1 Delta method

See also exercise "Delta Method".

Let $Y_1, Y_2, \ldots$ be a sequence of random variables such that $E(Y_i) = \mu$, $Var(Y_i) = \sigma^2 < \infty$ and

$$\sqrt{n}\,(Y_i - \mu) \to U \sim N(0, \sigma^2).$$

Of course, in this context Y_i is usually just the mean $\bar{X}_i$.

Define a new sequence of random variables

$$Z_i = f(Y_i)$$

where f is differentiable and measurable (at least at μ). What does the asymptotic distribution of Z_i look like?

The first order Taylor expansion of f around μ is

$$f\left(y\right) = f\left(\mu\right) + f'\left(\mu\right)\left(y - \mu\right) + \text{rest}$$

where the rest is negligible in the limit (and dropped from the notation). Thus

$$Z_i = f\left(\mu\right) + f'\left(\mu\right)\left(Y_i - \mu\right).$$

We see that Z_i is just a linear transformation of Y_i. Since Y_i is asymptotically normal, so is Z_i. The asymptotic mean of Z_i is

$$E\left(Z_i\right) = E\left(f\left(\mu\right) + f'\left(\mu\right)\left(Y_i - \mu\right)\right) = f\left(\mu\right)$$

and the asymptotic variance is

$$\begin{aligned} Var\left(Z_i\right) &= Var\left(f\left(\mu\right) + f'\left(\mu\right)\left(Y_i - \mu\right)\right) \\ &= \left[f'\left(\mu\right)\right]^2 Var\left(Y_i\right) \\ &= \left[f'\left(\mu\right)\right]^2 \sigma^2. \end{aligned}$$

2 Expected gradient vector

First we show that for all elements of G

$$E_\theta\left(G_{ij}\left(\theta, X_i\right)\right) = 0.$$

Consider the identity

$$\int e^{\ln f_X(x_i;\theta)} dx_i = 1$$

and differentiate both sides with respect to the j-th component of θ,

$$\begin{aligned} \int e^{\ln f_X(x_i;\theta)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i &= 0 \\ E_\theta\left(\frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j}\right) &= 0 \\ E\left(G_{ij}\left(\theta, X_i\right)\right) &= 0. \end{aligned}$$

Summing over $i = 1, \ldots, n$ yields

$$E\left(g_j\left(\theta\right)\right) = 0$$

and hence

$$E\left(g\left(\theta\right)\right) = 0.$$

3 Information matrix equality

Consider the expected gradient component equation $E\left(G_{ij}\left(\theta, X_i\right)\right) = 0$ and differentiate both sides once more with respect to θ. The left hand side is

$$E\left(G_{ij}\left(\theta, X_i\right)\right) = \int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i$$

and the derivative with respect to θ_k is

$$\begin{aligned}
\frac{\partial}{\partial \theta_k} \int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i &= \int \frac{\partial}{\partial \theta_k} e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i \\
&= \int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_k} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i \\
&\quad + \int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j \partial \theta_k} dx_i.
\end{aligned}$$

Thus

$$\int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_k} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j} dx_i = -\int e^{\ln f_X\left(x_i;\theta\right)} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j \partial \theta_k} dx_i$$

or

$$E_\theta\left(\frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_k} \frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j}\right) = -E_\theta\left(\frac{\partial \ln f_X\left(x_i;\theta\right)}{\partial \theta_j \partial \theta_k}\right).$$

Summing over all $i = 1, \ldots, n$ and arranging the terms as a matrix yields

$$Cov\left(g\left(\theta\right)\right) = -E\left(H\left(\theta\right)\right).$$

4 Consistency of ML estimator

See Davidson and MacKinnon (2004, section 10.3).

Finite sample identification condition: The log-likelihood must be different for different parameter values,

$$\ln L\left(\theta, x\right) \neq \ln L\left(\theta', x\right)$$

for $\theta \neq \theta'$ and all x. Let θ_0 denote the true parameter values. Then by Jensen's inequality

$$E_0 \ln\left(\frac{L\left(\theta'\right)}{L\left(\theta_0\right)}\right) < \ln E_0\left(\frac{L\left(\theta'\right)}{L\left(\theta_0\right)}\right) \tag{1}$$

since the logarithm is a concave transformation. The expectation on the right hand side is

$$\begin{aligned}
E_0\left(\frac{L(\theta')}{L(\theta_0)}\right) &= \int \frac{L(\theta')}{L(\theta_0)} L(\theta_0)\, dx \\
&= \int L(\theta')\, dx \\
&= 1
\end{aligned}$$

since $L(\theta')$ is a density function (the joint density of x at θ'). Hence, the right hand side of (1) vanishes,

$$E_0 \ln\left(\frac{L(\theta')}{L(\theta_0)}\right) < 0$$

or

$$E_0 \ln L(\theta') < E_0 \ln L(\theta_0)\,. \tag{2}$$

By a law of large numbers

$$plim\left(\frac{1}{n}\ln L(\theta)\right) = \lim\left(\frac{1}{n}E_0 \ln L(\theta)\right)$$

for all θ. Therefore, from (2) we conclude

$$plim\left(\frac{1}{n}\ln L(\theta')\right) \leq plim\left(\frac{1}{n}\ln L(\theta_0)\right) \tag{3}$$

for all θ' where the inequality is not strict because of the limits.

