

GMM and the CAPM

Non-normal and Non-i.i.d. Returns

- Why consider this? Normality is not a necessary condition.
 - Indeed, asset returns are not normally distributed (see e.g. Fama 1965, 1976)
 - Returns appear to have fat tails (see e.g. 1970's literature on mixtures of distribution – Stan Kon.)
 - Recall that returns have temporal dependence.
- In this environment, the CAPM will not hold, but we may want to examine empirical performance.

IV and GMM Estimation

- GMM estimation is essentially instrumental variables estimation where the model can be nonlinear. Our plan:
 - Introduce linear IV estimation.
 - Introduce linear test of overidentifying restrictions.
 - Generalize to nonlinear models.

Linear, Single Equation IV Estimation

- Suppose there is a linear relationship between y_t and the vector \mathbf{x}_t such that

$$y_t = \mathbf{x}_t' \boldsymbol{\theta}_0 + \varepsilon_t(\boldsymbol{\theta}_0)$$

- Where \mathbf{x}_t is $N_x \times 1$ and $\boldsymbol{\theta}_0$ is an $N_x \times 1$ parameter vector. Stacking T observations yields

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}(\boldsymbol{\theta}_0)$$

- Where \mathbf{y} is $T \times 1$, \mathbf{X} is $T \times N_x$, $\boldsymbol{\theta}_0$ is $N_x \times 1$, and $\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0)$ is $T \times 1$.

The System

$$\begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1N_x} \\ \vdots & & \vdots \\ x_{T1} & \cdots & x_{TN_x} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{N_x} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$$

- Note that ε is really a function of the parameter vector, so $\varepsilon(\hat{\theta}) = \hat{\varepsilon}$.
- For simplicity, assume for now that the errors are serially uncorrelated and homoskedastic:

$$\text{var}(\varepsilon(\theta_0)) = E[\varepsilon\varepsilon'] = \sigma^2 \mathbf{I}_T$$

- \mathbf{I}_T is a $T \times T$ identity matrix.

Instruments

- If you observe the regressors, x 's, there would be no need to do IV estimation.
- You would just use the x 's and run a standard regression.
- If you don't see the x 's, then you are in a situation where IV estimation is the most useful. You might have to use a general version of IV estimation, of which least squares is a special case.
- Usually, the instruments are a way to bring more structure to the estimation procedure and so get more precise parameter estimates.

Examples

- One place where IV estimation could be useful is if the regressors were correlated with the errors but you could find instruments correlated with the regressors but not with the errors.
 - If the instruments are uncorrelated with the regressors, they are never any help.
- Another example is estimating the non-linear rational expectations asset pricing model, where elements of the agents' information sets are used as instruments to help pin down the parameters of the asset pricing model.

Instruments

- There are N_Z instruments in an $N_Z \times 1$ column vector \mathbf{z}_t' , and there is an observation for each period, t . Hence, the matrix of instruments, \mathbf{Z} , is a $T \times N_Z$ matrix:

$$\begin{bmatrix} z_{11} & \cdots & z_{1N_Z} \\ \vdots & & \vdots \\ z_{T1} & \cdots & z_{TN_Z} \end{bmatrix}$$

- The instruments are contemporaneously uncorrelated with the errors so that $E[\mathbf{z}_t' \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}_0)] = \mathbf{0}$ is an $N_Z \times 1$ vector of zeros.

Usefulness of Instruments

- This depends on whether they can help identify the parameter vector.
- For instance, it might not be hard to generate instruments that are uncorrelated with the disturbances, but if those instruments weren't correlated with the regressors, the IV estimation would not help identify the parameter vector.
- This is illustrated in the formulation of the IV estimators.

Orthogonality Condition

- The statement that a particular instrument is uncorrelated with an equation error is called an *orthogonality condition*.
- IV estimation uses the N_Z available orthogonality conditions to estimate the model.
- Note that least squares is a special case of IV estimation because the first-order conditions for least squares are:

$$E[\mathbf{x}_t' \boldsymbol{\varepsilon}(\boldsymbol{\theta}_0)] = \mathbf{0}$$

an $N_x \times 1$ vector of zeros.

- Least squares is like an exactly identified IV system where the regressors are also the instruments.

The Error Vector

- Given an arbitrary parameter vector, $\boldsymbol{\theta}$, we can form an error vector $\boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) \equiv \mathbf{y}_t - \mathbf{x}'_t \boldsymbol{\theta}$, and write it as a stacked system:

$$\begin{bmatrix} \boldsymbol{\varepsilon}_1(\boldsymbol{\theta}) \\ \vdots \\ \boldsymbol{\varepsilon}_T(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} - \begin{bmatrix} x_{11} & \cdots & x_{1N_X} \\ \vdots & & \vdots \\ x_{T1} & \cdots & x_{TN_X} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{N_X} \end{bmatrix}$$

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}) \equiv \mathbf{Y} - \mathbf{X}\boldsymbol{\theta}$$

Orthogonality Conditions cont...

- Recall that we had N_Z instruments. Define an $N_Z \times 1$ vector

$$\mathbf{f}_t(\boldsymbol{\theta}) \equiv \mathbf{z}_t' \boldsymbol{\varepsilon}_t(\boldsymbol{\theta})$$

$$\begin{bmatrix} z_{1t} \\ \vdots \\ z_{N_Z t} \end{bmatrix} [\boldsymbol{\varepsilon}_t(\boldsymbol{\theta})]$$

- The expectation of this product is an $N_Z \times 1$ vector of zeroes at the true parameter vector $\boldsymbol{\theta}_0$: $E[\mathbf{f}_t(\boldsymbol{\theta})] = \mathbf{0}$.

