# NOTES: Generalized Method of Moments (GMM)

## 1 Introduction

To describe the GMM estimation methodology in some detail, I follow Hamilton (1994). Let us define an $(h \times 1)$ vector of observables $w_t$, an $(a \times 1)$ vector of coefficients $\theta$ with the true value $\theta_0$, and an $(r \times 1)$ vector valued function $h(\theta, w_t)$. $h(.)$ can be viewed as a residual from a model. Orthogonality conditions are defined as:

$$E[h(\theta, w_t)] = 0. \tag{1}$$

The sample equivalent of the orthogonality conditions (1) is given by

$$g(\theta, Y_T) \equiv \frac{1}{T} \sum_{t=1}^{T} h(\theta, w_t) \tag{2}$$

where $Y_T$ is an $(Th \times 1)$ vector $[w_1, ..., w_T]'$. Note that $g : \ R^a \rightarrow R^r$.. The idea behind GMM is to choose $\theta$ so as to make the sample moment $g(\theta, Y_T)$ as close as possible to the population moment of 0.

**Definition** The GMM estimator $\hat{\theta}_T$ is the value of $\theta$ that minimizes the scalar

$$Q(\theta, Y_T) = [g(\theta, Y_T)]' W_T [g(\theta, Y_T)] \tag{3}$$

where $\{W_T\}_{T=1}^{\infty}$ is a sequence of positive definite weighting matrices. $W_T$ is in general a function of the data $Y_T$.

For $a = r$, $g(\hat{\theta}_T, Y_T) = 0$. For $r > a$ (i.e. the number of restrictions greater then the number of parameters), $g(\hat{\theta}_T, Y_T) \neq 0$ in general. A variety of estimators can be viewed as examples of GMM: OLS, instrumental variable estimator, 2SLS, estimators of dynamic rational expectations models, etc.

## 2 Optimal weighting matrix

Under fairly general continuity and moment conditions, $\hat{\theta}_T$ minimizing (3) is a consistent estimator of $\theta_0$. However, the variance of the GMM estimator can be minimized for a proper choice of the weighting matrix.

**Theorem, Hansen (1982):**

$$S \equiv \sum_{\nu=-\infty}^{\infty} \Gamma_\nu \ \text{ where } \ \Gamma_\nu = E[(h(\theta_0, w_t)) \ (h(\theta_0, w_{t-\nu}))]'.$$

The minimum asymptotic variance for the GMM estimator $\hat{\theta}_T$ is obtained when $\hat{\theta}_T$ is chosen to minimize (3) with $W_T = S^{-1}$.

It can be shown that $S \equiv \lim_{T\to\infty} TE[(g(\theta_0, Y_T))\ (g(\theta_0, Y_T))]'$ i.e. $S$ is the asymptotic variance of the sample mean of $h(\theta_0, w_t)$. In a single variate case, consider $g = \sum X/T$ for some random variable $X$ with variance $\sigma^2$. $V(g) = \frac{1}{T^2}V(\sum X) = \frac{T}{T^2}\sigma^2 = \frac{\sigma^2}{T}$. $V(g)$ is our $S/T$ and $V(h) = TV(g)$.

**i. No serial correlation:** $\Gamma_\nu = 0$ for $\nu \neq 0$. White(1980) shows that the heteroskedasiticity-consistent estimator of $s$ is $S_T^* = 1/T \sum_{t-1}^{T}(h(\theta_0, w_t))\ (h(\theta_0, w))'$.

**ii. Serial correlation:** $\Gamma_\nu \neq 0$ for $\nu \neq 0$.

a. Kernel based (non-parametric) estimation.

$$\hat{S} = \hat{\Gamma}(0) + \sum_{j=1}^{T-1} k(j, q)(\hat{\Gamma}(j) + \hat{\Gamma}(j)')$$

where $k$ is the kernel set to weight the covariances so that $\hat{S}$ is positive semidefinite. $q$ is the bandwidth, it determines how the weights change with the lags of the $S$ estimator. Estimators differ by the choice of the number lags, by the kernel function and by the bandwidth - see Newey and West (1987,1994), Andrews and Mohanan (1992) for examples of kernel based, heteroskedasticity and autocorrelation consistent (HAC) estimation of $S$.

b. Parametric estimation. den Haan and Levin (1996) impose VAR structure on the variance-covariance matrix of residuals, the result is a VARHAC estimator.