Since the ML estimator $\hat{\theta}$ maximizes $\ln L(\theta)$ it must be true that

$$plim\left(\frac{1}{n}\ln L\left(\hat{\theta}\right)\right) \geq plim\left(\frac{1}{n}\ln L(\theta_0)\right). \tag{4}$$

The only way that (3) and (4) can both be true is if

$$plim\left(\frac{1}{n}\ln L\left(\hat{\theta}\right)\right) = plim\left(\frac{1}{n}\ln L(\theta_0)\right).$$

Since the asymptotic identification condition requires $plim\left(\frac{1}{n}\ln L(\theta')\right) \neq plim\left(\frac{1}{n}\ln L(\theta_0)\right)$ for all $\theta' \neq \theta_0$, consistency follows.

5 Asymptotic normality of ML estimator

Ignoring the residual term, the Taylor expansion is

$$g(\hat{\theta}) = g(\theta_0) + H(\theta_0)(\hat{\theta} - \theta_0) = 0.$$

Hence,

$$\begin{aligned}
(\hat{\theta} - \theta_0) &= -\left[H(\theta_0)\right]^{-1} g(\theta_0) \\
\sqrt{n}(\hat{\theta} - \theta_0) &= -\left[H(\theta_0)\right]^{-1} \sqrt{n} g(\theta_0) \\
\sqrt{n}(\hat{\theta} - \theta_0) &= -\left[\frac{1}{n}H(\theta_0)\right]^{-1} \sqrt{n}\bar{g}(\theta_0)
\end{aligned}$$

where $\bar{g}(\theta_0) = n^{-1} g(\theta_0)$. By the central limit theorem, $\sqrt{n}\bar{g}(\theta_0)$ is asymptotically normally distributed with

$$\begin{aligned} E\left(\sqrt{n}\bar{g}(\theta_0)\right) &= 0 \\ Cov\left(\sqrt{n}\bar{g}(\theta_0)\right) &= -E\left(\frac{1}{n}H(\theta_0)\right). \end{aligned}$$

Since $\sqrt{n}(\hat{\theta} - \theta_0)$ is just a linear transformation of $\sqrt{n}\bar{g}(\theta_0)$ it is asymptotically normal as well with

$$\begin{aligned} E\left(\sqrt{n}(\hat{\theta} - \theta_0)\right) &= 0 \\ Cov\left(\sqrt{n}(\hat{\theta} - \theta_0)\right) &= -\left[\frac{1}{n}H(\theta_0)\right]^{-1} E\left(\frac{1}{n}H(\theta_0)\right)\left[\frac{1}{n}H(\theta_0)\right]^{-1}. \end{aligned}$$

In the limit $plim\frac{1}{n}H(\theta_0) = E\left(\frac{1}{n}H(\theta_0)\right)$ and thus the covariance matrix simplifies to

$$Cov\left(\sqrt{n}(\hat{\theta} - \theta_0)\right) = -\left[E\left(\frac{1}{n}H(\theta_0)\right)\right]^{-1},$$

and hence approximately

$$\hat{\theta} \sim N\left(\theta_0, -\left[E\left(H(\theta_0)\right)\right]^{-1}\right).$$

6 Errors in variables

The model with the unobservable exogenous variable is

$$y_t = \alpha + \beta x_t^* + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma_\varepsilon^2).$$

We can only observe

$$x_t = x_t^* + v_t.$$

Estimate the model

$$\begin{aligned} y_t &= \alpha + \beta(x_t - v_t) + \varepsilon_t \\ &= \alpha + \beta x_t + \underbrace{\varepsilon_t - \beta v_t}_{=u_t}. \end{aligned}$$

The error u_t is correlated with the exogenous variable x_t and

$$E(u_t|x_t) = E(u_t|v_t) = -\beta v_t,$$

hence

$$\begin{aligned} Cov(x_t, u_t) &= E(x_t u_t) \\ &= E(x_t E(u_t|x_t)) \\ &= E((x_t^* + v_t)(-\beta v_t)) \\ &= -\beta\sigma_v^2. \end{aligned}$$

Generalization:

$$y = X^*\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

where a single column of the $T \times K$ matrix X^* is latent (say, the last column). We can observe

$$X = X^* + V$$

where the $T \times K$ matrix of measurement errors V is zero everywhere except in the last column. Estimate the model

$$\begin{aligned} y &= (X - V)\,\beta + \varepsilon \\ &= X\beta + \underbrace{(\varepsilon - V\beta)}_{=u}. \end{aligned}$$

Then $E(u|X) = E(u|V) = -V\beta$. Note that *all* parameters are estimated inconsistently – not just the last one. See `errorsinvars.R`.

7 Consistency of simple IV estimator

The probability limit of

$$\begin{aligned} \hat{\beta} &= \beta + (W'X)^{-1} W'u \\ &= \beta + \left(\frac{1}{n} W'X\right)^{-1} \frac{1}{n} W'u \end{aligned}$$

is

$$\begin{aligned} plim\hat{\beta}_{IV} &= \beta + plim\left(\left(\frac{1}{n} W'X\right)^{-1} \frac{1}{n} W'u\right) \\ &= \beta + plim\left(\frac{1}{n} W'X\right)^{-1} plim\frac{1}{n} W'u \\ &= \beta + \left(plim\frac{1}{n} W'X\right)^{-1} plim\frac{1}{n} W'u. \end{aligned}$$

If $plim\frac{1}{n}W'X = S_{WX}$ is deterministic and nonsingular, then

$$plim\hat{\beta}_{IV} = \beta + S_{WX}^{-1} \cdot plim\frac{1}{n} W'u.$$

Obviously, $\hat{\beta}_{IV}$ is consistent if $plim\frac{1}{n}W'u = 0$. Write

$$\frac{1}{n} W'u = \frac{1}{n} \sum_{t=1}^{n} W_t' u_t$$

where W_t is the t-th row of W (as a row vector). By the law of large numbers

$$\frac{1}{n} \sum_{t=1}^{n} W_t' u_t \xrightarrow{p} E(W'u) = 0$$

or $plim\frac{1}{n}W'u = 0$. This establishes the consistency of $\hat{\beta}_{IV}$.

