CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA | PORTO
Faculdade de Economia e Gestão

DOCUMENTOS DE TRABALHO **WORKING PAPERS**

ECONOMIA **ECONOMICS**

Nº 05/2008

A NOTE ON THE MONTE CARLO ASSESSMENT OF IMPULSE SATURATION WITH FAT TAILED DISTRIBUTIONS

Carlos Santos

Universidade Católica Portuguesa (Porto)

A note on the Monte Carlo assessment of Impulse Saturation with fat tailed distributions

Carlos Santos, Faculdade de Economia e Gestão,
Universidade Católica Portuguesa, and CEGE

August 8, 2008

1 Introduction

Impulse saturation (Santos, Hendry and Johansen, 2008) has become a major development in model selection in linear regression. The authors have established that a general-to-specific strategy is feasible to select from a set of T candidate indicator variables, one for each observation. Such an initial model cannot be estimated from the outset, so subset selection is used (where the subsets are sample partitions either in halves, thirds, etc), followed by searches across the union of the terminal models. For a split of $T/2$, this entails saturating half the sample and storing the significant indicators, and then examining the other half. Under the null hypothesis that no indicator matters, the impulse saturation procedure is shown to have the correct null rejection frequencies (NRFs) precluding overfitting, independently of the number of splits used for the subsets. For individual tests conducted on each indicator at a significance level α, the average retention rate is αT, matching exactly the binomial resul and showing low costs of search for low α. The asymptotic distribution of the post-selection estimators of the mean and variance, in a location-scale model with IID errors is derived, and extensive Monte Carlo evidence confirms the theoretical results.

A plethora of recent papers extends the earlier work of Santos et al. (2008): Santos (2008) discusses break tests based on the impulse saturation principle; Santos and Hendry (2008) extend the principle to develop a new test for super exogeneity. Santos and Hendry (2006) and Nielsen and Johansen (2007) extend impulse saturation to a class of dynamic models, namely those of the autoregressive type; Doornik and Sprudsz (2007) extend the analysis to study the possibility of having more variables than observations in wider settings than linear regression models.

In this paper, we explore some of the connections of impulse saturation with robust statistics (see, *inter alia*, Huber, 1980), namely with respect to efficiency results. A location-scale model is used as a baseline to assess how the impulse saturation estimator of the location parameter compares with the OLS, the

Maximum Likelihood (ML) and the Mehod of Moments (MM) estimator, for some error distributions with fat tails. Results in Santos et. al (2008) as well as in Hendry and Santos (2005) had already suggested that impulse saturation could work with nonnormal errors. Indeed, Monte Carlo evidence had shown low spurious retention rates for irrelevant indicators even when the error distribution had fat tails. Notwithstanding, in this note we are not concerned with model selection and variables retention rates, but rather with estimation properties, namely relative efficiency. In the next section, we discuss results for the Laplace distribution, whilst in section 3 we compare results with those arising from a $\mathsf{t}_{(4)}$ error distribution. In both sections, the ultimate purpose is to confront the variance of the impulse saturation estimator for the mean with the OLS estimator and with the median (which is the ML estimator in the first case, and the MM estimator in the second). Discussion is based on Monte Carlo evidence[1]. Section 4 concludes.

2 Estimation with the Laplace (Double Exponential) distribution

The impulse saturation estimator of the mean, in a simple location-scale model, of the form:

$$y_t = \mu + \varepsilon_t$$

where $\varepsilon_t \sim \mathsf{IID}$, $\mathsf{E}[\varepsilon_t] = 0$, and $\mathsf{V}[\varepsilon_t] \in R_+, \forall t$, is given by,

$$\widetilde{\mu} = \frac{\sum_{t=1}^{T_1} y_t 1_{\{|\mathsf{t}_{1,\hat{\delta}_t}|<c_\alpha\}} + \sum_{t=T_1+1}^{T} y_t 1_{\{|\mathsf{t}_{2,\hat{\delta}_t}|<c_\alpha\}}}{\sum_{t=1}^{T_1} 1_{\{|\mathsf{t}_{1,\hat{\delta}_t}|<c_\alpha\}} + \sum_{t=T_1+1}^{T} 1_{\{|\mathsf{t}_{2,\hat{\delta}_t}|<c_\alpha\}}} \tag{1}$$

as derived in Santos, Hendry and Johansen (2008). This is similar to an $(\alpha/2)$ − trimmed mean (see, *inter alia*, Stuart and Ord (1994)), where a fraction $\alpha/2$ of observations is annihilated in each tail. Nonetheless, there is a fundamental difference between the usual $\alpha/2$ trimmed mean and $\tilde{\mu}$ in (1): the trimming values r and s in the impulse saturation estimator are themselves parameters to be estimated. In that sense, for a given sample of size T, $\hat{r} = g(\mathbf{y})$ and $\hat{s} = h(\mathbf{y})$, where $\mathbf{y}$ is a $(T \times 1)$ column vector containing the y_ts. Hence, there is a selection of the order statistics in (1), and the selection algorithm is non-linear.