You can also choose if you want to use residuals filtered via an AR process. This filtering is called *prewhitening*. You can also opt for iterative vs. non-iterative GMM procedure. The non-iterative procedure is a two-step procedure when you get a GMM estimator $\hat{\theta}_T^1$ using $W_T = I$ and then re-optimize using $S_(\hat{\theta}_T)$ and get $\hat{\theta}_T^2$. If you re-optimize again using $\hat{\theta}_T^2$ to get $\hat{\theta}_T^3$ and so on, you are using the iterative procedure. Convergence is achieved according to some criteria related to the value of the objective function of the gradient in the optimization problem. Overall, if your software allows it, choose iterative VARHAC with prewhitening.

# 3 Asymptotic distribution of the GMM estimates

Let $\hat{\theta}_T$ is the argmin of (3) with $W_T = S^{-1}$. The first order conditions of this minimization problem are:

$$\left[\frac{\partial g(\theta, Y_T)}{\partial \theta'}|_{\theta=\hat{\theta}_T}\right]' S_T^{-1}[g(\hat{\theta}_T, Y_T)] = 0 \qquad (4)$$

where the first part of the formula has dimensions $(a \times r)$, the second $(r \times r)$ and the third $(r \times 1)$, respectively. Now define $h_t = h(\theta_0, w_t), E[h_t] = 0, V[h_t] = S$, and $g_T = \frac{1}{T}\sum_t h_t$.

The Central limit theorem implies that

$$Z_T = \frac{g_T - 0}{S/\sqrt{T}} \sim N(0,1).$$

This in turn implies that

$$\sqrt{T}g(\theta_0, Y_T) \rightarrow N(0,S).$$

Let us now rewrite the FOCs (4) as

$$\left[\frac{\partial g(\theta, Y_T)}{\partial \theta'}|_{\theta=\theta_0}\right]' S_0^{-1}[g(\hat{\theta}_T, Y_T)] = 0$$

Using a Taylor series expansion of $\sqrt{T}g(\theta_T, Y_T)$ around $\theta_0$ and implicitly defining derivatives $J_0$ results in:

$$J_0 S_0^{-1}[\sqrt{T}g(\theta_0, Y_T) + J_0'\sqrt{T}(\hat{\theta}_T - \theta_0)] = 0$$

Rearranging implies

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = (J_0 S_0^{-1} J_0')^{-1} J_0 S_0^{-1} \sqrt{T}g(\theta_0, Y_T).$$

Note that the variance of $\sqrt{T}g(\theta_0, Y_T)$ is $S_0$ and therefore

$$\sqrt{T}(\hat{\theta}_T - \theta_0) = N(0, V_0)$$

where $V_0 = (J_0 S_0^{-1} J_0')$ and can be estimated using consistent estimators. Finally, with some abuse of notation, we can write

$$\hat{\theta}_T \approx N(\theta_0, \frac{\hat{V}_T}{T}).$$

Define

$$Q(\theta_T, Y_T) = [g(\theta, Y_T)]' S_T^{-1}[g(\theta, Y_T)]. \tag{5}$$

Hansen(1982) shows that $TQ(\hat{\theta}_T, Y_T) \rightarrow \chi(r-a)$.This is the well-known test of over-identyfying restrictions.

# 4 Application

Hansen and Singleton (1982) apply the GMM methodology to test restrictions implied by the Lucas (1978) Consumption based Capital Asset Pricing Model (CCAPM) with the power utility function characterizing consumer preferences. The model implied residuals come from the first order conditions of the optimization problem maximizing the lifetime utility $\sum_{\tau=0}^{\infty} E_t[u(c_{t+\tau})]$ with the constant relative risk aversion utility function $u(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma}$. The first order conditions (the Euler equations) are given by

$$1 = \beta E_t[(1 + r_{i,t+1})(\frac{c_{t+1}}{c_t})^{-\gamma}]$$

for where $r_{i,t+1}$ is a rate of return on an asset and there are $i = 1, ..., m$ assets. To get an estimation counter part of the first order conditions, I take their unconditional expectation and introduce vectors of instruments $x_t^i$ as $3 \times 1$ vectors of $[1, c_t/c_{t-1}, , r_{i,t}]'$:

$$E[\{1 - \beta(1 + r_{i,t+1})(\frac{c_{t+1}}{c_t})^{-\gamma}\}x_t] = 0$$

The vector of parameters $\theta = [\beta, \gamma]$ where $\beta$ is the discount factor and $\gamma$ is the coefficient of relative risk aversion. $r_{i,t}$ is either the value-weighted or equally-weighted rate of return on the US stock market. $c$ is the per-capita consumption of non-durables and services. $x_t^i$ are a $5 \times 1$ vectors of instruments for the number of lagged instruments $l = 1$, for which $r = 2l + 1$ and the number of over-identifying restrictions is $r - a$. So, for one lag and $a = 2$, we have 3. This is for the number of lagged instruments $l = 1$, for which $r = 2l + 1 = 3$ and the number of over-identyfying restrictions is $r - a = 1$. So, for one lag and $a = 2$, we have 1 degree of freedom. Results are in Table 1 of Hansen and Singleton (1982) ($\alpha = -\gamma$, so the RRA coefficient is given by $-\alpha$). There is no equity premium puzzle here since there is no risk free rate. Other researchers have found the equity premium puzzle when the restrictions using the risk-free rate were imposed. To see if the puzzle still persists, I use the following error terms:

$$h(\theta, w_t) = \begin{bmatrix} \{1 - \beta(1 + r_{e,t+1})(c_{t+1}/c_t)^{-\gamma}\}x_t^1 \\ \{1 - \beta(1 + r_{f,t+1})(c_{t+1}/c_t)^{-\gamma}\}x_t^2 \end{bmatrix}. \quad (6)$$

$x_t^i, i = 1, 2$ are a $5 \times 1$ vectors of instruments $[1, c_t/c_{t-1}, c_{t-1}/c_{t-2}, r_{et}, r_{e,t-1}]'$ for the first equation and $[1, c_t/c_{t-1}, c_{t-1}/c_{t-2}, r_{ft}, r_{f,t-1}]'$ for the second equation, respectively. $w_t \equiv [r_{e,t+1}, r_{f,t+1}, c_{t+1}/c_t, x_t']]$.

Here I intend to confirm existence of the equity premium puzzle by estimating the standard power utility model by GMM using an updated dataset. To test restrictions (6), I use monthly US data from March 1967 to September 2008. For consumption, the data is taken from the St. Louis FED web page and for returns from the European Central Bank. The weighting matrix estimate $S$ is robust to heteroskedasticty (White correction) and autocorrelation (Barlett kernel). The parameter estimates are as follows:

| Parameter | Estimate | Error | t-statistic | P-value |
|---|---|---|---|---|
| $\beta$ | .87 | .262E-02 | 331.30 | [.00] |
| $\gamma$ | -1.38 | 1.65 | -.83 | [.40] |

We can see that the risk aversion is negative though theory suggests it is positive. The sign is not a problem however since it is insignificant. This seems to suggest that a highly significant risk aversion coefficient is not needed anymore to match the data, especially after a big drop in stock prices in October 2008. On the other hand, the Hansen J statistic is 153.82.It has a chi-square distribution with 10-2=8 degrees of freedom. The corresponding p-value for the test of over-identifying restrictions is then 0, which means the the the model is still rejected. Therefor the equity premium puzzle is weaker the before but it has not quite disappeared yet.

# References


Andrews, D.W.K. and Monahan, J.C. (1992).‘An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator,’ *Econometrica* 60, 953-966.

den Haan, W. and Levin, A. (1996). ‘Inferences from parametric and non-parametric covariance matrix estimation procedures,’ *Working paper t0195 , NBER.*

Hamilton, J.D. (1994).‘Time Series Analysis,’ *Princeton University Press.*

Hansen, L.P. and Singleton K.J. (1982).‘Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models,’ *Econometrica* 50, 1269-1286.

Lucas,R.E. (1978).‘Asset prices in an Exchange Economy,’ *Econometrica* 46, 1429-1445.