8 Asymptotic normality of simple IV estimator

Multiply

$$\hat{\beta}_{IV} - \beta = \left(\frac{1}{n}W'X\right)^{-1} \frac{1}{n}W'u$$

by $\sqrt{n}$ to get

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) = \left(\frac{1}{n}W'X\right)^{-1} \sqrt{n}\left(\frac{1}{n}W'u\right). \tag{5}$$

The right hand side of (5) has two terms. We already know that the first one, $\left(\frac{1}{n}W'X\right)^{-1}$ converges in probability to S_{WX}^{-1}. The second term is

$$\sqrt{n}\left(\frac{1}{n}W'u\right) = \sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} W_t'u_t\right).$$

We apply the central limit theorem to the mean $\frac{1}{n}\sum_{t=1}^{n} W_t'u_t$. The expectation of a single summand is $E\left(W_t'u_t\right) = 0$ (this is a vector of length K), and the $K \times K$ covariance matrix is

$$\begin{aligned}
Cov\left(W_t'u_t\right) &= E\left(W_t'u_tu_t'W_t\right) \\
&= E\left(E\left(W_t'u_tu_t'W_t|W_t\right)\right) \\
&= E\left(W_t'E\left(u_tu_t'|W_t\right)W_t\right) \\
&= E\left(W_t'\sigma^2 IW_t\right) \\
&= \sigma^2 E\left(W_t'W_t\right) \\
&= :\sigma^2 S_{WW}
\end{aligned}$$

If the instruments satisfy the asymptotic identification condition $\frac{1}{n}W'W \xrightarrow{p} S_{WW}$, then the central limit theorem implies

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} W_t'u_t\right) \to N\left(0, \sigma^2 S_{WW}\right).$$

Thus, the second term in (5) converges in distribution. By Cramer's rule, if one factor of a product converges in probability and the other factor converges in distribution, then the product converges in distribution. So,

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) &= \left(\frac{1}{n}W'X\right)^{-1} \sqrt{n}\left(\frac{1}{n}W'u\right) \\
&\xrightarrow{d} S_{WX}^{-1} \cdot U
\end{aligned}$$

where $U \sim N\left(0, \sigma^2 S_{WW}\right)$. The final step is to move the matrix S_{WX}^{-1} into the covariance matrix of the normal distribution,

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2\left(S_{WX}\right)^{-1} S_{WW}\left(S_{WX}'\right)^{-1}\right).$$

Consider the covariance matrix. By definition

$$\begin{aligned}
\sigma^2 (S_{WX})^{-1} S_{WW} (S'_{WX})^{-1} &= \sigma^2 plim \left[\left(\frac{1}{n} W'X \right)^{-1} \left(\frac{1}{n} W'W \right) \left(\frac{1}{n} X'W \right)^{-1} \right] \\
&= \sigma^2 plim \left[n (W'X)^{-1} (W'W) (X'W)^{-1} \right] \\
&= \sigma^2 plim \left[\frac{1}{n} X'W (W'W)^{-1} W'X \right]^{-1} \\
&= \sigma^2 plim \left[\frac{1}{n} X'P_W X \right]^{-1}
\end{aligned}$$

where $P_W = W (W'W)^{-1} W'$ is a projection matrix. Hence, for large samples, approximately

$$\hat{\beta}_{IV} \sim N \left(\beta, \sigma^2 (X'P_W X)^{-1} \right).$$

9 Durbin-Wu-Hausman test

The difference between the estimators is

$$\begin{aligned}
& \hat{\beta}_{IV} - \hat{\beta}_{OLS} \\
= \;& (X'P_W X)^{-1} X'P_W y - (X'X)^{-1} X'y \\
= \;& (X'P_W X)^{-1} \left(X'P_W y - (X'P_W X) (X'X)^{-1} X'y \right) \\
= \;& (X'P_W X)^{-1} \left(X'P_W \left(I - X (X'X)^{-1} X' \right) y \right) \\
= \;& (X'P_W X)^{-1} (X'P_W M_X \, y)
\end{aligned}$$

We need to test if $X'P_W M_X \, y$ is significantly different from 0. This term is identically equal to zero for all variables in X that are instruments (i.e. that are also in W). Denote the exogenous variable in X by Z, then

$$Z'P_W M_X y = 0$$

because $Z'P_W = Z'$ since Z is a subset of W and $P_W W = W (W'W)^{-1} W'W = W$ and P_W is symmetric. Next, $Z'M_X = 0$ since Z is a subset of X and $M_X X = \left(I - X (X'X)^{-1} X' \right) X = 0$ and M_X is symmetric.