Overidentification

- We have N_X parameters to estimate and N_Z restrictions, where $N_Z \geq N_X$.
- The idea is to choose parameters, $\hat{\boldsymbol{\theta}}$, to satisfy this orthogonality restriction as closely as possible.
- If $N_Z > N_X$, unless the model were literally true, we won't be able to satisfy the restriction exactly – in finite samples, we won't be able to do so even if the model is true.
 - In this case the model is overidentified.
- When $N_Z = N_X$, we can choose $\hat{\boldsymbol{\theta}}$ to satisfy the restriction exactly. Such a system is exactly identified (i.e. OLS).

Constructing the Estimator

- We don't see $E[\mathbf{f}_t(\boldsymbol{\theta}_0)]$, so we must work instead with the sample average.
- Define $\mathbf{g}_T(\boldsymbol{\theta})$ to be the sample analog of $E[\mathbf{f}_t(\boldsymbol{\theta}_0)]$:
$$\mathbf{g}_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T \mathbf{f}_t(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t' \boldsymbol{\varepsilon}_t(\boldsymbol{\theta}) = T^{-1} \mathbf{Z}'_{N_Z \times T} \boldsymbol{\varepsilon}_{T \times 1}(\boldsymbol{\theta})$$
- Again, because when the system is overidentified, there are more orthogonality conditions than there are parameters to be estimated, we can't select parameter estimates to set all the elements of $\mathbf{g}_T(\boldsymbol{\theta})$ to zero.
- Instead, we minimize a quadratic form – a weighted sum of squares and cross-products of the elements of $\mathbf{g}_T(\boldsymbol{\theta})$.

The Quadratic Form

- We can look at the linear IV problem as one of minimizing the quadratic form. Call this $Q_T(\theta)$ where

$$\begin{aligned} Q_T(\theta) &= g_T(\theta)'_{1 \times N_Z} \mathbf{W}_{T_{N_Z \times N_Z}} g_T(\theta)_{N_Z \times 1} \\ &= [T^{-1} \boldsymbol{\varepsilon}(\theta)' \mathbf{Z}] \mathbf{W}_T [T^{-1} \mathbf{Z}' \boldsymbol{\varepsilon}(\theta)] \end{aligned}$$

- \mathbf{W}_T is a symmetric, positive definite weighting matrix.
- IV regression chooses the parameter estimates to minimize $Q_T(\theta)$.

Why a Weighting Matrix?

- One could just use a $N_Z \times N_Z$ identity matrix instead and still perform the optimization.
- The reason you don't is that this approach would not minimize the variance of the estimator.
- We will perform the optimization for an arbitrary \mathbf{W}_T and then at the end, pick the one that leads to the estimator with the smallest asymptotic variance.

Solution

- Now, substitute into $Q_T(\theta)$ for ε , yielding:

$$Q_T(\theta) = [T^{-1}[\mathbf{y} - \mathbf{X}\theta]' \mathbf{Z}] \mathbf{W}_T [T^{-1} \mathbf{Z}' [\mathbf{y} - \mathbf{X}\theta]]$$

- The first-order conditions for minimizing w.r.t. θ are:

$$\mathbf{X}' \mathbf{Z} \mathbf{W}_T \mathbf{Z}' (\mathbf{y} - \mathbf{X}\hat{\theta}) = \mathbf{0}$$

- Which solve as:

$$\mathbf{X}' \mathbf{Z} \mathbf{W}_T \mathbf{Z}' \mathbf{y} = \mathbf{X}' \mathbf{Z} \mathbf{W}_T \mathbf{Z}' \mathbf{X} \hat{\theta}$$

Simplification

- Same number of regressors as instruments (exactly identified).
 - Then, $\mathbf{Z}'\mathbf{X}$ is invertible, and two of the $\mathbf{Z}'\mathbf{X}$'s cancel as does \mathbf{W}_T , leaving
$$\hat{\theta} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{y}$$
 - Here, there is no need to take particular combinations of instruments, because $N_Z = N_X$, the FOC can be satisfied exactly, i.e. \mathbf{W}_T does not appear in the solution.

Simplification cont...

- It may be clearer why we have to use the weighting matrix if we look at the problem in another way.

- If we write out the minimization problem for OLS, we are minimizing the sum of squared residuals.

- Taking the first-order condition leads to our N_X sample orthogonality conditions:

$$E[x_1' \times e] = 0$$

$$\vdots = \vdots$$

$$E[x_{N_X}' \times e] = 0$$

Note that the first x might be the constant vector, $\mathbf{1}$.

- There are N_X parameters to estimate, and N_X equations, so you don't need to weight the information in them in any special way.

Simplification cont...

- Everything is fine, and those equations were just the OLS normal equations.

- But what if we tried the same trick with the instruments, and just tried to form the analog to the OLS normal equations?

- i.e. if you tried:

$$E[x_1' \times e] = 0$$

$$\vdots = \vdots$$

$$E[x_{N_Z}' \times e] = 0$$

you'd have N_Z equations and N_X unknowns. The system would not have a solution.

- So what we do is pick a weighting matrix, \mathbf{W}_T , that minimizes the variance of the estimator.

Simplification cont...

- So, When $N_Z > N_X$, the model is overidentified and the \mathbf{W}_T stays in the solution:
 - That is, while $\mathbf{Z}'\mathbf{e}$ is $N_Z \times 1$, $\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{e}$ is $N_X \times 1$, and we can solve for the N_X parameters.
- Now the solution looks like:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{y}$$

Large Sample Properties

- Consistency:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= (\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'(\mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\theta}_0 + (\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\boldsymbol{\varepsilon}\end{aligned}$$

and $\mathbf{Z}'\boldsymbol{\varepsilon}$ is zero by assumption.

Large Sample Properties cont...

