Asymptotic Distribution of M-estimator

The following topics are covered today: Today we briefly covered global and local consistency and asymptotic distribution of general M-estimators, including maximum likelihood(ML) and generalized method of moments(GMM). We also dicuss briefly quantile regression and the issue of asymptotic efficiency. The relevant reading for today's class is Ch 4 of Amemiya and Section 1-6 of the Newey and McFadden Chapter in Vol 4 of Handbook of Econometrics.

Consistency: continued from last time. There is a distinction between global consistency and local consistency. Assuming the parameter space Θ is compact.

1. Global Condition(Thm 4.1.1 in Amemiya, pp106): $\sup_{\theta\in\Theta}|Q_n(\theta)-Q(\theta)|overset{p}{\longrightarrow}0$. And $Q(\theta)<Q(\theta_0)$ for $\theta\neq\theta_0$. Then there is $\hat{\theta}\xrightarrow{p}\theta_0$ for $\hat{\theta}=argmax_{\theta\in\Theta}Q_n(\theta)$, the parameter vector that globally maximized the sample objective function.

2. Local Condition(Thm 4.1.2 in Amemiya, p110): Assuming there exists a small neighborhood N around θ_0 such that $\sup_{\theta\in N}\left|\frac{\partial Q_n(\theta)}{\partial\theta}-\frac{\partial Q(\theta)}{\partial\theta}\right|\xrightarrow{p}0$, and that $Q(\theta)>Q(\theta_0)$ for $\theta\neq\theta_0$ and $\theta\in N$. Then if we use $\hat{\Theta}$ to denote the set of θ for which $\frac{\partial Q_n(\theta)}{\partial\theta}=0$, then the claim is that some point in $\hat{\Theta}$ will be a consistent estimate for θ_0, although which point is the consistent estimate is not known in advance, i.e., $\forall\epsilon>0$, $\lim_{n\to\infty}P\left(\inf_{\theta\in\hat{\Theta}}||\theta-\theta_0||>\epsilon\right)=0$. Practically, for the local consistency condition, you only need to check two properties, (1) $\frac{\partial Q(\theta_0)}{\partial\theta}=0$ and (2) $\frac{\partial^2 Q(\theta_0)}{\partial\theta\partial\theta'}$ negative definite.

Consistency for MLE: Read sec 4.2 pp115 Ch4 in Amemiya. Let $L(y_1,\ldots,y_n,\theta)$ be the JOINT density for the data $y_1,\ldots,y_n$. Then $Q_n(\theta)\equiv\frac{1}{n}\log L(y_1,\ldots,y_n,\theta)$.

1. For the particular case of iid data, there is $Q_n(\theta)=\frac{1}{n}\sum_{t=1}^{n}\log f(y_t,\theta)$. If θ_0 is identified in the sense that for $\theta\neq\theta_0$, there is a positive probabiity of y_t(under θ_0) where $f(y_t,\theta)\neq f(y_t,\theta_0)$, then the following holds

$$E\log f(y;\theta)-E\log f(y;\theta_0)<\log E\frac{f(y;\theta)}{f(y;\theta_0)}=\log\int f(y;\theta)\,dy=\log 1=0.$$

where the second inequality comes from the Jensen's inequality plus the identification condition. Usually, to justify $\sup_{\theta\in\Theta}Q_n(\theta)\xrightarrow{p}0$, we need some dominance condition like $E\sup_{\theta\in\Theta}|\log f(y;\theta)|<\infty$, see in particular Thm 4.2.1 in p116 in Amemiya, which we discuss in the last note when applying stochastic equicontinuity. However, the condition that $E\sup_{\theta\in\Theta}|\log f(y;\theta)|<\infty$ will be violated if the support of the y_t depends on the parameters, for example if y_t is distributed in (θ_0,∞). because for $\theta>\theta_0$, there is be some y such that $f(y,\theta)=0$, in which case $\log f(y,\theta)=-\infty$ so that the dominance conditions fails. However, as long as you imposed that constraint that $\theta\geq\min(y_1,\ldots,y_n)$ in maximizing $Q_n(\theta)$ to obtain $\hat{\theta}$, recognizing $Q_n(\theta)$ is undefined for $\theta<\min(y_1,\ldots,y_n)$. This is for two reasons, (1) because $\min(y_1,\ldots,y_n)\xrightarrow{a.s.}\theta_0$, as we discussed in last lecture for the uniform case, so that any $\theta<\theta_0$ will be ruled out by the data eventually. (2) Then if we restricted ourselve to the subset of the parameter space for which $\theta\leq\theta_0$, then both $E\sup_{\theta\leq\theta_0}|\log f(y;\theta)|<\infty$ and $E\log f(y;\theta)<E\log f(y;\theta_0)$ will still be true. So the rule is that you always get consistent estimate from maximum likelihood estimation, even if you have a parameter-dependent support of the data, as long as you incorporate the constraints defined by the data on the parameter space properly into optimizing $Q_n(\theta)$.

2. In the general case when y_t is not iid, there is still $E \log L(y_1, \ldots, y_n; \theta) \leq EL(y_1, \ldots, y_n; \theta_0)$ by the same application of Jensen's equality. However, to justify the strict inequality $<$ is harder, see pp115-116 Amemiya.