Denote by $\tilde{X}$ all possibly endogenous regressors. To test if $\tilde{X}'P_W M_X \, y$ is significantly different from zero, perform a Wald test of $\delta = 0$ in the regression

$$y = X\beta + P_W \tilde{X} \delta + u.$$

The OLS estimator of δ can be obtained by regression $M_X y$ on $M_X P_W \tilde{X}$, i.e.

$$\hat{\delta} = \left(\tilde{X}'P_W M_X P_W \tilde{X} \right)^{-1} \tilde{X}'P_W M_X y.$$

Since $\left(\tilde{X}'P_W M_X P_W \tilde{X} \right)^{-1}$ is positive definite, testing if $\delta = 0$ is equivalent to testing if $\tilde{X}'P_W M_X y$ is significantly different from 0. This reasoning is copied from Davidson and MacKinnon (2004, section 8.7).

10 GMM estimation of the linear regression model

Part I: Model description. The unknown parameter vector to be estimated is β (we do not attempt to estimate σ^2 at this point). The observations are $y_t = (X_t, y_t)$ where, of course, y_t and y_t are different things. The elementary zero functions are

$$f_t(\theta, y_t) = f_t(\beta, y_t, X_t) = y_t - X_t'\beta$$

because

$$E(y_t - X_t'\beta) = 0.$$

Note that this equation defines the model but does not say anything about the estimating equation.

Part II: Covariance matrix. The covariance matrix of $f(\theta, y)$ has elements

$$E(f_t(\theta, y_t) f_s(\theta, y_s)) = \begin{cases} \sigma^2 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases}$$

since f_t and f_s are independent, and $E(f_t^2) = E(u_t^2) = \sigma^2$.

11 GMM estimation of log-normal distribution

Part I: Model description. Suppose there is a random sample $X_1, \ldots, X_n$ from $X \sim LN(\mu, \sigma^2)$. The raw moment of order p is

$$E(X^p) = \exp\left(p\mu + \frac{1}{2}p^2\sigma^2\right).$$

The parameters to be estimated are $\theta = (\mu, \sigma^2)$. The observations are $y_t = X_t$. The elementary zero functions are

$$f_t(\theta, y_t) = \begin{bmatrix} f_{t1}(\theta, y_t) \\ f_{t2}(\theta, y_t) \end{bmatrix} = \begin{bmatrix} X_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ X_t^2 - \exp\left(2\mu + 2\sigma^2\right) \end{bmatrix}$$

because

$$\begin{aligned} E(f_t(\theta, y_t)) &= E\left(\begin{bmatrix} X_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ X_t^2 - \exp\left(2\mu + 2\sigma^2\right) \end{bmatrix}\right) \\ &= \begin{bmatrix} E\left(X_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right)\right) \\ E\left(X_t^2 - \exp\left(2\mu + 2\sigma^2\right)\right) \end{bmatrix} \\ &= \begin{bmatrix} E(X_t) - \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ E\left(X_t^2\right) - \exp\left(2\mu + 2\sigma^2\right) \end{bmatrix} \\ &= 0 \end{aligned}$$

Part II: Covariance matrix. The covariance matrix of $f(\theta, y)$ is

$$E(f(\theta, y)\ f(\theta, y)) = E\begin{pmatrix} f_{11}^2 & f_{11}f_{12} & \cdots & f_{11}f_{n1} & f_{11}f_{n2} \\ f_{12}f_{11} & f_{12}^2 & \cdots & f_{12}f_{n1} & f_{12}f_{n2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_{n1}f_{11} & f_{n1}f_{12} & \cdots & f_{n1}^2 & f_{n1}f_{n2} \\ f_{n2}f_{11} & f_{n2}f_{12} & \cdots & f_{n2}f_{n1} & f_{n2}^2 \end{pmatrix}.$$

Since the elements of the random sample are independent,

$$E\left(f_{ti}\left(\theta, y_t\right) f_{sj}\left(\theta, y_t\right)\right) = E\left(f_{tj}\left(\theta, y_t\right)\right) \cdot E\left(f_{sj}\left(\theta, y_t\right)\right) = 0$$

for $s \neq t$ and $i, j \in \{1, 2\}$. Hence, the covariance matrix is of block-diagonal form,

$$E\left(f(\theta, y)\ f\left(\theta, y\right)\right) = E\begin{pmatrix} f_{11}^2 & f_{11}f_{12} & & & \\ f_{12}f_{11} & f_{12}^2 & & & \\ & & \ddots & & \\ & & & f_{n1}^2 & f_{n1}f_{n2} \\ & & & f_{n2}f_{n1} & f_{n2}^2 \end{pmatrix}.$$

The blocks have elements

$$\begin{aligned} E\left(f_{t1}\left(\theta, y_t\right)^2\right) &= E\left(\left(X_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right)\right)^2\right) \\ &= E\left(X_t^2 - 2X_t \exp\left(\mu + \frac{1}{2}\sigma^2\right) + \exp\left(2\mu + \sigma^2\right)\right) \\ &= E\left(X_t^2\right) - 2E\left(X_t\right)\exp\left(\mu + \frac{1}{2}\sigma^2\right) + \exp\left(2\mu + \sigma^2\right) \\ &= \exp\left(2\mu + 2\sigma^2\right) - 2\exp\left(\mu + \frac{1}{2}\sigma^2\right)\exp\left(\mu + \frac{1}{2}\sigma^2\right) + \exp\left(2\mu + \sigma^2\right) \\ &= \exp\left(2\mu + 2\sigma^2\right) - \exp\left(2\mu + \sigma^2\right), \end{aligned}$$