- Asymptotic Normality

$$\begin{aligned}\sqrt{T}(\hat{\theta} - \theta_0) &= \sqrt{T}(\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_T\mathbf{Z}'\boldsymbol{\varepsilon}(\theta_0) \\ &= \left(\frac{\mathbf{X}'\mathbf{Z}}{T}\mathbf{W}_T\frac{\mathbf{Z}'\mathbf{X}}{T}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Z}}{T}\mathbf{W}_T\frac{1}{\sqrt{T}}\mathbf{Z}'\boldsymbol{\varepsilon}(\theta_0)\right)\end{aligned}$$

- So what happens as $T \rightarrow \infty$? As long as:
 - $\mathbf{Z}'\mathbf{Z}/T \rightarrow \mathbf{M}_{ZZ}$, finite and full rank,
 - $\mathbf{X}'\mathbf{Z}/T \rightarrow \mathbf{M}_{XZ}$, finite and rank N_X , and
 - \mathbf{W}_T limits out to something finite and full rank, all is well.
 - Then, if $\boldsymbol{\varepsilon}(\theta)$ is serially uncorrelated and homoskedastic,

$$\frac{1}{\sqrt{T}}\mathbf{Z}'\boldsymbol{\varepsilon}(\theta_0) \rightarrow N(0, \sigma^2\mathbf{M}_{ZZ})$$

Asymptotic Normality cont...

- Then \sqrt{T} times the sample average of the orthogonality conditions is asymptotically normal.
- Note: If the $\boldsymbol{\varepsilon}$'s are serially correlated and/or heteroskedastic, asymptotic normality is still possible.

Asymptotic Normality cont...

- Define S as:
$$S = \lim_{T \rightarrow \infty} \text{Var} \left[\frac{1}{\sqrt{T}} \mathbf{Z}' \boldsymbol{\varepsilon}(\boldsymbol{\theta}) \right] = \sigma^2 \mathbf{M}_{ZZ}$$
- More generally, S is the variance of $T^{1/2}$ times the sample average of $\mathbf{f}()$, or $T^{1/2} \mathbf{g}_T$. That is,

$$S = \lim_{T \rightarrow \infty} \text{Var} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{f}_t(\boldsymbol{\theta}_0) \right] = \lim_{T \rightarrow \infty} \text{Var} \left[\sqrt{T} \mathbf{g}_T(\boldsymbol{\theta}_0) \right]$$

- where again, $\mathbf{f}_t(\boldsymbol{\theta}) = \mathbf{z}_t' \boldsymbol{\varepsilon}_t(\boldsymbol{\theta})$, which is an $N_Z \times 1$ column vector, of the orthogonality conditions in a single period evaluated at the parameter vector, $\boldsymbol{\theta}$, and

$$\mathbf{g}_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T \mathbf{f}_t(\boldsymbol{\theta})$$

- is the sample average of the orthogonality conditions.

Asymptotic Normality cont...

- With these assumptions,

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow N(\mathbf{0}, \mathbf{V})$$

- where,

$$\begin{aligned} \mathbf{V} &= (\mathbf{M}_{XZ} \mathbf{W} \mathbf{M}_{ZX})^{-1} (\mathbf{M}_{XZ} \mathbf{W} \mathbf{S} \mathbf{W} \mathbf{M}_{ZX}) (\mathbf{M}_{XZ} \mathbf{W} \mathbf{M}_{ZX})^{-1} \\ &= \sigma^2 (\mathbf{M}_{XZ} \mathbf{W} \mathbf{M}_{ZX})^{-1} (\mathbf{M}_{XZ} \mathbf{W} \mathbf{M}_{ZZ} \mathbf{W} \mathbf{M}_{ZX}) (\mathbf{M}_{XZ} \mathbf{W} \mathbf{M}_{ZX})^{-1} \end{aligned}$$

Optimal Weighting Matrix

- Let's pick the matrix, \mathbf{W}_T , that minimizes the asymptotic variance of our estimator.
 - It turns out that V is minimized by picking \mathbf{W} (the limiting value of \mathbf{W}_T) to be any scalar times \mathbf{S}^{-1} .
 - \mathbf{S} is the asymptotic covariance matrix of the sample average of the orthogonality conditions $g_T(\theta)$.
 - Using the inverse of \mathbf{S} means that to minimize variance you want to down-weight the noisy orthogonality conditions and up-weight the precise ones.
 - Here, since $\mathbf{S}^{-1} = \sigma^{-2} \mathbf{M}_{ZZ}^{-1}$, it's convenient to set our optimal weighting matrix to be $\mathbf{W}^* = \mathbf{M}_{ZZ}^{-1}$

Optimal Weighting Matrix

- Plugging in to get the associated asymptotic covariance matrix, \mathbf{V}^* , yields:

$$\begin{aligned} \mathbf{V}^* &= \sigma^2 (\mathbf{M}_{XZ} \mathbf{M}_{ZZ}^{-1} \mathbf{M}_{ZX})^{-1} \times \\ &\quad (\mathbf{M}_{XZ} \mathbf{M}_{ZZ}^{-1} \mathbf{M}_{ZZ} \mathbf{M}_{ZZ}^{-1} \mathbf{M}_{ZX}) (\mathbf{M}_{XZ} \mathbf{M}_{ZZ}^{-1} \mathbf{M}_{ZX})^{-1} \\ &= \sigma^2 (\mathbf{M}_{XZ} \mathbf{M}_{ZZ}^{-1} \mathbf{M}_{ZX})^{-1} \end{aligned}$$

- In practice, $\mathbf{W}_T^* = T^{-1}(\mathbf{Z}'\mathbf{Z})^{-1}$ and as T increases $\mathbf{W}_T^* \rightarrow \mathbf{W}^*$.
- Now, with the optimal weighting matrix, our estimator becomes:

$$\hat{\theta}_T^* = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1} \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}'\mathbf{y}$$

Optimal Weighting Matrix

- You will notice that this is the 2SLS estimator.
- Thus 2SLS is just IV estimation using an optimal weighting matrix.
- If we had used \mathbf{I}_{Nz} as our weighting matrix, the orthogonality conditions would not have been weighted optimally, and the variance of the estimator would have been too large.