The OLS estimator of μ would be the sample mean:

$$\hat{\mu} = \bar{y} = \frac{\sum_{t=1}^{T} y_t}{T} \tag{2}$$

where T is the sample size. According to the Gauss-Markov theorem, $\hat{\mu}$ would be the best linear unbiased estimator of μ (BLUE).

[1] All simulations were conducted using Ox (Doornik, 2001).

Let us now suppose that

$$f(\varepsilon_t) = \frac{1}{2}e^{-|\varepsilon_t|} \tag{3}$$

that is, the errors are drawn independently from a Laplace (or double exponential) distribution, with mean zero and variance 2. More generally, we could write:

$$f(\varepsilon_t) = \frac{1}{2\phi}e^{-|\varepsilon_t-\theta|/\phi} \tag{4}$$

as the density for a Laplace random variable, where $\phi > 0$. In fact, given the location-scale model above,

$$f(y_t) = \frac{1}{2\phi}e^{-|y_t-\mu|/\phi} \tag{5}$$

The maximum likelihood estimator of μ in this case is the value that minimizes with respect to μ

$$\sum_{t=1}^{T} |y_t - \mu|$$

and is given by $\check{\mu}_{ML} = median(y_1, ..., y_T)$, a result due to Keynes (1911). For an even T, the ML estimator is the arithmetic mean of observations $\frac{1}{2}T$ and $\left(\frac{1}{2}T + 1\right)$. For odd T, $\check{\mu}_{ML} = y_{\frac{T+1}{2}}$. In both cases, we obtain unbiased estimators of μ. As discussed in Johnson, Kotz and Balakrishnan (1995), the median is not necessarily the minimum variance, linear unbiased estimator of μ. Indeed, only for $T \geq 7$ do we have $\mathsf{V}[\bar{y}] \geq \mathsf{V}[\check{\mu}_{ML}]$ (Chu and Hotelling, 1955). This result in turn shows that the arithmetic mean can't also be the BLUE estimator of μ (Johnson et al. 1995).

Govindarajulu (1966) deduces the BLUE estimator as a trimmed mean. Hence $\tilde{\mu}$, the impulse saturation estimator of the mean, is closer in spirit to the BLUE estimator.

2.1 Simulation Results for Bias and Efficiency

It is therefore of interest to compare $\tilde{\mu}$, $\hat{\mu}_{OLS}$ and $\check{\mu}_{ML}$ in terms of unbiasedness and efficiency. For this purpose we consider a Monte Carlo design where the Data Generating Proccess (DGP) is given by the location-scale model, where the errors are independent draws from a Laplace distribution with parameters $\phi = 0.5$ and $\mu = 2$, so that (5) holds.

$M = 10000$ replications are conducted. We consider significance levels of $\alpha = 0.01$, $\alpha = 0.025$ and $\alpha = 0.05$ to retain impulses, and a sample size of $T = 100$.[2] Table (1) reports the results. In order to proxy for the bias, say, of estimator $\check{\mu}$, $\mathsf{E}[\check{\mu}] - \mu$, we use the difference between the mean value of $\check{\mu}$ across

[2]Results for other sample sizes and DGP parameter values were also obtained. They are not included here as they added little value. Nonetheless, all results are available from the author on request.

$T = 100$	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
$\mathsf{E}[\breve{\mu}] - \mu$	6.9451e-005	1.9123e-005	0.00016297
$\mathsf{V}[\breve{\mu}]$	0.0045882	0.0045130	0.0043788
$\mathsf{E}[\hat{\mu}_{OLS}] - \mu$	-0.00030139		
$\mathsf{V}[\hat{\mu}_{OLS}]$	0.0050780		
$\mathsf{E}[\breve{\mu}_{ML}] - \mu$	-8.2640e-006		
$\mathsf{V}[\breve{\mu}_{ML}]$	0.0029196		

Table 1: Laplace Distribution: MC bias and variance comparison, $T = 100$

the $M = 10000$ replications and the true parameter value. In order to proxy for the estimator's variance, $\mathsf{V}[\breve{\mu}]$, we compute the variance of $\breve{\mu}$ across Monte Carlo replications.