In a similar way, one can derive the other two covariances

$$\begin{aligned} E\left(f_{t1}f_{t2}\right) &= \exp\left(3\mu + \frac{9}{2}\sigma^2\right) - \exp\left(3\mu + \frac{5}{2}\sigma^2\right) \\ E\left(f_{t2}\left(\theta, y_t\right)^2\right) &= \exp\left(4\mu + 8\sigma^2\right) - \exp\left(4\mu + 4\sigma^2\right) \end{aligned}$$

12 GMM estimation of a general asset pricing model

Part I: Model description. The parameters to be estimated depend on the specific case. For instance, if the stochastic discount factor is

$$m_t = \beta\left(\frac{c_t}{c_{t-1}}\right)^{-\gamma}$$

then $\theta = (\beta, \gamma)'$. The observations are time series of prices $p_1, \ldots, p_n$ and payoffs $x_1, \ldots, x_n$. If required, other variables may be included, e.g. time series of consumption c_t. The elementary zero functions are

$$f_t\left(\theta, y_t\right) = m_{t+1}\left(\theta\right) x_{t+1} - p_t$$

because $E\left(f_t\left(\theta, y_t\right)\right) = E\left(m_{t+1}\left(\theta\right) x_{t+1}\right) - E\left(p_t\right)$ and asset pricing models imply the general moment condition

$$E\left(p_t\right) = E\left(m_{t+1}\left(\theta\right) x_{t+1}\right).$$

Part II: Covariance matrix. Since time series are dependent over time, the observations are not independent,

$$E\left(f(\theta,y)\ f\left(\theta,y\right)\right)=E\begin{pmatrix} f_{11}^{2} & \cdots & f_{11}f_{n1} \\ \vdots & \ddots & \vdots \\ f_{n1}f_{11} & \cdots & f_{n1}^{2} \end{pmatrix}.$$

A typical element of this matrix is

$$\begin{aligned} E\left(\left(f_t(\theta,y_t)\ f_s\left(\theta,y_s\right)\right)\right) &= E\left(\left(m_{t+1}\left(\theta\right)x_{t+1}-p_t\right)\left(m_{s+1}\left(\theta\right)x_{s+1}-p_s\right)\right) \\ &= E\left(m_{t+1}\left(\theta\right)m_{s+1}\left(\theta\right)x_{t+1}x_{s+1}\right)-E\left(m_{t+1}\left(\theta\right)x_{t+1}p_s\right) \\ &\quad -E\left(m_{s+1}\left(\theta\right)x_{s+1}p_t\right)+E\left(p_tp_s\right). \end{aligned}$$

The exact form depends on the time series properties of the series.

13 Consistency of GMM estimator

Limiting estimation functions

$$\alpha\left(\theta\right)=plim\frac{1}{n}Z'f\left(\theta,y\right)$$

and limiting estimation equations

$$\alpha\left(\theta\right)=0.$$

Here we assume that $\frac{1}{n}Z'f\left(\theta,y\right)$ actually converges in probability, i.e. that a law of large numbers holds. In addition, we assume the asymptotic identification condition, $\alpha\left(\theta\right)\neq\alpha\left(\theta_0\right)$ for all $\theta\neq\theta_0$.

We do not prove consistency rigorously, but only heuristically (following Davidson and MacKinnon, 2004, p. 218f). To do so, assume that $\hat{\theta}$ has a deterministic probability limit, $\hat{\theta}\rightarrow\theta_\infty$. We now show that θ_∞ must equal the true value θ_0 if the asymptotic identification condition is satisfied. We start by assuming the opposite, i.e. that $\theta_\infty\neq\theta_0$. By definition, the GMM estimator satisfies

$$\frac{1}{n}Z'f\left(\hat{\theta},y\right)=0.$$

Applying the *plim*-operator on both sides gives

$$\begin{aligned} plim\frac{1}{n}Z'f\left(\hat{\theta},y\right) &= plim0 \\ \frac{1}{n}Z'f\left(plim\hat{\theta},y\right) &= 0 \\ \frac{1}{n}Z'f\left(\hat{\theta}_\infty,y\right) &= 0 \\ \alpha\left(\theta_\infty\right) &= 0 \end{aligned}$$

However, the asymptotic identification condition requires that $\alpha\left(\theta\right)\neq 0$ for all $\theta\neq\theta_0$. It follows that $\theta_\infty=\theta_0$, i.e. the probability limit is the true value. The GMM estimator is consistent.

14 Asymptotic normality of GMM estimator

A first order Taylor series expansion of

$$\frac{1}{n} Z' f(\theta) = 0$$

in $\hat{\theta}$ around θ_0 gives

$$\frac{1}{n} Z' f\left(\hat{\theta}\right) = \frac{1}{n} Z' f(\theta_0) + \frac{1}{n} Z' F(\theta_0) \left(\hat{\theta} - \theta_0\right) = 0$$

with Jacobian matrix

$$F(\theta) = \frac{\partial f(\theta)}{\partial \theta}.$$

A typical element of the $n \times K$ matrix $F(\theta)$ is

$$F_{ti}(\theta) = \frac{\partial f_t(\theta)}{\partial \theta_i}.$$

[What is $F(\theta)$ in the linear regression model?]