3. Sometimes the local condition may be important even when the global condition fails, especially when Θ is not compact. Look at the mixture of normal example in pp119 Example 4.2.2 in Amemiya, where $y_t \sim \lambda N(\mu_1, \sigma_1^2) + (1-\lambda) N(\mu_2, \sigma_2^2)$. The likelihood function is $L = \prod_{t=1}^n \left[\frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp\left(-(y_t - u_1)^2 / (2\sigma_1^2) \right) + \frac{1-\lambda}{\sqrt{2\pi}\sigma_2} \exp\left(-(y_t - u_2)^2 / (2\sigma_2^2) \right) \right]$. Set $u_1 = y_1$(or any y_t for that matter), and let $\sigma_1 \to 0$(but never touch 0), then L increases to ∞ without bound. You can't do this with the nonmixture model where there is just $y_t \sim N(\mu_1, \sigma_1^2)$, because the other terms $t \neq 1$ will go to 0 at an exponential rate, fast that the linear rate at which the first $t = 1$ goes to ∞. In the mixture model, the existence of the mixture part for $1 - \lambda$ prevents the other $t \neq 1$ terms from decreasing to 0, thus creating the above problem. However, a local root of the log likelihood function can still be consistent, see pp120 Amemiya.

Consistency for GMM: Read Newey and McFadden pp2132. Now $Q_n(\theta) = g_n(\theta)' W g_n(\theta)$, for $g_n(\theta) = \frac{1}{n} \sum_{t=1}^n g(z_t, \theta)$, and W is the positive definite weighting matrix. If $\sup_{\theta \in \Theta} |g_n(\theta) - Eg(z_t, \theta)| \xrightarrow{p} 0$ and $Eg(z_t, \theta) = 0$ iff $\theta = \theta_0$. Then $\hat{\theta} \equiv argmax_\theta Q_n(\theta) \xrightarrow{p} 0$. (Note that if W is not full rank, you can essentially drop moment conditions until it is full rank.)

Identification in Linear Models: Global identification in nonlinear GMM model is usually difficult and is usually "assumed". However, identification in linear models usually reduces to condition that the sample var-cov matrix for regressors is full rank, this is $Ex_t x_t'$ for iid models, and $\lim_{T \to \infty} \frac{1}{n} \sum_{t=1}^n x_t x_t'$ for fixed regressors. For least Square, $\frac{1}{n} \sum_{t=1}^n (y_t - x_t'\beta)^2 \xrightarrow{p} E(y - x'\beta)^2$. Iff $Ex_t x_t'$ full rank

$$E(y - x'\beta)^2 - E(y - x'\beta_0)^2 = E[x'(\beta - \beta_0)]^2 = (\beta - \beta_0)' Ex_t x_t' (\beta - \beta_0) > 0 \quad \text{if} \quad \beta \neq \beta_0$$

Quantile Regression: The identifying assumption for quantile regression models is to assume that the conditional τth quantile of y_t given x_t is a linear regression function $x_t'\beta_0$, i.e. $Pr(y_t \leq x_t'\beta_0 | x_t) \equiv F_y(x_t'\beta_0 | x_t) = \tau$. The $\tau = \frac{1}{2}$th quantile is the median. The following moment conditions are satisfied for a linear model identified by conditional quantile assumption: $E(\tau - 1(y \leq x_t'\beta_0)) x_t = Ex_t(\tau - Pr(y_t \leq x_t'\beta_0 | x_t)) = 0$. Motivated by the population moment condition, in the sample we look for $\hat{\beta}$ such that

$$0 \approx \frac{1}{n} \sum_{t=1}^n x_t \left(\tau - 1\left(y_t \leq x_t'\hat{\beta} \right) \right) = \frac{1}{n} \sum_{t=1}^n x_t \left[\tau 1\left(y > x_t'\hat{\beta} \right) - (1-\tau) 1\left(y_t \leq x_t'\hat{\beta} \right) \right].$$

Integrate this first order condition back to obtain the convex objective function $Q_n(\beta)$ for quantile regression: (Note $x^+ = \max(0, x)$ and $x^- = \min(0, x)$).

$$Q_n(\beta) = \frac{1}{n} \sum_{t=1}^n [\tau - 1(y_t \leq x_t'\beta)](y_t - x_t'\beta) = \frac{1}{n} \sum_{t=1}^n \left[\tau (y_t - x_t'\beta)^+ + (1-\tau)(y - x'\beta)^- \right]$$

To think of relating the "derivative" of this objective function to the moment condition above, think of $\frac{\partial x^+}{\partial x} = 1(x > 0)$ and $\frac{\partial x^-}{\partial x} = -1(x < 0)$, so that $\frac{\partial (y - x'\beta)^+}{\partial \beta} = 1(y - x'\beta > 0) x_t$, and $\frac{\partial (y - x'\beta)^-}{\partial \beta} = -1(y - x'\beta \leq 0) x_t$ etc. When $\tau = \frac{1}{2}$, $Q_n(\beta) = \frac{1}{n} \sum_{t=1}^n |y_t - x_t'\beta|$ becomes the Least Absolute Deviation (LAD) regression, which looks for the conditional median.