In table (1), the Monte Carlo results for the mean bias and for the variance of both the OLS and the ML estimators do not vary with α. Indeed, neither the full sample mean (the OLS estimator) nor the median (the ML estimator) depend on the choice of the significance level used for impulse saturation. On the other hand, the impulse saturation estimate of the location parameter varies with α. We conclude from table (1) that $\mathsf{V}[\tilde{\mu}]$ decreases as α increases.

The main conclusion to draw from table (1) is that $\mathsf{V}[\tilde{\mu}] < \mathsf{V}[\hat{\mu}_{OLS}]$. For all estimators and significance levels considered, the Monte Carlo evidence suggests the unbiasedness predicted by theory (Santos et. al, 2008). Unless one uses a definition of a linear estimator different from the one implied in Johnson et al. (1995), namely precluding predetermining order statistics, one is bound to conclude the Gauss-Markov theorem does not apply with Laplace errors.

It is also worth noticing that $\mathsf{V}[\breve{\mu}_{ML}]$ is the lowest of the three presented. Defining relative efficiency as the inverse ratio of an estimator's variance to the variance of the ML estimator, and considering $T = 100$, we obtain 57.5% for the OLS estimator and 64%, 65% and 67% for the impulse saturation estimator, referring respectively to $\alpha = 0.01$, $\alpha = 0.025$ and $\alpha = 0.05$.

More interestingly, for $\alpha = 0.01$, $\mathsf{V}[\tilde{\mu}]$ is 90% of $\mathsf{V}[\hat{\mu}_{OLS}]$ when $T = 100$. These values would reveal an even greater gap with α increasing. Then, on average, our Monte Carlo results reveal that the impulse saturation estimator of the mean induces an efficiency gain of about 10% relative to the OLS estimator (this value is stable across sample sizes[3]).

In conclusion, the confrontation of mean bias and variance for the impulse saturation and the OLS estimators of the mean, reveals that the first is relatively more efficient than the second. They both lose to the ML estimator in terms of relative efficiency, but here the ML estimator requires a considerable degree of prior information about the DGP, namely with respect to the error distribution.

[3] Additional evidence is available upon request.

$T = 100$	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
$\mathsf{E}[\tilde{\mu}] - \mu$	-0.0013207	-0.0019079	-0.0016083
$\mathsf{V}[\tilde{\mu}]$	0.016288	0.016271	0.016253
$\mathsf{E}[\hat{\mu}_{OLS}] - \mu$	-0.0012718		
$\mathsf{V}[\hat{\mu}_{OLS}]$	0.020058		
$\mathsf{E}[\breve{\mu}_{MM}] - \mu$	-0.0010736		
$\mathsf{V}[\breve{\mu}_{MM}]$	0.017708		

Table 2: $\mathsf{t}_{(4)}$ Distribution: MC bias and variance comparison, $T = 100$

3 Estimation with a $\mathbf{t}_{(4)}$: simulation results on bias and efficiency

In this section, the DGP is again a location-scale model of the form:

$$y_t = \mu + \varepsilon_t$$

where ε_t are IID and $\varepsilon_t \sim \mathsf{t}_{(4)}$, $\forall t$. Hence, $\mathsf{E}[\varepsilon_t] = 0$ and $\mathsf{V}[\varepsilon_t] = 2$. We consider the impulse saturation estimator for the mean, $\tilde{\mu}$, given by (1), as well as the OLS estimator, $\hat{\mu}_{OLS} = \bar{y}$, and the method of moments estimator (which is now the sample median), $\breve{\mu}_{MM}$.

For the Monte Carlo experiments, the defaults from the previous section apply. Table (2) reports the results.

In the same way as before, the finite-sample distribution of the OLS and the MM estimators does not depend on the significance level used for impulse saturation. Therefore, their moments do not vary with α. However, the impulse saturation estimate of μ depends on α. In fact, once again we observe that $\mathsf{V}[\tilde{\mu}]$ diminishes as α increases.

Again, the impact on bias of using $\tilde{\mu}$ or $\hat{\mu}_{OLS}$ is irrelevant, as the Monte Carlo results suggest that both are unbiased.