Note that $F(\theta)$ also depends on y and, thus, it is stochastic. Multiplication by $\sqrt{n}$ gives

$$\sqrt{n} \frac{1}{n} Z' f(\theta_0) + \frac{1}{n} Z' F(\theta_0) \sqrt{n} \left(\hat{\theta} - \theta_0\right) = 0$$

or

$$\sqrt{n} \left(\hat{\theta} - \theta_0\right) = -\left(\frac{1}{n} Z' F(\theta_0)\right)^{-1} \sqrt{n} \frac{1}{n} Z' f(\theta_0).$$

Now, let $n \to \infty$. By the law of large numbers the mean

$$\frac{1}{n} Z' F(\theta_0)$$

converges in probability. We assume that the limit $plim \frac{1}{n} Z' F(\theta_0)$ is deterministic and nonsingular. As to the second term on the right hand side, i.e.

$$\sqrt{n} \left(\frac{1}{n} Z' f(\theta_0)\right)$$

and we can invoke the central limit theorem. Hence, asymptotic normality of $n^{-1/2} Z' f(\theta_0)$ is established. What are the expectation and the covariance matrix of the asymptotic normal distribution? The expectation vector is obviously 0, since $E(f(\theta_0)) = 0$. As to the covariance matrix,

$$E\left[\left(\sqrt{n} \frac{1}{n} Z' f(\theta_0)\right) \left(\sqrt{n} \frac{1}{n} Z' f(\theta_0)\right)'\right] = E\left[\frac{1}{n} Z' f(\theta_0) f'(\theta_0) Z\right].$$

For $n \to \infty$, this converges to

$$plim \frac{1}{n} Z' \Omega Z.$$

Note that the convergence by the law of large numbers breaks down if the second moment of $f(\theta_0) f'(\theta_0)$ does not exist, i.e. if the fourth moment of $f(\theta_0)$ does not exist. This normally happens for return distributions!

The asymptotic distribution of $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ also has expectation vector 0. Its covariance matrix is the "sandwich covariance matrix"

$$\left(plim\frac{1}{n}Z'F\left(\theta_0\right)\right)^{-1}\left(plim\frac{1}{n}Z'\Omega Z\right)\left(plim\frac{1}{n}F\left(\theta_0\right)'Z\right)^{-1}$$

or

$$\left(plim\frac{1}{n}J'W'F\left(\theta_0\right)\right)^{-1}\left(plim\frac{1}{n}J'W'\Omega WJ\right)\left(plim\frac{1}{n}F\left(\theta_0\right)'WJ\right)^{-1}.$$

15 Asymptotic efficiency

The estimating equations are

$$\frac{1}{n}Z'f(\theta)=0.$$

The matrix Z can be chosen by the user. The optimal choice depends on the assumptions about $F(\theta)$ and Ω. We consider three cases.

15.1 Case 1

We start with the simplest case, $\Omega=\sigma^2 I$ and $F(\theta_0)$ is predetermined, i.e. F_t and f_t are contemporaneously uncorrelated,

$$E\left(F_t(\theta_0)f_t(\theta_0)\right)=0.$$

This assumption is similar to the assumption that X_t and u_t are contemporaneously uncorrelated, in the linear regression model. To keep the notation simple, ignore the probability limit and the factors of n^{-1}. The asymptotic covariance matrix of the GMM estimator is

$$\left(Z'F_0\right)^{-1}Z'\Omega Z\left(F_0{}'Z\right)^{-1}$$

where F_0 is an abbreviation for $F(\theta_0)$. We now make use of the following lemma: Let A and B be two positive definite matrices; then $A-B$ is positive semidefinite if and only if $B^{-1}-A^{-1}$ is positive semidefinite.

The inverse of the covariance matrix (often called precision matrix, as a small variance means a large precision) is

$$\begin{aligned}
&\left(F_0{}'Z\right)\left(Z'\Omega Z\right)^{-1}\left(Z'F_0\right)\\
=\;&\frac{1}{\sigma^2}F_0{}'Z\left(Z'Z\right)^{-1}Z'F_0\\
=\;&\frac{1}{\sigma^2}F_0{}'P_Z F_0
\end{aligned}\tag{6}$$

where $P_Z=Z\left(Z'Z\right)^{-1}Z'$ is a projection matrix. If we let $Z=F_0$ then the precision matrix becomes

$$\frac{1}{\sigma^2}F_0{}'F_0.\tag{7}$$

The difference between (7) and (6) is

$$\frac{1}{\sigma^2}F_0{}'F_0 - \frac{1}{\sigma^2}F_0{}'P_Z F_0 = \frac{1}{\sigma^2}F_0'(I - P_Z)F_0.$$

Since the "hat matrix" $I - P_Z$ is a positive semidefinite matrix, the best possible precision matrix is $\frac{1}{\sigma^2}F_0{}'F_0$, and hence the optimal choice is $Z = F(\theta_0)$.

