Notes on Generalized Method of Moments
Estimation

© Bronwyn H. Hall

March 1996 (revised February 1999)

1. Introduction

These notes are a non-technical introduction to the method of estimation popularized by Hansen and others, commonly referred to as Generalized Method of Moments. Theoretical issues are mentioned but not discussed thoroughly and no proofs are given (nor are assumptions fully stated): see Hansen (1982), Chamberlain (1987), and the texts by Ruud (2000) and Davidson and MacKinnon (1993) for details on these. The emphasis here is on explication of the basic method and notes on how to perform this kind of estimation in TSP.

2. Method of Moments

This is a consistent, but not generally efficient estimator. Assume the data $y = (y_1, y_2, \ldots, y_n)$ are generated by a probability distribution indexed by a parameter vector $\theta$ with $k$ elements.

In the method of moments, $\theta$ is estimated by computing $k$ sample moments of $y$, setting them equal to population moments derived from the assumed probability distribution, and solving for $\theta$. Because sample moments are generally consistent estimators of population moments (under some regularity conditions), $\hat{\theta}$ will be consistent for $\theta$ (if it exists).

2.1. Example 1: Sample Mean

The population moment is $\mu$, the expectation of $y$ and the sample moment is the sample mean of $y$. $y$ is assumed to be independently and identically distributed
with mean $\mu$ and finite variance. Then we can estimate $\mu$ by setting it equal to the sample first moment:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

### 2.2. Example 2: Gamma Distribution

The data $y$ are assumed to be generated according to the probability density

$$f(y_i) = \frac{\lambda^p}{\Gamma(p)} e^{-\lambda y_i} y_i^{p-1}$$

There are two parameters to be estimated, $\lambda$ and $p$. The first two population moments are the following:

$$E(y_i) = \frac{p}{\lambda} \quad \text{and} \quad V(y_i) = \frac{p}{\lambda^2}$$

The first two sample moments are the mean $\bar{y}$ and the variance $s_y^2$. We set these equal to the population moments and solve for $\lambda$ and $p$:

$$\hat{\lambda} = \frac{\bar{y}}{s_y^2} \quad \text{and} \quad \hat{p} = \frac{\bar{y}^2}{s_y^2}$$

These estimates are consistent if the sample variance of $y$ is finite; for example,

$$p \lim \hat{\lambda} = p \lim \frac{\bar{y}}{s_y^2} = E(Y)/V(Y) = \lambda$$

### 3. Generalized Method of Moments - Theory

In the previous two examples, we had exactly as many moment conditions as parameters, so we simply solved the $k$ equation in $k$ unknowns to obtain our estimates. What happens when we have more moment conditions than parameters? How can we combine them optimally? That is what the GMM estimator does. It estimates the parameter vector by minimizing the sum of squares of the differences between the population moments and the sample moments, using the variance of the moments as a metric. This is the minimum variance estimator in the class of estimators that use these moment conditions. Here is some notation:

Write the theoretical moment condition as

$$E[m_j(y_i, \theta)] = 0 \quad j = 1, ..., J \quad \text{and} \quad k < J \text{ elements in } \theta$$
Then the sample moments of the data are

\[ \overline{m}_j(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_j(y_i, \theta) \]

The set of J equations \( \overline{m}_j(\theta) = 0 \) will not have a unique solution for \( \theta \) in general. Now define

\[ \overline{m}(\theta) = [\overline{m}_1(\theta), \overline{m}_2(\theta), ..., \overline{m}_J(\theta)] \]

The GMM estimator is then

\[ \hat{\theta} = \arg \min_\theta \overline{m}(\theta)'W^{-1}\overline{m}(\theta) \]

where \( W \) is the asymptotic variance of \( \overline{m}(\theta) \) or an estimate. This estimator is similar to, but not the same as, a generalized least squares (GLS) estimator. When \( W \) is either the true variance or a consistent estimate of it, \( \hat{\theta} \) is consistent and asymptotically efficient for \( \theta \), in the class of estimators that use only the information in the moment conditions (for example, there may be a more efficient ML estimator if the data \( y \) are assumed to be generated by a specific probability distribution - see Davidson and MacKinnon, pp. 597-600 for details). That is, if \( W \) is positive definite, \( p \lim \overline{m} = 0 \), and certain regularity conditions hold,

\[ p \lim \hat{\theta} = \theta \]

The asymptotic variance of this estimator is

\[ \Sigma(\theta) = [G'W^{-1}G]^{-1} \]

where \( G \) is the matrix of derivatives of the moment conditions with respect to the parameters; the jth row of \( G \) is given by

\[ G^j = \frac{\partial \overline{m}_j}{\partial \theta} \]

**Comments:**

- When \( J = k \), the choice of \( W \) does not matter, since \( \overline{m} \equiv 0 \) at the solution.
- When \( J < k \), \( \theta \) is not identified.
• Under the null that the moment restrictions hold, the criterion function evaluated at the estimated $\hat{\theta}$ has the chi-squared distribution with degrees of freedom equal to the number of independent moment conditions ($J$) less the number of estimated parameters ($k$). This is the basis of the omnibus specification test called the J-test or the Sargan test. TSP reports the value of this test statistic and its p-value in the GMM output.

• $m_j(y_i, \theta)$ can be an arbitrary nonlinear function of data and parameters.