- The covariance matrix with the optimal \mathbf{W} is
$$\hat{\sigma}^2 (T^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1}$$

Simplification

- This formula is also valid for just-identified IV and also for OLS, where $\mathbf{X} = \mathbf{Z}$ so that

$$\hat{\sigma}^2 (T^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X})^{-1} = \hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1}$$

Test of Overidentifying Restrictions

- Hansen (1982) has shown that T times the minimized value of the criterion function, Q_T , is asymptotically distributed as a χ^2 with $N_Z - N_X$ degrees of freedom under the null hypothesis.
- The intuition is that under the null, the instruments are uncorrelated with the residuals so that the minimized value of the objective function should be close to zero in sample.

Example – OLS

- We have

$$\mathbf{y}_{Tx1} = \mathbf{X}_{TxN_X} \boldsymbol{\theta}_{N_X \times 1} + e_{Tx1}$$

- With the usual OLS assumptions:

- $E[e] = 0$
- $E[ee'] = \sigma^2 \mathbf{I}$
- $E[\mathbf{X}'e] = 0$

- The quadratic form to be minimized with OLS is:

$$\min_{\hat{\beta}} \mathbf{e}'\mathbf{e}$$

or

$$\min_{\hat{\beta}} [\mathbf{y} - \mathbf{X}\boldsymbol{\theta}]'[\mathbf{y} - \mathbf{X}\boldsymbol{\theta}]$$

Example – OLS

- The first-order conditions to that problem are

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \mathbf{0}_{N \times 1}$$

which implies that

$$\mathbf{X}'\mathbf{e} = \mathbf{0}$$

- Now, suppose that we have a single regressor, x and a constant, 1.
- Then,

$$\mathbf{X} = [1 \quad X]$$

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}$$

Example – OLS

- First-order conditions:

1. $1 \times (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = 0$

2. $x \times (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = 0$

- These are the two orthogonality conditions which are the OLS normal equations. The solution is, of course:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

Example 2: IV Estimation

- Let's do IV estimation the way you have seen it before.
 - Recall that your \mathbf{X} matrix is correlated with the disturbances.
 - To get around this problem, you regress \mathbf{X} on \mathbf{Z} , and form

$$\hat{\mathbf{X}}_{T \times N_X} = \mathbf{Z}_{T \times N_Z} (\mathbf{Z}' \mathbf{Z})_{N_Z \times N_Z}^{-1} (\mathbf{Z}' \mathbf{X})_{N_Z \times N_X}$$

- Then
$$\hat{\boldsymbol{\theta}}_{IV} = (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}' \mathbf{y}$$
$$= [\mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{X}]^{-1} \mathbf{X}' \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}$$
- This is exactly what we got before when we did IV estimation with an optimal weighting matrix.

Comments on This Estimator

- To form $\hat{\mathbf{X}}$, what one does in practice is take each regressor, x_i , and regress it on all of the \mathbf{Z} variables to form \hat{x}_i .
 - This is important because it may be that only some of the x 's are correlated with the disturbances. Then, if x_j were uncorrelated with ε , one can simply use it as its own instrument.
 - Notice that by regressing \mathbf{X} on \mathbf{Z} , we are collapsing down from N_Z instruments to N_X regressors.
 - Put another way, we are picking particular combinations of the instruments to form $\hat{\mathbf{X}}$
 - This procedure is optimal in the sense that it produces the smallest asymptotic covariance matrix for the estimators.
 - Essentially, by performing this regression, we are optimally weighting the orthogonality conditions to minimize the asymptotic covariance matrix of the estimator.

Generalizations

- Next we generalize the model to non-spherical distributions by adding in
 - Heteroskedasticity
 - Serial correlation
- This will be important for robust estimation of covariance matrices, something that is usually done in asset pricing in finance. The heteroskedasticity-consistent estimator is the White (1980) estimator, and the estimator that is robust to serial correlation as well is due to Newey and West (1987).

Heteroskedasticity and Serial Correlation

- Start with the linear model where

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

$$E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$$

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \boldsymbol{\Omega}$$

where $\boldsymbol{\Omega}_{T \times T}$ is positive definite.

Heteroskedasticity and Serial Correlation

- Heteroskedastic disturbances have different variances but are uncorrelated across time.

$$\sigma^2 \mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_T^2 \end{bmatrix}$$

- Serially correlated disturbances are often found in time series where the observations are not independent across time. The off-diagonal terms in $\sigma^2 \mathbf{\Omega}$ are not zero – they depend on the model used.
 - If memory fades over time, the values decline as you move away from the diagonal.
 - A special case is the moving average, where the value equals zero after a finite number of periods.

Example: OLS

- With OLS

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

$$E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$$

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}] = \sigma^2 \mathbf{I}$$

- The OLS estimator is just

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= \boldsymbol{\theta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \end{aligned}$$

Example: OLS cont...

- The sampling (or asymptotic) variance of the estimator is:

$$\begin{aligned} \text{Var}[\hat{\boldsymbol{\theta}} | \mathbf{X}] &= E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' | \mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\sigma^2 \boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\sigma^2}{T} \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{T} \mathbf{X}'\boldsymbol{\Omega}\mathbf{X} \right) \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1} \end{aligned}$$

- This is not the same as OLS. We're using OLS here when some kind of GLS would be appropriate.