To show formally that $Ex_t x_t'$ implies global consistency for the linear quantile regression model, consider for $\sup_{\theta\in\Theta}|Q_n(\beta)-Q(\theta)|\xrightarrow{p}0$ where $Q(\beta)=E(\tau-1(y\leq x'\beta))(y-x'\beta)$ there is

$$\begin{aligned}
&Q(\beta)-Q(\beta_0)\\
=&E\left[(\tau-1(y\leq x'\beta))(y-x'\beta)-(\tau-1(y<x'\beta_0))(y-x'\beta_0)\right]\\
=&E\tau(x'\beta_0-x'\beta)+E(y-x'\beta)\left[1(y\leq x'\beta_0)-1(y\leq x'\beta)\right]+E(x'\beta-x'\beta_0)1(y<x'\beta_0)\\
=&E(y-x'\beta)\left[1(y\leq x'\beta_0)-1(y<x'\beta)\right]\\
=&E_x\int_{x'\beta}^{x'\beta_0}(y-x'\beta)f(y|x)\,dy
\end{aligned}$$

Assuming that the set of x_t and β is bounded, and that $\exists\delta>0$ such that $f_y(y|x)>\delta$ uniformly in x for all $|y-x'\beta_0|\leq\delta$, so that $\exists M$ large such that $M\delta>>|x'(\beta-\beta_0)|$ for all x and β, we can continue the inequalities as:

$$\begin{aligned}
E_x\int_{x'\beta}^{x'\beta_0}(y-x'\beta)f(y|x)\,dy=E_x\int_{x'(\beta-\beta_0)}^{0}(u-x'(\beta-\beta_0))f(u|x)\,du&\geq E\int_{x'(\beta-\beta_0)}^{0}(u-x'(\beta-\beta_0))f(u|x)\,du\\
&\geq E_x\int_{\frac{x'(\beta-\beta_0)}{M}}^{0}\left(u-\frac{x'(\beta-\beta_0)}{M}\right)f(u|x)\,du\geq\frac{\delta}{M^2}E(x'(\beta-\beta_0))^2\\
&=\frac{\delta}{M^2}(\beta-\beta_0)'(Exx')(\beta-\beta_0)>0\quad\text{for}\quad\beta\neq\beta_0\quad\text{if}\quad Exx'\quad\text{nonsingular}
\end{aligned}$$

The objective function $Q_n(\beta)$ for quantile regression has two features, (1) the objective function is convex so that pointwise convergence is sufficient for uniform convergence over any compact set Θ and the parameter space does not have to be compact. More on this below. (2) No moment conditions are needed for y_t to obtain pointwise convergence, this is done by subtracting $Q_n(\beta_0)$, a function that does not depend on β from $Q_n(\beta)$, and show that $Q_n(\beta)-Q_n(\beta_0)\xrightarrow{p}Q(\beta)-Q(\beta_0)$. The reason that this scaling is useful is because the absolute value is a norm for which we can apply the triangular inequality: $||a|-|b||\leq|a-b|$, which allows us to write for the LAD,

$$\frac{1}{n}\sum_{t=1}^{n}\left||y_t-x_t'\beta|-|y_t-x_t'\beta_0|\right|\leq\frac{1}{n}\sum_{t=1}^{n}|x_t'\beta-x_t'\beta_0|=\frac{1}{n}\sum_{t=1}^{n}|x_t||\hat{\beta}-\hat{\beta}_0|$$

Similarly for the general τth quantile regression, we can use $|a^+-b^+|\leq|a-b|$ and $|a^--b^-|\leq|a-b|$ to put a bound on:

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left|\left[\tau(y_t-x_t'\beta)^+(1-\tau)(y-x'\beta)^-\right]-\left[\tau(y_t-x_t'\beta_0)^+(1-\tau)(y-x'\beta_0)^-\right]\right|\\
\leq&\frac{1}{n}\sum_{t=1}^{n}\tau|x'\beta-x'\beta_0|+(1-\tau)|x'\beta-x'\beta_0|=\frac{1}{n}\sum_{t=1}^{n}|x'\beta-x'\beta_0|=\frac{1}{n}\sum_{t=1}^{n}|x_t||\hat{\beta}-\hat{\beta}_0|
\end{aligned}$$

We can't do the same trick for least square because the square is not a norm for which the triangular inequality is not applicable.

Concavity and Noncompact parameter set: when $Q_n(\theta)$ is concave for maximization problem(or convex for minimization problem), then two good features are available,

1. If $Q_n(\theta)$ is concave in θ almost surely, then pointwise convergence of $Q_n(\theta)$ to $Q(\theta)$ implies uniform convergence over any compact set Θ.

2. It is only neccessary for $Q(\theta)$ to be uniquely maximized at θ_0 over some neighborhood around θ_0 to obtain global consistency.

For details read Newey and McFadden pp2133, for the basic intuition think of the simple picture we draw in class.

Asymptotic Normality:

The General Framework: Read Amemiya, Thm 4.1.3, pp111. Everything is just some form of first order Taylor Expansion:

$$\frac{\partial Q_n\left(\hat{\theta}\right)}{\partial\theta}=0\Longleftrightarrow\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}+\sqrt{n}\left(\hat{\theta}-\theta_0\right)\frac{\partial^2 Q_n\left(\theta^*\right)}{\partial\theta\partial\theta'}$$

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=\left(\frac{\partial^2 Q_n\left(\theta^*\right)}{\partial\theta\partial\theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}\overset{LD}{=}\left(\frac{\partial^2 Q\left(\theta_0\right)}{\partial\theta\theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}\xrightarrow{d}N\left(0,A^{-1}BA^{-1}\right)$$

$$A=\left(\frac{\partial^2 Q\left(\theta_0\right)}{\partial\theta\theta'}\right)\qquad B=Var\left(\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}\right)$$