The relative gain in efficiency from using $\tilde{\mu}$ instead of the OLS estimator is greater for the case of a $\mathsf{t}_{(4)}$ error distribution. Notice that, for $\alpha = 0.01$, the ratio of $\mathsf{V}[\tilde{\mu}]$ to $\mathsf{V}[\hat{\mu}_{OLS}]$ is 81% when $T = 100$. That is, on average there is an efficiency gain of 19% when using the impulse saturation mean estimator instead of the full sample mean. It is once more clear that the assessment of the gain would be increased for higher α (although this effect would be of 1 or 2 percentage points only[4]).

Furthermore, for a $\mathsf{t}_{(4)}$, the impulse saturation estimator is relatively more efficient than the sample median itself. Indeed, for $\alpha = 0.01$ the inverse ratio of the variance of the sample median to the variance of the impulse saturation estimator for the mean is 92%, for a sample size of $T = 100$. On average, there is a gain of about 8% in efficiency when using $\tilde{\mu}$ instead of $\breve{\mu}_{MM} = median(y_1, ..., y_T)$.

[4] Additional evidence not reported here but available from the corresponding author on request.

For this error distribution, Monte Carlo evidence suggests that:

$$\mathsf{V}[\tilde{\mu}] < \mathsf{V}[\breve{\mu}_{MM}] < \mathsf{V}[\hat{\mu}_{OLS}]$$

In conclusion, the impulse saturated mean estimator seems more robust than the OLS estimator: retaining efficiency properties under departures from the assumption of normality (see the difinition of robustness in Peracchi, 2001).

4 Conclusion

The main conclusion to draw from this note is that the impulse saturation estimator of the location parameter is more efficient the OLS estimator for a some classes of nonnormal error distributions. Furthermore, although it can be less efficient than the ML estimator, maximum likelihood requires a high degree of prior information on the distribution of the unobserved errors. Therefore, it seems advisable, at least when the residuals distribution suggests the erros might not be normally distributed, to use impulse saturation. This result enforces, from the point of view of efficiency, the good properties of this procedure: it is safe to assume it is more efficient under nonnormality, and we had already established it did not cause over selection (Santos et. al, 2008).

An open point of debate has to do with the definition of a BLUE estimator. In the sense that ordering is not a linear operation, OLS would remain BLUE. This, however, is not the interpretation in Johnson et. al (1995). Notwithstanding, we believe linearity is not such an interesting property once we have established that the impulse saturation estimator is unbiased and more efficient than OLS under some nonnormal error distributions.

REFERENCES

Chu, J. T. and Hotelling, H. (1955) The Moments of the Sample Median, *Annals of Mathematical Statistics*, 26, 593-606.

Doornik, J. A. (2001) *OX, An Object Oriented Matrix Programming Language*, 4th edition, London: Timberlake Consultants Press.

Doornik, J. A. and U. Sprudsz (2007) Econometric Modelling when there are more Variables than Observations, Nuffield College, mimeo

Govindarajulu, Z. (1966) Best Linear Estimates under Symmetric Censoring of the Parameters of a Double Exponential Population, *Journal of the American Statistical Association*, 61, 248-258 (Correction: 71, 255)

Huber, P. J. (1980) *Robust Statistics,* New York, Wiley Inter-Science

Hendry, D. F. (2000) *Econometrics: Alchemy or Science?*, Oxford: Oxford University Press

Hendry, D. F. and Reade, J. (2005) Problems with Model Averaging with Dummy Variables, Univ. of Oxford. Department of Economics, mimeo

Hendry, D. F. and Santos, C. (2005) Regression Models with Data-Based Indicator Variables, *Oxford Bulletin of Economics and Statistics*, 67, 571-595

Hendry, D. F.and Santos, C. (2008) An Automatic Test for Super Exogeneity, Oxford University, Department of Economics, mimeo

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions* –2, 2nd edition, New York: John Wiley and Sons.

Nielsen, B. and S. Johsanen (2007) Saturation by Indicators in Autoregressive Models, paper presented at the Conference in Honour of David F. Hendry, Oxford

Peracchi, F. (2001) *Econometrics*, New York: John Wiley and Sons.

Santos, C. (2008) Impulse Saturation Break Tests, *Economic Letters*,98 (2), pp. 136-143

Santos, C. and Hendry, D. F. (2006) Saturating Autoregressive Models, *Notas Económicas*, 24 , pp.8-19

Santos, C., Hendry, D. F. and S. Johansen (2008) Automatic Selection of Indicators in Fully Saturated Regressions, *Computational Statistics*, 23(2), 317-335

Stuart, A. and Ord, K. (1994) *Kendall's Advanced Theory of Statistics*, Volume 1: Distribution Theory, 6th edition, London: Arnold.