Special case: Standard linear regression model with covariance matrix $Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

15.2 Case 2

Suppose, $\Omega = \sigma^2 I$ but $F(\theta_0)$ is not predetermined, i.e.

$$E(F_t(\theta_0)f_t(\theta_0)) \neq 0.$$

Define

$$\begin{aligned} \bar{F}_t \quad &: \quad = E\left(F_t(\theta_0)|\Omega_t\right) \\ V_t \quad &: \quad = F_t(\theta_0) - \bar{F}_t. \end{aligned} \tag{8}$$

Consider the covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ if we set $Z_t = \bar{F}_t$. It is

$$\sigma^2\left(plim\frac{1}{n}\bar{F}'F_0\right)^{-1}\left(plim\frac{1}{n}\bar{F}'\bar{F}\right)\left(plim\frac{1}{n}F_0'\bar{F}\right)^{-1}. \tag{9}$$

By definition of (8) we have

$$plim\frac{1}{n}\bar{F}'F_0 = plim\frac{1}{n}\bar{F}'\left(\bar{F} + V\right) = plim\frac{1}{n}\bar{F}'\bar{F},$$

so (9) simplifies to

$$\sigma^2\left(plim\frac{1}{n}\bar{F}'\bar{F}\right)^{-1}. \tag{10}$$

For other instrument matrices Z we find

$$\sigma^2\left(plim\frac{1}{n}Z'\bar{F}\right)^{-1}\left(plim\frac{1}{n}Z'Z\right)\left(plim\frac{1}{n}\bar{F}'Z\right)^{-1}$$

where $plim\frac{1}{n}Z'\bar{F} = plim\frac{1}{n}Z'F_0$. Using the shorthand notation, the limit of the precision matrix is $1/\sigma^2$ times

$$\bar{F}'Z\left(Z'Z\right)^{-1}Z'\bar{F} = \bar{F}'P_Z\bar{F}$$

while the precision matrix of (10) is $1/\sigma^2$ times

$$\bar{F}'\bar{F}$$

so their difference is always positive semidefinite and the optimality of $Z = \bar{F}$ is proved. In practice, we substitute $\bar{F}$ (which is normally unobserved) by its linear projection $P_W F(\theta)$. Then the estimating equations become

$$\frac{1}{n}F(\theta)'P_W f(\theta) = 0$$

and the asymptotic covariance matrix of $\hat{\theta}$ is estimated by

$$\begin{aligned} Cov(\hat{\theta}) &= \hat{\sigma}^2 \left(F(\hat{\theta})' P_W P_W F(\hat{\theta}) \right)^{-1} \\ &= \hat{\sigma}^2 \left(F(\hat{\theta})' P_W F(\hat{\theta}) \right)^{-1} \end{aligned}$$

where $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} f_t^2(\hat{\theta})$.

15.3 Case 3

Suppose, Ω is unknown. Since it contains $n\,(n+1)\,/2$ elements it cannot be estimated consistently. The asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ has the sandwich form

$$\left(plim \frac{1}{n} J'W'F_0 \right)^{-1} \left(plim \frac{1}{n} J'W'\Omega W J \right) \left(plim \frac{1}{n} F_0{}'WJ \right)^{-1}.$$

For $J = (W'\Omega W)^{-1} W'F_0$ we get the optimal covariance matrix

$$\begin{aligned} Cov\left(\sqrt{n}(\hat{\theta} - \theta_0) \right) &= \left(plim \frac{1}{n} F_0' W \left(W'\Omega W \right)^{-1} W'F_0 \right)^{-1} \\ &\quad \times \left(plim \frac{1}{n} F_0' W \left(W'\Omega W \right)^{-1} W'\Omega W \left(W'\Omega W \right)^{-1} W'F_0 \right) \\ &\quad \times \left(plim \frac{1}{n} F_0{}' W \left(W'\Omega W \right)^{-1} W'F_0 \right)^{-1} \\ &= \left(plim \frac{1}{n} F_0' W \left(W'\Omega W \right)^{-1} W'F_0 \right)^{-1}. \end{aligned}$$

We now show (a bit heuristically) that $J = (W'\Omega W)^{-1} W'F_0$ is in fact the optimal weighting matrix, see also exercise 9.1 in Davidson and MacKinnon (2004). The optimal precision matrix of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ is (without plims and n^{-1})

$$F_0' W \left(W'\Omega W \right)^{-1} W'F_0$$

which can be written as

$$F_0' \Omega^{-1/2} P_{\Omega^{1/2} W} \Omega^{-1/2\prime} F_0$$

with the projection matrix

$$P_{\Omega^{1/2} W} = \Omega^{1/2} W \left(W'\Omega W \right)^{-1} W' \Omega^{1/2\prime}.$$

The non-optimal precision matrix for arbitrary J is (without plims and n^{-1})

$$F_0{}' W J \left(J' W' \Omega W J \right)^{-1} J' W' F_0$$

which can be written as

$$F_0' \Omega^{-1/2} P_{\Omega^{1/2} W J} \Omega^{-1/2\prime} F_0$$

where

$$P_{\Omega^{1/2}WJ} = \Omega^{1/2} WJ \left(J'W'\Omega WJ\right)^{-1} J'W'\Omega^{1/2\prime}.$$

The difference between the precision matrices is

$$\begin{aligned} & F_0'\Omega^{-1/2} P_{\Omega^{1/2}W} \Omega^{-1/2\prime} F_0 - F_0'\Omega^{-1/2} P_{\Omega^{1/2}WJ} \Omega^{-1/2\prime} F_0 \\ = \; & F_0'\Omega^{-1/2} \left(P_{\Omega^{1/2}W} - P_{\Omega^{1/2}WJ}\right) \Omega^{-1/2\prime} F_0. \end{aligned}$$

The matrix $P_{\Omega^{1/2}W} - P_{\Omega^{1/2}WJ}$ is obviously symmetric; and it is easy to show that it is idempotent, and hence it is a positive semidefinite orthogonal projection matrix.