Consistency and Asymptotic Normality

- Consistency follows as long as the variance of $\hat{\boldsymbol{\theta}} \rightarrow 0$. This means that $(1/T(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}))$ can't blow up.
- Asymptotic normality follows if

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{T} \right)^{-1} \frac{1}{\sqrt{T}} \mathbf{X}'\boldsymbol{\varepsilon} \text{ is normal.}$$

- We have that

$$\lim_{T \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{T} \right) = M_{XX}$$

Consistency and Asymptotic Normality

- This means that the limiting distribution of $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is the same as that of

$$\begin{aligned} & \mathbf{M}_{XX}^{-1} \frac{1}{\sqrt{T}} \mathbf{X}' \boldsymbol{\varepsilon} \\ &= \mathbf{M}_{XX}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \end{aligned}$$

- If the disturbances are just heteroskedastic, then

$$\text{Var} \left[\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{x}_t \varepsilon_t \right] = \frac{1}{T} \sum_{t=1}^T \sigma^2 \omega' \mathbf{x}_t \mathbf{x}_t'$$

Consistency and Asymptotic Normality

- As long as the diagonal elements of $\boldsymbol{\Omega}$ are well behaved, the Lindberg-Feller CLT applies so that the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is

$$\frac{\sigma^2}{T} \mathbf{M}_{XX}^{-1} \text{plim} \left(\frac{1}{T} \mathbf{X}' \boldsymbol{\Omega} \mathbf{X} \right) \mathbf{M}_{XX}^{-1}$$

and asymptotic normality of the estimator holds.

- Things are harder with serial correlation, but there are conditions given by both Amemya (1985) and Anderson (1971) that are sufficient for asymptotic normality and are thought to cover most situations found in practice.

Example: IV Estimation

- We have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{IV} &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\ &= \boldsymbol{\theta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}\end{aligned}$$

- Consistency and asymptotic normality follow, with (asymptotically):

$$\hat{\boldsymbol{\theta}} \sim N[\boldsymbol{\theta}, \mathbf{V}]$$

where

$$\begin{aligned}\mathbf{V}_{IV} &= \frac{\sigma^2}{T} ((\mathbf{M}_{XZ}\mathbf{M}_{ZZ}^{-1}\mathbf{M}_{XZ})^{-1}\mathbf{M}_{XZ}\mathbf{M}_{ZZ}^{-1}) \text{plim} \left(\frac{1}{T} \mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z} \right) \times \\ &\quad ((\mathbf{M}_{XZ}\mathbf{M}_{ZZ}^{-1}\mathbf{M}_{XZ})^{-1}\mathbf{M}_{XZ}\mathbf{M}_{ZZ}^{-1})'\end{aligned}$$

Why Do We Care?

- We wouldn't care if we knew a lot about $\boldsymbol{\Omega}$.
 - If we actually knew $\boldsymbol{\Omega}$, or at least the form of the covariance matrix, we could run GLS.
- In this case, we're desperate.
 - We don't know much about $\boldsymbol{\Omega}$ but we want to do statistical tests.
 - What if we just wanted to use IV estimation and we hadn't the foggiest notion what amount of heteroskedasticity and serial correlation there was.
 - However, we suspected that there was some of one or both.
 - This is when robust estimation of asymptotic covariance matrices comes in handy. This is exactly what is done with GMM estimation.

Example: OLS

- Let's do this with OLS to illustrate.
- The results generalize, and everywhere we use the asymptotic covariance matrix we derived for OLS under serial correlation and heteroskedasticity, just replace it with \mathbf{V}_{IV} derived immediately above.
- Recall that if $\sigma^2\mathbf{\Omega}$ were known, \mathbf{V}_{OLS} , the estimator of the asymptotic covariance matrix of the parameter estimates with heteroskedasticity and serial correlation is given by:

$$\text{Var}[\hat{\boldsymbol{\theta}} | \mathbf{X}] = \mathbf{V}_{OLS} = \frac{1}{T} \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{T} \mathbf{X}'[\sigma^2\mathbf{\Omega}]\mathbf{X} \right) \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1}$$

Example: OLS cont...

- However, $\sigma^2\mathbf{\Omega}$ must be estimated here.
- Further, we can't estimate σ^2 and $\mathbf{\Omega}$ separately.
 - $\mathbf{\Omega}$ is unknown, and can be scaled by anything.
 - Greene scales by assuming that the trace of $\mathbf{\Omega}$ equals T , which is the case in the classical model when $\mathbf{\Omega} = \mathbf{I}$.
 - So, let $\boldsymbol{\Sigma} \equiv \sigma^2\mathbf{\Omega}$.

A Problem

- So, we need to estimate

$$\left(\frac{1}{T} \mathbf{X}' \Sigma \mathbf{X} \right)$$

- To do this, it looks like we need to estimate Σ , which has $T(T+1)/2$ (since Σ is a symmetric matrix) parameters.
- With only T observations, we'd be stuck, except that what we really need to estimate is the $N_X(N_X+1)/2$ elements in the matrix:

$$\text{plim } \mathbf{M}_* = \text{plim } \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^T \sigma_{ij} x_i x_j'$$

A Problem cont...

- The point is that \mathbf{M}_* is a much smaller matrix that involves sums of squares and cross-products that involve σ_{ij} and the rows of \mathbf{X} .
- The least-squares estimator of θ is consistent, which implies that the least squares residuals e_i are pointwise consistent estimators of the population disturbances.
- So we ought to be able to use \mathbf{X} and e to estimate \mathbf{M}_* .

Heteroskedasticity

- With heteroskedasticity alone, $\sigma_{ij} = 0$ for $i \neq j$. That is, there is no serial correlation.
- We therefore want to estimate

$$\mathbf{M}_* = \frac{1}{T} \sum_{i=1}^T \sigma_i^2 x_i x_i'$$

- White has shown that under very general conditions, the estimator

$$\mathbf{S}_0 = \frac{1}{T} \sum_{i=1}^T e_i^2 x_i x_i'$$

has

$$\text{plim } \mathbf{S}_0 = \text{plim } \mathbf{M}_*$$

Heteroskedasticity

- The end result is the White (1980) heteroskedasticity consistent estimator:

$$\text{Asy. Var}[\hat{\theta}] =$$

$$\begin{aligned} V_{OLS} &= \frac{1}{T} \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{T} \sum_{i=1}^T e_i^2 x_i x_i' \right) \left(\frac{1}{T} \mathbf{X}'\mathbf{X} \right)^{-1} \\ &= T(\mathbf{X}'\mathbf{X})^{-1} \mathbf{S}_0 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- This is an extremely important and useful result.
 - It implies that without actually specifying the form of the heteroskedasticity, we can make appropriate inferences using least squares. Further, the results generalize to linear and nonlinear IV estimation.