Maxmimum Likelihood Estimator: $\frac{\partial Q(\theta)}{\partial\theta}=\frac{1}{n}\frac{\log L(\theta)}{\partial\theta}$. $\frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'}=\frac{1}{n}\frac{\partial^2\log L(\theta_0)}{\partial\theta\theta'}$. Note that the following information matrix for the log likelihood function: $E\frac{\partial^2\log L(\theta_0)}{\partial\theta\partial\theta'}=-E\frac{\partial\log L(\theta_0)}{\partial\theta}\frac{\partial\log L(\theta_0)}{\partial\theta'}$, which results from:

$$0=E\frac{\partial\log L\left(\theta_0\right)}{\partial\theta}\Longrightarrow E\frac{\partial L\left(\theta_0\right)}{\partial\theta}=\int\frac{\partial L\left(\theta\right)}{\partial\theta}\frac{1}{L\left(\theta\right)}L\left(\theta\right)dy=\int\frac{\partial L\left(\theta\right)}{\partial\theta}dy=\frac{\partial}{\partial\theta}\int L\left(\theta\right)dy=0.$$

whenever interchange of integration and differentiation is justified(by DOM for example). Furthermore differentiate the leftmost equality:

$$\begin{aligned}0=&\frac{\partial}{\partial\theta}E\frac{\partial\log L\left(\theta\right)}{\partial\theta}=\frac{\partial}{\partial\theta}\int\frac{\partial\log L\left(\theta\right)}{\partial\theta}L\left(\theta\right)dy\\&=\int\frac{\partial^2\log L\left(\theta\right)}{\partial\theta\partial\theta'}L\left(\theta\right)dy+\int\frac{\partial\log L\left(\theta\right)}{\partial\theta}\frac{\partial\log L\left(\theta\right)}{\partial\theta'}L\left(\theta\right)dy=E\frac{\partial^2\log L\left(\theta\right)}{\partial\theta\partial\theta'}+E\frac{\partial\log L\left(\theta\right)}{\partial\theta}\frac{\partial\log L\left(\theta\right)'}{\partial\theta}\end{aligned}$$

So that in this case we have $A=B$, and therefore $\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,A^{-1}\right)=N\left(0,\left(\lim\frac{1}{n}E\frac{\partial^2\log L(\theta)}{\partial\theta\partial\theta'}\right)^{-1}\right]$.

This is true for either iid data or heterogenuous and dependent data, as long as regularity conditions for LLN and CLT are satisfied, as long as L represents the joint density for the data. In the case where the data is iid, this would simply to $\frac{1}{n}E\frac{\partial^2 f(y;\theta)}{\partial\theta\partial\theta'}=-E\frac{\partial\log f(y;\theta)}{\partial\theta}\frac{\partial\log f(y;\theta)}{\partial\theta'}$.

However, there are times when interchanging integration and differentiation is not possible, and there will be an extra term that breaks the inequality. For example, if $y\in(\theta,\infty)$, then

$$0=\frac{\partial}{\partial\theta}\int_\theta^\infty f\left(y\right)dy=-f\left(\theta\right)+\int_\theta^\infty\frac{\partial f\left(y\right)}{\partial\theta}dy=-f\left(\theta\right)+E\frac{\partial\log f\left(y;\theta\right)}{\partial\theta}$$

which leads to $E\frac{\partial\log f(y;\theta)}{\partial\theta}=f\left(\theta\right)$, which may or may not be 0 depending on whether $f\left(\theta\right)=0$.

GMM: $Q_n\left(\theta\right)=g_n\left(\theta\right)'Wg_n\left(\theta\right)$, $g_n\left(\theta\right)=\frac{1}{n}\sum_{t=1}^n g\left(z_t,\theta\right)$. Let $G_n\left(\theta\right)=\frac{\partial g_n(\theta)}{\partial\theta}$. Because in this case $Q_n\left(\theta\right)$ is already a quadratic form in $g_n\left(\theta\right)$, it is possible to avoid taking 2nd order expansion of $Q_n\left(\theta\right)$ by taking 1st order expansion of $g_n\left(\theta\right)$, thus avoiding imposing regularity condition on $\frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'}$, Read Newey and McFadden pp2145. Write for $\theta^*\in\left[\theta_0,\hat{\theta}\right]$: $\hat{G}_n\equiv G_n\left(\hat{\theta}\right)$, $G_n^*\equiv G_n\left(\theta^*\right)$,

$G = EG_n(\theta_0)$, $\Omega = E\left(g(z,\theta_0)\, g(z,\theta_0)'\right)$:

$$\begin{aligned} 0 = \hat{G}_n' W g_n\left(\hat{\theta}\right) &= \hat{G}_n' W\left(g_n(\theta_0) + G_n^*\left(\hat{\theta}-\theta_0\right)\right) = \hat{G}_n' W g_n(\theta_0) + \hat{G}_n' W G_n^*\left(\hat{\theta}-\theta_0\right) \\ \Longrightarrow \sqrt{n}\left(\hat{\theta}-\theta_0\right) &= \left(\hat{G}_n' W G_n^*\right)^{-1} \hat{G}_n' W \sqrt{n} g_n(\theta_0) \overset{LD}{=} \left(G'WG\right)^{-1} G'W\sqrt{n} g_n(\theta_0) \\ &\overset{LD}{=} \left(G'WG\right)^{-1} G'W \times N(0,\Omega) = N\left(0, \left(G'WG\right)^{-1} G'W\Omega W\left(G'WG\right)^{-1}\right) \end{aligned}$$