The optimal asymptotic covariance matrix

$$\left(plim \frac{1}{n} F_0' W \left(W'\Omega W\right)^{-1} W' F_0 \right)^{-1}$$

contains the unknown Ω which cannot be estimated consistently (due to its $n(n+1)/2$ elements). However, the $L \times L$ matrix $\frac{1}{n} W'\Omega W$ can be estimated consistently. Suppose that $\hat{\Sigma}$ is a consistent estimator of $\frac{1}{n} W'\Omega W$. Then

$$\left(plim \frac{1}{n^2} F_0' W \hat{\Sigma}^{-1} W' F_0 \right)^{-1}$$

is the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$ and

$$\widehat{Cov}(\hat{\theta}) = n \left(F(\hat{\theta})' W \hat{\Sigma}^{-1} W' F(\hat{\theta}) \right)^{-1}$$

is the estimated covariance matrix of $\hat{\theta}$.

16 Alternative notation for GMM

As usual, we start with a Taylor series expansion of $\bar{g}$ in $\hat{\theta}$ around the true value θ_0,

$$\bar{g}\left(\hat{\theta}\right) = \bar{g}\left(\theta_0\right) + G\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right).$$

Pre-multiply from the left by $G(\hat{\theta})'A$,

$$G(\hat{\theta})'A\bar{g}\left(\hat{\theta}\right) = G(\hat{\theta})'A\bar{g}\left(\theta_0\right) + G(\hat{\theta})'AG\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right).$$

Note that $G(\hat{\theta})'A\bar{g}\left(\hat{\theta}\right) = 0$ are the first order conditions for the minimization problem $\bar{g}(\hat{\theta})'Ag(\hat{\theta})$. Hence

$$G(\hat{\theta})'A\bar{g}\left(\theta_0\right) + G(\hat{\theta})'AG\left(\theta_0\right)\left(\hat{\theta} - \theta_0\right) = 0$$

or

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -\left(G(\hat{\theta})'AG\left(\theta_0\right)\right)^{-1} G(\hat{\theta})'A\sqrt{n}\bar{g}\left(\theta_0\right).$$

By assumption, $\sqrt{n}\bar{g}\left(\theta_0\right)$ is asymptotically normal with expectation 0 and covariance matrix V. Asymptotically, $G(\hat{\theta})$ and $G(\theta_0)$ are the same since $\hat{\theta}$ is consistent. We may denote both simply by G. Hence, $\sqrt{n}\left(\hat{\theta}-\theta\right)$ is also asymptotically normal with expectation 0. Its covariance matrix is

$$(G'AG)^{-1}\, G'AVA'G\,(G'AG)^{-1}\,.$$

Hence, the weighting matrix A has an impact on the asymptotic covariance matrix and should be chosen optimally to minimize the covariance matrix.

17 Optimal weighting matrix in the alternative notation

It can be shown that

$$(G'AG)^{-1}G'AVAG(G'A'G)^{-1} - (G'V^{-1}G)^{-1}$$

is positive semidefinite for all A. Hence, the best possible covariance matrix is $(G'V^{-1}G)^{-1}$, and the optimal weighting matrix is

$$A = V^{-1}.$$

Since V is usually unknown, use the sequence of weight matrices $A_n = \hat{V}_n^{-1}$.

18 List of equivalences

First a brief review of the notation used e.g by Davidson and MacKinnon.

elementary zero functions	$f_t\,(\theta, y_t)$
	$E(f) = 0$
	$E(ff') = \Omega$
instrumental variables	W
moment conditions	$E(W'f) = 0$
estimation equations	$J'W'f = 0$
asymptotic covariance of $\hat{\theta}$	$(J'W'F)^{-1}\,(J'W'\Omega WJ)\,(F'WJ)^{-1}$
to be estimated	$W'\Omega W \quad (\text{or } \Sigma)$
optimal J matrix	$(W'\Omega W)^{-1}\, W'F$
optimal asympt. covariance	$\left(F'W\,(W'\Omega W)^{-1}\, W'F\right)^{-1}$

And a brief review of the notation used by many others.

moment conditions	$E(g) = 0$
	$\bar{g} = 0$
estimation by minimizing	$\bar{g}'A\bar{g}$
or estimation equations	$G'A\bar{g} = 0$
asymptotic covariance of $\hat{\theta}$	$(G'AG)^{-1}\, G'AVA'G\,(G'AG)^{-1}$
to be estimated	V
optimal A matrix	V^{-1}
optimal asympt. covariance	$\left(G'V^{-1}G\right)^{-1}$

List of equivalences between both notations.

Davidson & MacKinnon	Other notation
$E(W'f) = 0$	$E(g) = 0$
$W'f$	$\bar{g}$
$J'W'f = 0$	$G'A\bar{g} = 0$
J	$A'G$
$W'F$	G
$W'\Omega W$ or Σ	V
$(J'W'F)^{-1}(J'W'\Omega WJ)(F'WJ)^{-1}$	$(G'AG)^{-1}G'AVA'G(G'AG)^{-1}$
$J = (W'\Omega W)^{-1}W'F$	$A'G = V^{-1}G$
$\left(F'W(W'\Omega W)^{-1}W'F\right)^{-1}$	$\left(G'V^{-1}G\right)^{-1}$