Extending to Serial Correlation

- The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^T \sigma_{ij} \mathbf{x}_i \mathbf{x}_j'$$

would be

$$\hat{\mathbf{Q}}_* = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^T e_i e_j \mathbf{x}_i \mathbf{x}_j'$$

- But there are two problems.

Extending to Serial Correlation

1. The matrix in the above equation is $1/T$ times a sum of T^2 terms (the $e_i e_j$ terms are not zero for $i \neq j$ as in the heteroskedasticity case), which makes it hard to conclude that it converges to anything at all.
 - What we need so that we can count on convergence is that as i and j get far apart, the $e_i e_j$ terms get smaller, reaching zero in the limit.
 - This happens in a time series setting. So...
 - Put another way, we need the rows of \mathbf{X} to be well behaved in the sense that correlations between the errors diminish with increasing temporal separation.

Extending to Serial Correlation

2. Practically speaking, $\hat{\mathbf{Q}}_*$ need not be positive definite (and covariance matrices have to be).

- Newey and West have devised an autocorrelation consistent covariance estimator that overcomes this:

$$\hat{\mathbf{Q}}_* = \mathbf{S}_0 + \frac{1}{T} \sum_{l=1}^L \sum_{t=l+1}^T w_l e_t e_{t-l} (\mathbf{x}_t \mathbf{x}_{t-l}' + \mathbf{x}_{t-l} \mathbf{x}_t')$$

$$w_l = \frac{l}{L+1}$$

- The weights are such that the closer are the residuals in time the higher the weight. It is also true that you limit the “span” of the dependence.
 - What is L? There is little theoretical guidance.

Asymptotics

- We have estimators that are asymptotically normally distributed.
- We have a robust estimator of the asymptotic covariance matrix.
- We have not specified distributions for the disturbances.
- Hence, using the F statistic is not a good idea.
- The best thing to do is to use the Wald statistic with asymptotic t ratios for statistical inference.

GMM

- The discussion here follows closely that in Greene.
- We proceed as follows:
 - Review method of moments estimation.
 - Generalize method of moments estimation to overidentified systems (nonlinear analogs to the systems we just considered).
 - Relate back to linear systems.

Method of Moments Estimators

- Suppose the model for the random variable y_i implies certain expectations. For example:

$$E[y_i - \mu] = 0$$

- The sample counterpart is

$$\frac{1}{T} \sum_{i=1}^T (y_i - \mu)$$

- The estimator is the value of $\hat{\mu}$ that satisfies the sample moment conditions.
- This example is trivial.

An Apparently Different Case: OLS

- Among the OLS assumptions is:

$$E[\mathbf{x}_j e_i] = \mathbf{0}$$

- The sample analog is:

$$\frac{1}{T} \sum_{i=1}^T \mathbf{x}_i e_i = \frac{1}{T} \sum_{i=1}^T \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) = \frac{1}{T} \mathbf{X}' \mathbf{e} = \mathbf{0}$$

- The estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, satisfies these moment conditions.
- These moment conditions are just the normal equations for the least squares estimator.

Linear IV Estimation

- For linear IV estimation:

$$E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$$

- We resolved the problem of having more moments than parameters by solving:

$$\left(\frac{1}{T} \mathbf{X}' \mathbf{Z} \right) \left(\frac{1}{T} \mathbf{Z}' \mathbf{Z} \right)^{-1} \left(\frac{1}{T} \mathbf{Z}' \boldsymbol{\varepsilon} \right) = \left(\frac{1}{T} \hat{\mathbf{X}}' \boldsymbol{\varepsilon} \right) = \mathbf{0}$$

ML Estimators

- All of the maximum likelihood estimators we looked at for testing the CAPM involve equating the derivatives of the log-likelihood function with respect to the parameters to zero. For example, if:

$$\ln(L) = \sum_{i=1}^T \ln(f(y_i, \mathbf{x}_i | \boldsymbol{\theta})),$$

- then

$$E\left[\frac{\partial \ln(f(y_i, \mathbf{x}_i | \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

- and the MLE is found by equating the sample analog to zero:

$$\frac{1}{T} \sum_{i=1}^T \frac{\partial \ln(f(y_i, \mathbf{x}_i | \hat{\boldsymbol{\theta}}))}{\partial \hat{\boldsymbol{\theta}}} = \mathbf{0}$$

The Point

- The point is that everything we have considered is a method of moments estimator.

GMM

- The preceding examples (except for the linear IV estimation) have a common aspect.
- They were all exactly identified.
- But where there are more moment restrictions than parameters, the system is *overidentified*.
- That was the case with linear IV estimators, and we needed a weighting matrix so that we could solve the system.
- That's what we have to do for the general case as well.

Intuition for Weighting

- What we want to do is minimize a criterion function such as the sum of squared residuals by choosing parameters.
- Then, we'll only have as many first-order conditions as parameters, and we'll be able to solve the system.
- That's what the optimal weighting matrix did for us in linear IV estimation.
- If there are N_Z instruments and N_X parameters, the matrix took the N_Z orthogonality conditions and weighted them appropriately so that there were only N_X equations that were set to zero.
- These N_X equations are the first-order conditions of the criterion function with respect to the parameters.