1. Efficient choice of $W = \Omega^{-1}$(or $W \propto \Omega^{-1}$, in which case $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right)$.
2. When G is invertible, W is irrelevant, $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, G^{-1}\Omega G'^{-1}\right) = N\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right)$
3. Sometimes Ω is known up to a scalar, e.g. in homoscedastic regression models, $\Omega = aG$, in which case we have $G \propto \Omega$ and $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0, aG^{-1}\right)$. Consider the following examples,
4. Least square(LS): $y = x'\beta+\varepsilon$, $g(z,\theta) = x(y-x\beta)$. $G = Exx'$, $\Omega = E\varepsilon^2 xx'$. If $E\left[\varepsilon^2|x\right] = \sigma^2$, then $\Omega = \sigma^2 Exx' \Longrightarrow \Omega = \sigma^2 G$ and $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0, \sigma^2\left(Exx'\right)^{-1}\right)$. Otherwise we must use $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0, \left(Exx'\right)^{-1}\left(E\varepsilon^2 xx'\right)\left(Exx'\right)^{-1}\right)$, the so-called White's heteroscedasticity consistency standard error.
5. Weighted LS: $g(z_t,\beta) = \frac{1}{E(\varepsilon^2|x)}(y-x'\beta)$. $G = E\frac{1}{E(\varepsilon^2|x)}xx' = \Omega \Longrightarrow \sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N(0,G)$.
6. Linear 2SLS: $g(z,\beta) = z(y-x\beta)$, $G = Ezx'$, $\Omega = E\varepsilon^2 zz'$, $W = (Ezz')^{-1}$, so that $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N(0,V)$ for
$$V = \left[Exz'\left(Ezz'\right)^{-1}Ezx'\right]^{-1} Exz'\left(Ezz'\right)^{-1} E\varepsilon^2 zz'\left(Ezz'\right)^{-1} Ezx'\left[Exz'\left(Ezz'\right)^{-1}Ezx'\right]^{-1}.$$
Only when in homoscedastic models, where $E\varepsilon^2 zz' = \sigma^2 Ezz'$, the variance simplies to
$$V = \sigma^2\left[Exz'\left(Ezz'\right)^{-1}Ezx'\right]^{-1}.$$
7. Linear 3SLS: $g(z,\beta) = z(y-x\beta)$, $G = Ezx'$, $\Omega = E\varepsilon^2 zz'$, $W = \left(E\varepsilon^2 zz'\right)^{-1}$, so that
$$\sqrt{n}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N(0,V) \quad \text{for} \quad V = \sigma^2\left[Exz'\left(E\varepsilon^2 zz'\right)^{-1}Ezx'\right]^{-1}.$$
8. MLE as GMM: $g(z,\theta) = \frac{\partial \log f(z,\theta)}{\partial\theta}$, $G = -E\frac{\partial^2 \log f(z,\theta)}{\partial\theta\partial\theta'} = \Omega = E\frac{\partial \log f(z,\theta)}{\partial f(z,\theta)}'$, W is irrelevant. So that $\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{d} N\left(0, G^{-1}\right) = N(0,\Omega)$.
9. GMM again: Although there can be more moment conditions in $g(z,\theta)$ than the dimension of θ, it is always possible to take linear combinations of the moment conditions so that there are exactly the same number of $g(z,\theta)$ as the number of θ. In particular, take $h(z,\theta) = G'Wg(z,\theta)$ and use $h(z,\theta)$ as the new moment conditions defining the GMM estimate. Then
$$\hat{\theta} = argmax_\theta \left[\frac{1}{n}\sum_{t=1}^n h(z_t,\theta)\right]' \left[\frac{1}{n}\sum_{t=1}^n h(z_t,\theta)\right]$$

is asymptotically equivalent to $\hat{\theta} = argmax_\theta g_n' W g_n$. Then it is also possible to write $G = E\frac{\partial m(z,\theta)}{\partial \theta} = G'WG$, $\Omega = Eh(z,\theta)h(z,\theta)' = G'W\Omega WG$.

10. Quantile Regression as GMM: For $y = x'\beta + \mu$, $Pr(u \leq 0|x) = \tau$.

$$\hat{\beta} = argmin_\beta Q_n(\beta) \sum_{t=1}^{n} (\tau - 1(y_t \leq x_t'\beta))(y_t - x_t'\beta),$$

roughly the first order condition hold: $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\tau - 1\left(y \leq x_t'\hat{\beta}\right)\right)x_t = o_p(1)$. Therefore roughly speaking, $g(z,\beta) = (\tau - 1(y \leq x'\beta))x$, and $G = E\frac{g(z,\beta)}{\partial\beta} = -E\frac{\partial 1(y \leq x'\beta)}{\partial\beta}$, W is irrelevant. The problem here is that it is not possible to differentiate $\frac{\partial 1(y \leq x'\beta)}{\partial\beta}$. The "quick and dirty" way to get around with this problem and obtain the result is to "pretend" that we can take expectation before taking differentiation. The formal way to justify doing this is the theory of empirical process, which you could read in great detail from the Andrews Chapter 37 in the Handbook, if you are interested. Proceeding with the "quick and dirty" way,

$$G = \frac{\partial E1(y \leq x'\beta)x}{\partial\beta} = \frac{\partial ExF(y \leq x'\beta|x)}{\partial\beta} = Ex\frac{\partial F(y \leq x'\beta|x)}{\partial\beta} = Ef_y(x'\beta|x)xx' = Ef_u(0|x)xx'.$$