3.1. Example 3: Instrumental Variable Estimation with Homoskedastic Disturbances

The model is the usual linear model:

$$y_i = X_i \beta + e_i \quad i = 1, ..., n \quad Var(e_i) = \sigma^2$$

$$E[e_i | X_i] \neq 0 \text{ but } E[e_i | Z_i] = 0 \text{ for a set of } J \text{ instruments } Z_i$$

Assume $J > k$. Then the $J$ moment conditions are

$$\bar{m} = Z'e = 0$$

We can estimate the variance of $\bar{m}$ by the following:

$$W = Var(Z'e) = E[Z'e'e'Z] - E[Z'e]E[e'Z] = E[Z'\sigma^2 I Z] = \sigma^2 Z'Z$$

Therefore the GMM estimator using these moment conditions, is given by

$$b_{IV} = \arg\min_{\beta} e'Z'(Z'Z)^{-1}Z'e$$

Note two things: we have dropped $\sigma^2$ because it is a scalar and does not affect the minimization\(^1\) and the resulting criterion function is the instrumental variable criterion function. The solution to this minimization problem has a first order condition given by

\(^1\)This is not strictly true in finite samples. If we are estimating $\sigma^2$ along with the rest of the model, because the estimate will depend on $b_{IV}$, it will make a difference if we include it in the criterion function. If we do include it, and we update it continuously during the iterations, the resulting estimator will be LIML rather than IV; they are asymptotically equivalent but may differ in finite samples.
\[ \frac{\partial e'}{\partial \beta} Z(Z'Z)^{-1}Z'e = -X'Z(Z'Z)^{-1}Z'(y - Xb_{IV}) = 0 \]

which yields the IV estimator:

\[ b_{IV} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \]

An estimate of the variance of this estimator is given by

\[ \Sigma(b_{IV}) = [G'\hat{W}^{-1}G]^{-1} = [X'Z(\sigma^2Z'Z)^{-1}Z'X]^{-1} \]

This is the estimator computed by INST (TSLS) in TSP for linear two stage least squares or instrumental variables.

In general, the model can be nonlinear with additive disturbances, so that \( e = h(y, X, \beta) \). In that case the estimator is computed by iterative methods and the variance is

\[ \Sigma(b_{IV}) = \left[ \frac{\partial e'}{\partial \beta} Z(\sigma^2Z'Z)^{-1}Z' \frac{\partial e}{\partial \beta} \right]^{-1} \]

This is the estimator computed by LSQ in TSP for nonlinear two stage least squares.

### 3.2. Example 4: Three Stage Least Squares with Heteroskedastic Disturbances (the Hansen-Singleton Estimator)

This is the estimator commonly known in the econometric literature as GMM. It is usually applied in situations where there are several equations with correlated disturbances in the model (e.g., multiple assets as in Hansen-Singleton, or multiple time periods for panel data), and where it would be inappropriate to impose homoskedasticity on the disturbances. For simplicity, I discuss the case where the same set of instruments are available for all equations. See Appendix B to Mairesse and Hall (1996) for details on what to do when you want to use different instruments in different equations (use the MASK option in GMM in TSP).

The moment conditions for this estimator can be written

\[ m_i = z_i \otimes e_i \]
where $z_i$ is a vector of $z$’s for the $i$th firm (including all time periods) and $e_i$ is a vector of disturbances for the $i$th firm. The data for each firm are assumed to be independently but not identically distributed conditional on the $z$’s. The moment conditions for estimation are therefore

$$\bar{m} = \sum_{i=1}^{n} (z_i \otimes e_i)$$

The variance of $\bar{m}$ is $Var[\sum_{i=1}^{n} z_i \otimes e_i]$ and this can be approximated by the following expression (given the conditional independence of the firms):

$$W = \frac{1}{n} \sum_{i=1}^{n} (z_i \otimes e_i)(z_i \otimes e_i)' = \frac{1}{n} \sum_{i=1}^{n} z_i z_i' \otimes e_i e_i'$$

In general, this formula allows for heteroskedasticity of the $e$’s, but when they are homoskedastic, it reduces to the following:

$$E[W|Z] = \frac{1}{n} \sum_{i=1}^{n} z_i z_i' \otimes E[e_i e_i'|Z] = \frac{1}{n} Z'Z \otimes \Sigma$$

where $\Sigma$ is the covariance matrix of the disturbances $e$. Thus one can show that GMM with the assumption of homoskedasticity for this model is the same as three-stage least squares estimation. In TSP, it can be done using 3SLS (LSQ) by listing the $z$’s in the instrument list and defining the equations of the model using FRML.

However, under heteroskedasticity, the variance of the $e$’s does not factor out of this expression, and we obtain the proper GMM estimator. Although estimation is straightforward (use the GMM procedure), several difficulties arise:

- Estimates of the variance $W$ based on fourth moments converge slowly, so you may need quite large samples to get consistent estimates. In finite samples, three-stage type estimates may be preferred if the amount of heteroskedasticity is not too large.

- Like 2SLS and 3SLS, finite sample estimates are biased toward OLS (estimates without instruments), but in this case finite may be fairly large.

- As in the case of 2SLS and 3SLS, finite sample estimates are sensitive to the choice of normalization (which, if any, endogenous variable has its coefficient set to unity) unless the estimate of $W$ used is continuously updated during iteration. A LIML-type estimator is available that may be preferred in this case.
4. References