Intuition for Weighting

- Hansen (1982) showed that we can use as a criterion function a weighted sum of squared orthogonality conditions.
 - What does this mean?
 - Suppose we have $\bar{m}(\theta) = \mathbf{0}$ as a set of l (possibly non-linear) orthogonality conditions in the population.
 - Then a criterion function q looks like: $q = \bar{m}(\theta)' \mathbf{B} \bar{m}(\theta)$ where \mathbf{B} is any positive definite matrix that is not a function of θ , such as the identity matrix.
 - Any such \mathbf{B} will produce a consistent estimator of θ .
 - Choosing an optimal \mathbf{B} is essentially choosing an optimal weighting matrix.

Testing for a Given Distribution

- Suppose we want to test whether a set of observations x_t , ($t = 1, \dots, T$) come from a given distribution $y = F(X, \theta)$.
- Under the null, the moments should coincide.
- This means $E[x_t^r - y^r] = 0 \quad \forall r = 1, \dots, R$
- Assume the x_t are i.i.d. (we can get by with less). Then, sample moments converge to population moments:

- Under the null
$$\frac{1}{T} \sum_{t=1}^T x_t^r \rightarrow E[x_t^r] \quad \forall r \in R$$

$$\frac{1}{T} \sum_{t=1}^T x_t^r - E[y^r] \rightarrow 0 \quad \forall r \in R \quad **$$

Testing for a Given Distribution cont...

- Define $f(x_t, \theta)$ as an R vector with elements $x_t^r - E[y^r]$ and let

$$g_T(\theta) = \frac{1}{T} \sum_{t=1}^T f(x_t, \theta).$$

Hence, $g_T(\theta)$ has elements given by the equation ** above.

- The idea is to find parameters θ so that the vector

$$\{E[y], E[y^2], \dots, E[y^R]\}$$

satisfies the condition **.

- If the number of parameters is less than R , the system is overidentified and we must choose θ_T to set

$$A_T g_T(\theta) = \mathbf{0}$$

Applying Hansen's Results

- The optimal choice of the $l \times R$ matrix A_0 is

$$\mathbf{D}_0' \mathbf{S}_0^{-1}$$

where

$$\mathbf{D}_0 = E \left[\frac{\partial f(x_t, \theta)}{\partial \theta'} \right]$$

and

$$\mathbf{S}_0 = E[f(x_t, \theta) f(x_t, \theta)']$$

- Then, we can use Hansen's test of overidentifying restrictions

$$J_T = T g_T(\theta_T)' \mathbf{S}_0^{-1} g_T(\theta_T)$$

which is distributed χ^2_{r-l} under the null, to test the distributional assumption.

The Normal Distribution

- Let

$$x_t \sim N(\mu, \sigma^2)$$

so that

$$x_t - \mu \sim N(0, \sigma^2)$$

- Using the moment generating function for a normal distribution, the moments of $x_t - \mu$ are given by:

$$E \left[\begin{array}{l} (x_t - \mu)^{2n-1} \\ (x_t - \mu)^{2n} - \frac{\sigma^{2n}(2n)!}{2^n n!} \end{array} \right] = 0$$

for all integers greater than zero.

The Normal Distribution cont...

- Defining sample moments yields

$$g_T(\mu, \sigma^2) = \frac{1}{T} \sum_{t=1}^T \left[\begin{array}{l} (x_t - \mu)^{2n-1} \\ (x_t - \mu)^{2n} - \frac{\sigma^{2n}(2n)!}{2^n n!} \end{array} \right]$$

for all integers greater than zero.

- Now we can test the normal model. We want to choose θ such that $\mathbf{D}_0' \mathbf{S}_0^{-1} g_T(\mu, \sigma^2) = 0$
- WLOG, test for normality with $n=2$. Then,

$$g_T(\mu, \sigma^2) = \frac{1}{T} \sum_{t=1}^T \left[\begin{array}{l} (x_t - \mu) \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{array} \right]$$

The Normal Distribution cont...

- Now, we need the covariance matrix of the moment conditions, S_0 and the derivative matrix D_0 . So first:

$$\mathbf{S}_0 = E[f f']$$

which is a 4x4 matrix.

- What do the f 's look like?

$$\begin{bmatrix} (x_t - \mu) \\ (x_t - \mu)^2 - \sigma^2 \\ (x_t - \mu)^3 \\ (x_t - \mu)^4 - 3\sigma^4 \end{bmatrix}$$

- So the 1,1 element of S_0 is $E[(x_t - \mu)(x_t - \mu)] = \sigma^2$

The Normal Distribution cont...

- The 1,2 element is

$$E[(x_t - \mu)((x_t - \mu)^2 - \sigma^2)] = 0$$

and so on.

- Therefore

$$\mathbf{S}_0 = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}$$

The Normal Distribution cont...

- Now, $\mathbf{D}_0 = \partial \mathbf{g} / \partial \theta =$

$$\mathbf{D}_0 = \begin{bmatrix} \frac{\partial g_1}{\partial \mu} = -1 & \frac{\partial g_1}{\partial \sigma^2} = 0 \\ \frac{\partial g_2}{\partial \mu} = -2(0) = 0 & \frac{\partial g_2}{\partial \sigma^2} = -1 \\ \frac{\partial g_3}{\partial \mu} = -3\sigma^2 & \frac{\partial g_3}{\partial \sigma^2} = 0 \\ \frac{\partial g_4}{\partial \mu} = 0 & \frac{\partial g_4}{\partial \sigma^2} = -6\sigma^2 \end{bmatrix}$$

so that

$$\mathbf{D}_0' = \begin{bmatrix} -1 & 0 & -3\sigma^2 & 0 \\ 0 & -1 & 0 & -6\sigma^2 \end{bmatrix}$$

The Normal Distribution cont...

- Now, in sample, we really have \mathbf{D}_T and \mathbf{S}_T . So what we do is plug in sample moments for the population moments:

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T x_t, \quad \hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \hat{\mu})^2$$

- The corresponding asymptotic covariance matrix for the estimators is

$$(\mathbf{D}_0' \mathbf{S}_0^{-1} \mathbf{D}_0)^{-1},$$

which equals

$$\begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

The Normal Distribution cont...