On the other hand, use the fact that conditional on x, $\tau - 1(y \leq x'\beta_0) = \tau - 1(u \leq 0)$ is a Bernoulli r.v. which $= \tau - 1$ w.p. τ and $= \tau$ w.p. $\tau - 1$, there is $E\left[(\tau - 1(y \leq x'\beta_0))^2|x\right] = \tau(1-\tau)$, so that

$$\Omega = Eg(z,\beta_0)g(z,\beta_0)' = EE\left[(\tau - 1(y \leq x'\beta_0))^2\right]xx' = \tau(1-\tau)Exx'.$$

The final conclusion is $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \tau(1-\tau)\left[Ef_u(0|x)xx'\right]^{-1}Exx'\left[Ef_u(0|x)xx'\right]^{-1}\right)$. In homoscedastic models where $f(0|x) = f(0)$, this simplies to $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f(0)}Exx'\right)$.

11. Consistent estimation of G and Ω:

    (a) G is the easy part, estimate by $G \doteq \frac{1}{n}\sum_{t=1}^{n}\frac{\partial g\left(z_t,\hat{\theta}\right)}{\partial\theta}$. For nonsmooth problems as quantile regression, use numerical derivative of the objective function $\frac{Q_n\left(\hat{\theta}+2h_n\right)+Q_n\left(\hat{\theta}-2h_n\right)-2Q\left(\hat{\theta}\right)}{4h_n^2}$ to approximate G. Require $h_n = o(1)$ and $\frac{1}{h_n} = o\left(\frac{1}{\sqrt{n}}\right)$. Essentially, h_n needs to go to 0 slower than $\hat{\beta}$ to β_0 in order to smooth out the noise of $\hat{\beta}$ in estimating β_0.

    (b) For stationary data, heteroscedasticity and dependence will only affect estimation of Ω. For independent data, use White's heteroscedasticity-consistent estimate of Ω, for dependent data, use Newey-West's autocorrelation-consistent estimate of Ω. For more detail, read Newey and McFadden pp 2135–2155.

Iteration and One Step Estimation: Read Amemiya pp 137-141, Newey and McFadden pp 2150–2152. Starting from an initial guess $\tilde{\theta}$, There are two ways of iteration to obtain the next round guess $\bar{\theta}$.

1. Newton-Raphson, Use quadratic approximation for $Q_n(\theta)$:

$$Q_n(\theta) \approx Q_n\left(\tilde{\theta}\right) + \frac{\partial Q\left(\tilde{\theta}\right)}{\partial\theta}'\left(\theta - \tilde{\theta}\right) + \frac{1}{2}\left(\theta - \tilde{\theta}\right)'\frac{\partial^2 Q\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\left(\theta - \tilde{\theta}\right) = 0.$$

$$\Longrightarrow \frac{\partial Q_n\left(\tilde{\theta}\right)}{\partial\theta} + \frac{\partial^2 Q_n\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\left(\bar{\theta} - \tilde{\theta}\right) = 0 \Longrightarrow \bar{\theta} = \tilde{\theta} - \left[\frac{\partial^2 Q_n\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\right]^{-1}\frac{\partial Q_n\left(\tilde{\theta}\right)}{\partial\theta}$$

2. Gauss-Newton, use linear approximation for the first-order condition, e.g. GMM:

$$Q_n(\theta) \approx \left(g_n\left(\tilde{\theta}\right) + \tilde{G}\left(\theta - \tilde{\theta}\right)\right)' W \left(g_n\left(\tilde{\theta}\right) + \tilde{G}\left(\theta - \tilde{\theta}\right)\right)'$$
$$\Longrightarrow \tilde{G}' W g_n\left(\tilde{\theta}\right) + \tilde{G}' W \tilde{G}\left(\bar{\theta} - \tilde{\theta}\right) = 0. \Longrightarrow \bar{\theta} = \tilde{\theta} - \left(\tilde{G}' W \tilde{G}\right)^{-1} \tilde{G} W g_n\left(\tilde{\theta}\right)$$

3. If the initial guess is a $\sqrt{n}$ consistent estimate, e.g. $\left(\tilde{\beta} - \beta_0\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$, then $\sqrt{n}\left(\bar{\theta} - \theta_0\right) \overset{LD}{=} \sqrt{n}\left(\hat{\theta} - \theta_0\right)$, for $\hat{\theta} = argmax_\theta Q_n(\theta)$. More iteration will not increase (first-order) asymptotic efficiency:

   (a) For Newton-Raphson:

$$\sqrt{n}\left(\bar{\theta} - \theta_0\right) = \sqrt{n}\left(\tilde{\theta} - \theta_0\right) - \left[\frac{\partial^2 Q\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\right]^{-1} \sqrt{n}\frac{\partial Q\left(\tilde{\theta}\right)}{\partial\theta}$$
$$= \sqrt{n}\left(\tilde{\theta} - \theta_0\right) - \left[\frac{\partial^2 Q\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\right]^{-1} \left[\sqrt{n}\frac{\partial Q(\theta_0)}{\partial\theta} + \left(\tilde{\theta} - \theta_0\right)\frac{\partial^2 Q(\theta^*)}{\partial\theta\partial\theta'}\right]$$
$$= \left(I - \left[\frac{\partial^2 Q\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial^2 Q(\theta^*)}{\partial\theta\partial\theta'}\right)\sqrt{n}\left(\tilde{\theta} - \theta_0\right) - \left[\frac{\partial^2 Q\left(\tilde{\theta}\right)}{\partial\theta\partial\theta'}\right]^{-1} \sqrt{n}\frac{\partial Q(\theta_0)}{\partial\theta} = o_p(1) + \sqrt{n}\left(\hat{\theta} - \theta_0\right)$$

   (b) For Gauss-Newton:

$$\sqrt{n}\left(\bar{\theta} - \theta_0\right) = \sqrt{n}\left(\tilde{\theta} - \theta_0\right) - \left(\tilde{G}' W \tilde{G}\right)^{-1} \tilde{G}' W \sqrt{n}\left[g_n(\theta_0) + G^*\left(\tilde{\theta} - \theta_0\right)\right]$$
$$= \left(I - \left(\tilde{G}' W \tilde{G}\right)^{-1} \tilde{G}' W G^*\right)\sqrt{n}\left(\tilde{\theta} - \theta_0\right) - \left(\tilde{G}' W \tilde{G}\right)^{-1} \tilde{G}' W \sqrt{n} g_n(\theta_0) = o_p(1) + \sqrt{n}\left(\tilde{\theta} - \theta_0\right)$$

Influence Function: Read Newey and McFadden pp 2142, (z_t) is called an influence function when $\sqrt{n}\left(\hat{\theta} - \theta_0\right) = \frac{1}{\sqrt{n}}\sum_{t=1}^n \phi(z_t) + o_p(1)$, and $E\phi(z_t) = 0$, $E\phi(z_t)\phi(z_t)' < \infty$. With slight abuse notation, you can think of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ distributed as a vector $\phi(z_t)$ that is normal $N(0, \Omega = E\phi\phi')$. Examples include:

1. For MLE, $\phi(z_t) = \left[-E\frac{\partial^2 \ln f(y_t,\theta_0)}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}$. Or $\left[E\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}\frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}'\right]^{-1} \frac{\partial \ln f(y_t,\theta_0)}{\partial\theta}$
2. For GMM, $\phi = -(G'WG)^{-1} G'W g(z_t,\theta_0)$, or $\phi = -\left(E\frac{\partial h}{\partial\theta}\right)^{-1} h(z_t,\theta_0)$ for $h(z_t,\theta_0) = G'W g(z_t,\theta_0)$.
3. Quantile Regression: $\phi(z_t) = [Ef(0|x)xx']^{-1}(\tau - 1(u \le 0))x_t$.

Influence function representation is particularly used for discussion of asymptotic efficiency, two step or multistep estimation, etc.

Asymptotic Efficiency: Read Amemiya pp123-pp125, pp146. Newey and McFadden pp 2162–2173.

1. Superefficient estimator: suppose θ_0 is just a scalar, there is some estimate $\hat{\theta}$ such that $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N(0, V)$ for all θ. Now define

$$\theta^* = \begin{cases} \hat{\theta} & \text{if} \quad |\hat{\theta}| \ge n^{-1/4} \\ 0 & if \quad |\hat{\theta}| < n^{-1/4} \end{cases}$$

Question: Show that if $\theta_0 = 0$, then $\sqrt{n}(\theta^* - \theta_0) \xrightarrow{d} N(0,0)$. But if $\theta_0 \neq 0$, then $\sqrt{n}(\theta^* - \theta_0) \overset{LD}{=} \sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0,V)$. θ^* does better than $\hat{\theta}$ at $\theta_0 = 0$ and does no worse than $\hat{\theta}$ at any other point. This type of behavior needs to be ruled out by regularity conditions which requires convergence of $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ to be locally uniform in θ_0, for details read the Newey(1990) article. $\hat{\theta}$ is regular if for any data generated by $\theta_n = \theta_0 + \delta/\sqrt{n}$, for $\delta \geq 0$, $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ has a limit distribution that does not depend on δ. In particular, θ^* is not regular according to this definition, you can verify that the limiting distribution depends on whether $\delta = 0$.

2. For regular estimators that have an influence function representation, indexed by τ:

$$\sqrt{n}\left(\hat{\theta}(\tau) - \theta_0\right) \overset{LD}{=} \phi(z,\tau) \sim N\left(0, E\phi(\tau)\phi(\tau)'\right),$$

a neccessary condition for $\hat{\theta}(\bar{\tau})$ being efficient than any other $\hat{\theta}(\tau)$, meaning that it has a smaller var-cov matrix, is that $Cov(\phi(z,\tau) - \phi(z,\bar{\tau}), \phi(z,\bar{\tau})) = 0$ for all τ including $\bar{\tau}$ itself.