TABLE I

INSTRUMENTAL VARIABLE ESTIMATES FOR THE PERIOD 1959:2–1978:12

| Cons | Return | NLAG | $\hat{\alpha}$ | $\widehat{SE}(\hat{\alpha})$ | $\hat{\beta}$ | $\widehat{SE}(\hat{\beta})$ | $\chi^2$ | DF | Prob |
|---|---|---|---|---|---|---|---|---|---|
| NDS | EWR | 1 | −.9457 | .3355 | .9931 | .0031 | 4.9994 | 1 | .9746 |
| NDS | EWR | 2 | −.9281 | .2729 | .9929 | .0031 | 7.5530 | 3 | .9438 |
| NDS | EWR | 4 | −.7895 | .2527 | .9925 | .0031 | 9.1429 | 7 | .7574 |
| NDS | EWR | 6 | −.8927 | .2138 | .9934 | .0030 | 15.726 | 11 | .8484 |
| NDS | VWR | 1 | −.9001 | .3130 | .9979 | .0025 | 1.1547 | 1 | .7174 |
| NDS | VWR | 2 | −.8133 | .2298 | .9981 | .0025 | 3.2654 | 3 | .6475 |
| NDS | VWR | 4 | −.6795 | .1855 | .9973 | .0024 | 6.3527 | 7 | .5008 |
| NDS | VWR | 6 | −.7958 | .1763 | .9980 | .0023 | 14.179 | 11 | .7767 |
| ND | EWR | 1 | −.9737 | .1245 | .9922 | .0031 | 5.9697 | 1 | .9854 |
| ND | EWR | 2 | −.9664 | .1074 | .9919 | .0031 | 8.9016 | 3 | .9694 |
| ND | EWR | 4 | −.9046 | .0926 | .9918 | .0031 | 11.084 | 7 | .8650 |
| ND | EWR | 6 | −.9466 | .0793 | .9422 | .0030 | 15.663 | 11 | .8459 |
| ND | VWR | 1 | −.8985 | .1057 | .9971 | .0025 | 1.5415 | 1 | .8756 |
| ND | VWR | 2 | −.8757 | .0856 | .9974 | .0025 | 3.2654 | 3 | .6475 |
| ND | VWR | 4 | −.8174 | .0742 | .9967 | .0024 | 7.8776 | 7 | .5008 |
| ND | VWR | 6 | −.8514 | .0629 | .9973 | .0024 | 14.938 | 11 | .8147 |

The estimates of $\alpha$ range from −.95 to −.68 when ND is used as the measure of consumption, and from −.97 to −.82 when ND is used as the measure of consumption. The estimated standard errors for $\alpha$, $\widehat{SE}(\hat{\alpha})$, are smaller when consumption is measured as ND than when consumption is measured as NDS. As expected, all of the estimates of $\beta$ exceed .99 but are less than unity. The chi-square tests are also displayed in Table I, where the number of overidentifying restrictions is indicated by DF and Prob is the probability that a $\chi^2(\mathrm{DF})$ random variate is less than the computed value of the test statistic under the hypothesis that the restrictions (3.1) are satisfied. These tests provide greater evidence against the model when EWR is included as the return, and when the instrument vector is formed from a small number of lagged values of $x$.

For comparison, we present some results in Table II from estimating $\alpha$ and $\beta$ using the method of maximum likelihood under the assumption that $x$ is lognormally distributed. They were obtained using the procedure described in Section 4 assuming that log $x$ has a sixth-order vector autoregressive representation. The corresponding estimates of $\alpha$ and $\beta$ from the two methods of estimation are similar. However, the estimated standard errors of $\alpha$ and $\beta$ from the instrumental variables procedure are smaller than the corresponding standard errors from the maximum likelihood procedure.[13] Three possible explanations for this result are that the asymptotic standard errors are being estimated imprecisely, the economic model of stock returns is misspecified, or the auxiliary assumptions underlying the maximum likelihood procedure are incorrect. The maximum likelihood procedure assumes that $x$ is lognormally distributed and that the lag length specification of the vector autoregression is correct. Since the

[13] Analytical differentiation, as opposed to numerical differentiation, was used to calculate the standard errors.