurtosis is unbounded for the t(1) and t(3) distributions, and equals 9 for the t(5). For each configuration we consider 9 versions of the the Falstaff estimator with the q-dimension of the augmentation matrix varying from 0 to 8. Obviously, $q = 0$ provides the benchmark, ordinary least squares estimator against which we will compare the performance of the other estimators. For each configuration the augmentation matrix, D, is generated as iid from the standard normal distribution. And 10,000 replications are done for each configuration.

Figure 1 presents the results of the simulation. Columns of the array of figures correspond to the three t distributions, and rows to the six sample sizes. The horizontal line in each figure represents performance of the sample mean, corresponding to the Falstaff estimator with $q = 0$. The curve plotted in each panel represents the performance, measured by mean squared error, of the other Falstaff estimators relative to the sample mean for each configuration. The vertical bars represent 95 percent confidence intervals for the point estimates represented by the curve. When the line drops below the horizontal line it indicates improvement in performance over that of the sample mean. Thus, for the Cauchy, $\nu = 1$, cases there is dramatic improvement over the entire range of q simulated in the experiment. For the Student on 3 degrees of freedom configurations, there is improvement for modest q when the sample size is small, and improvement over the entire range when the sample size is larger. In the last column of the array, corresponding to the Student on 5 degrees of freedom, the results show no improvement from the Falstaff estimator for the smallest sample sizes, 10 and 15, slight improvement at $n = 25$ for $q = 2$, and significant improvement for the larger sample sizes, for moderate q. These results confirm the theoretical conclusions of the second-order asymptotics, and also indicate that the choice of an optimal q is rather delicately tied to the degree of non-normality of the error density and the sample size. See Koenker and Machado (1996) for a more serious theoretical consideration of this aspect of the GMM problem.


Student t on 1 n = 10
Student t on 3 n = 10
Student t on 5 n = 10
Student t on 1 n = 15
Student t on 3 n = 15
Student t on 5 n = 15
Student t on 1 n = 25
Student t on 3 n = 25
Student t on 5 n = 25
Student t on 1 n = 50
Student t on 3 n = 50
Student t on 5 n = 50
Student t on 1 n = 75
Student t on 3 n = 75
Student t on 5 n = 75
Student t on 1 n = 100
Student t on 3 n = 100
Student t on 5 n = 100
MSE
q

4. Morals

What have we learned from this pathological parable? Our first moral is that we should be wary of reliance on conventional first-order asymptotics. It is tempting to adopt the view that our ability to consistently estimate V implies that we can safely ignore the consequences of this estimation. This is true when n is sufficiently large relative to q, but it may seriously misrepresent the situation commonly faced for moderate n. Our second moral is that it is possible to substantially improve upon the performance of the ordinary least squares estimator of the linear model when the error distribution is non-Gaussian. The Falstaff estimator is only one of many nonlinear estimators which stand ready to challenge the putative superiority of the Gauss-Markov estimator. If pressed, we would not choose Falstaff to lead us into battle, but like Jack Falstaff this portly estimator contains a little wisdom.

Finally, we must confess that the Falstaff estimator is really only a variant of the estimator proposed by Cragg (1983) designed to deal with heteroscedasticity of unknown form. While Cragg's estimator was intended to deal with *real* heteroscedasticity and exploited the existence of overidentifying restrictions based on measurable functions of the regressors already appearing in the model, Falstaff makes something similar out of thin air.

5. Acknowledgements

This paper began on a napkin at the Falstaff Cafe on the Grand Place in Brussels. We wish to express our appreciation to the proprietors of this august institution for bibulous encouragement, and more circumspectly to the NSF for support under grant SBR-93-20555, and program Praxis XXI under grant 2/2.1/ECO/28/94.

Department of Economics, University of Illinois, Champaign, IL 61820
Faculdade de Economia, Universidade Nova de Lisboa, Tr. Estevão Pinto, P1070 Lisbon

REFERENCES

CRAGG, J. (1983) More efficient estimation in the presence of heteroscedasticity, *Econometrica*, **51**, 751-764.

JUDGE, G. AND BOCK, M.E. (1978) *Implications of Pre-Test and Stein Rule Estimators in Econometrics*, North-Holland.

KAKWANI, (1967), The unbiasedness of Zellner's SUR Estimators, *J. of the Am. Stat. Association*, **82**, 141-142.

KOENKER,R., MACHADO, J.A.F., AND SKEELS, C. AND WELSH A.H. (1994) Momentary lapses: Moment expansions and the robustness of minimum distance estimation, *Econometric Theory*, **10,** 172-197.

KOENKER,R., AND MACHADO, J.A.F. (1996) GMM inference when the number of moment conditions is large, unpublished.

ROTHENBERG, T. (1984) Approximate normality of generalized least squares estimates, *Econometrica*, **52,** 811-826.

SCHMIDT, P. (1976), *Econometrics*, Dekker.

SIMON, G. (1976) Computer simulation swindles, with applications to estimates of location and dispersion, *Applied Statistics*, **25**, 266-274.

STEIN, C. (1955), Inadmissability of the usual estimator of the mean of a multivariate normal distribution, in *Proceedings of the 3rd Berkeley Symposium*, **1**, 197-206.