- The covariance matrix for the estimates is given by

$$(\mathbf{D}_T' \mathbf{S}_T^{-1} \mathbf{D}_T)^{-1},$$

- Which equals

$$\begin{bmatrix} \hat{\sigma}^2 & 0 \\ 0 & 2\hat{\sigma}^4 \end{bmatrix}$$

- The GMM estimates are the MLE's. Note that the optimal weights, $\mathbf{D}_0' \mathbf{S}_0^{-1}$, pick out only the first two moment conditions.

$$\mathbf{D}_0' \mathbf{S}_0^{-1} = \mathbf{A}_0 = \begin{bmatrix} -\frac{1}{\sigma^2} & 0 & 0 & 0 \\ 0 & -\frac{1}{2\sigma^4} & 0 & 0 \end{bmatrix}$$

The Normal Distribution cont...

- Why is this? Recall GMM picks the linear combinations of moments that minimizes the covariance matrix of the estimators.
- In the normal case, the MLE's achieve the Cramer-Rao lower bound. Thus GMM is going to find the MLE's.
- What about the test of overidentifying restrictions?
 - Because the first two moment conditions are set identically to zero, J_T tests whether the higher order moment conditions are statistically equal to zero.

Tests of the CAPM using GMM

- Robust tests of the CAPM can be performed using GMM.
- With GMM, we can have conditional heteroskedasticity and serial dependence of returns.
- Need only that returns (not errors) are stationary and ergodic with finite fourth moments.

How to Proceed

- First, set up the moment conditions.
- We know that we need to set things up so that “errors” have zero expectations.
 - Start with
$$\mathbf{Z}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}Z_{mt} + \boldsymbol{\varepsilon}_t$$
where \mathbf{Z}_t is an N-vector of asset excess returns at time t.
 - Then, $\boldsymbol{\varepsilon}_t$ equals
$$\boldsymbol{\varepsilon}_t = \mathbf{Z}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}Z_{mt}$$
 - We know also that $\boldsymbol{\varepsilon}_t$ and Z_{mt} are orthogonal.

CAPM cont...

- This gives us two sets of N orthogonality conditions:
 - $E[\epsilon_t] = 0$
 - $E[Z_{mt} \epsilon_t] = 0$

- Now, let $\mathbf{h}_t' = [1 \ Z_{mt}]$.

- Further, let $\boldsymbol{\theta}' = [\alpha' \ \beta']$.

- Then, using the GMM notation

$$f_t(\boldsymbol{\theta}) = \mathbf{h}_t \otimes \boldsymbol{\epsilon}_t$$

- Where \otimes is the Kronecker product.

- Now, we are in the standard GMM setup. The sample average of f_t is

$$g_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^T f_t(\boldsymbol{\theta})$$

CAPM cont...

- The GMM estimator minimizes the quadratic form,

$$Q_T(\boldsymbol{\theta}) = g_T(\boldsymbol{\theta})' \mathbf{W} g_T(\boldsymbol{\theta})$$

where \mathbf{W} is the $2N \times 2N$ weighting matrix.

- The system is exactly identified, so that \mathbf{W} drops out and we are left with the ML (and OLS) estimators from before.
- So what's new?

What's New

- What's new is not the estimator, it's the variance-covariance matrix of the estimator.
- This is basically GMM on a linear system where the instruments are the regressors, 1 and Z_{mt} , we already showed our GMM estimator reduces to OLS in that case.
- What about the covariance matrix?
- What's important is that it's robust. We have already shown that the V-C matrix for $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is, with an optimal weighting matrix, (ours was optimal)

$$\mathbf{V} = [\mathbf{D}_0' \mathbf{S}_0^{-1} \mathbf{D}_0]^{-1}$$

What's New cont...

- where

$$\mathbf{D}_0 = E \left[\frac{\partial f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right]$$

- and

$$\mathbf{S}_0 = E[f(x_t, \boldsymbol{\theta})f(x_t, \boldsymbol{\theta})']$$

- Recall the need to use the finite sample analogs.

Asymptotic Distribution of $\hat{\theta}$.

- It's given by:

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{T}[\mathbf{D}_0' \mathbf{S}_0^{-1} \mathbf{D}_0]^{-1}\right)$$

- We know that

$$\mathbf{D}_0 = \begin{bmatrix} 1 & \mu_m \\ \mu_m & (\sigma_m^2 + \mu_m^2) \end{bmatrix} \otimes \mathbf{I}_N$$

- A consistent estimator \mathbf{D}_T can be constructed using MLE's of μ_m and σ_m^2 .
- For \mathbf{S}_0 , it's not so obvious. You need to reduce the summation to a finite number of terms. The appendix provides a number of assumptions.

- These assumptions essentially mean that one ignores the persistence past a certain number of lags.
 - Newey-West had it at L lags.
- Once you have an \mathbf{S}_T , then one can construct a χ^2 test of the N restrictions obtained by setting $\alpha = 0$. That is:

$$\hat{\alpha} = \mathbf{R}\hat{\theta}$$

where

$$\mathbf{R} = [1 \quad 0] \otimes \mathbf{I}_N$$

- Then,

$$Var[\hat{\boldsymbol{\alpha}}] = \frac{1}{T} \mathbf{R}[\mathbf{D}_T' \mathbf{S}_T^{-1} \mathbf{D}_T]^{-1} \mathbf{R}'$$

and

$$J_\gamma = T \hat{\boldsymbol{\alpha}}' [\mathbf{R}[\mathbf{D}_T' \mathbf{S}_T^{-1} \mathbf{D}_T]^{-1} \mathbf{R}']^{-1} \hat{\boldsymbol{\alpha}}$$

which under the null is distributed $\chi^2(N)$.