Think about the simple picture drawn in class, the basic geometric intuition is that $\hat{\theta}(\tau) - \hat{\theta}(\bar{\tau}) \approx \phi(z,\tau) - \phi(z,\bar{\tau})$ should be "orthogonal" to $\hat{\theta}(\bar{\tau}) \approx \phi(z,\bar{\tau})$, in order for $\hat{\theta}(\bar{\tau}) \approx \phi(z,\bar{\tau})$ to have the shortest distance from the origin to the subspace spaned by the random variables of all possible regular estimators $\hat{\theta}(\tau)$. You can verify that the following are equivalent:

$$\begin{aligned} Cov(\phi(z,\tau) - \phi(z,\bar{\tau}), \phi(z,\bar{\tau})) = 0 &\Longleftrightarrow Cov(\phi(z,\tau), \phi(z,\bar{\tau})) = Var(\phi(z,\bar{\tau})) \\ &\Longleftrightarrow E\phi(z,\tau)\phi(z,\bar{\tau})' = E\phi(z,\bar{\tau})\phi(z,\bar{\tau})' \end{aligned}$$

3. If $Cov(\phi(z,\tau) - \phi(z,\bar{\tau}), \phi(z,\bar{\tau})) \neq 0$, then you can find $\phi(z,\tau^*)$ with a smaller variance than $\phi(z,\bar{\tau})$ by projecting $\phi(z,\bar{\tau})$ onto the $\phi(z,\tau) - \phi(z,\bar{\tau})$, and set $\phi(z,\tau^*)$ equal to the residual of this least square projection. See Amemiya P146 for the formula.

4. Newey's efficiency framework: Most of the estimators that can be classified into the GMM framework have an influence function with the form: $\phi(z,\tau) = D(\tau)^{-1} m(z,\tau)$.

   (a) For the class of GMM estimator indexed by $\tau = W$ the weighting matrix, given a vector $g(z,\theta_0)$, has the form above with $D(\tau) \equiv D(W) = (G'WG)$ and $m(z,\tau) \equiv m(z,W) = G'Wg(z,\theta_0)$.

   (b) Consider MLE among the class of GMM estimators with all possible $g(z,\theta_0)$ and possible W, so that τ indexes any vector of moment function having the same dimension as θ. In this case, $D(\tau) \equiv D(h) = E\frac{\partial h}{\partial \theta}$ and $m(z,\tau) = h(z_t,\theta_0)$.

5. For this particular case where $\phi(z,\tau) = D(\tau)^{-1} m(z,\tau)$, the condition $E\phi(z,\tau)\phi(z,\bar{\tau})' = E\phi(z,\bar{\tau})\phi(z,\bar{\tau})'$ for $\bar{\tau}$ indexing an efficient estimators becomes:

$$D(\tau)^{-1} Em(z,\tau)m(z,\bar{\tau}) D(\bar{\tau})^{-1} = D(\bar{\tau})^{-1} Em(z,\bar{\tau})m(z,\bar{\tau}) D(\bar{\tau})^{-1}$$

It is easy to check that if $\bar{\tau}$ is such that $D(\tau) = Em(z,\tau)m(z,\bar{\tau})$ for all τ, then both sides above are the same $D(\bar{\tau})^{-1}$. Examples:

   (a) GMM with optimal weighting matrix: Remember

$$D(\tau) = D(W) = G'WG, \quad Em(z,\tau)m(z,\bar{\tau})' = Em(z,W)m\left(z,\bar{W}\right) = G'W\Omega\bar{W}G.$$

So that you want

$$0 = D(\tau) - Em(z,\tau)\, m(z,\bar{\tau}) = G'WG - G'W\Omega\bar{W}G = G'W\left(G - \Omega\bar{W}\right)G,$$

which is satisfied by setting $\bar{W} = \Omega^{-1}$.

(b) MLE better than any GMM: For $D(\tau) = -E\frac{\partial h}{\partial\theta}$, $m(z,\tau) = h(z,\theta_0)$, the optimal choice of $\bar{h}$ is one for which $D(\tau) = -E\frac{\partial h(z,\theta_0)}{\partial\theta} = Eh(z,\theta_0)\,\bar{h}(z,\theta_0)$. You can verify that this is satisfied if $\bar{h}(z,\theta_0) = \frac{\partial \ln f(y,\theta_0)}{\partial\theta}$, the score function for MLE. This is due to the generalized information matrix equality, which says that

$$\begin{aligned} 0 &= \frac{\partial Eh(z,\theta_0)}{\partial\theta} = \frac{\partial}{\partial\theta}\int h(z,\theta)\, f(z,\theta)\, dz \\ &= \int \frac{\partial h(z,\theta)}{\partial\theta} f(z,\theta)\, dz + \int h(z,\theta)\frac{\partial \ln f(z,\theta)}{\partial\theta} f(z,\theta)\, dz = E\frac{\partial h(z,\theta_0)}{\partial\theta} + Eh(z,\theta_0)\frac{\partial \ln f(z,\theta_0)}{\partial\theta} \end{aligned}$$

(c) For the example on efficient instrument in nonlinear GMM model, look at Newey and McFadden pp 2171–2173. You can practice by looking at the previous examples of WLS, 2SLS, 3SLS in this setup.

Two Step Estimator: Read Sec 6 in Newey and McFadden, the influence function is particularly convenient, simply do Taylor expansion and plug in the influence function representation:

1. General Framework: You have a first step estimator where $\sqrt{n}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\phi(z_t) + o_p(1)$. You estimate $\hat{\theta}$ by $\frac{\partial Q_n(\hat{\theta},\hat{\gamma})}{\partial\theta} = \frac{1}{n}\sum_{t=1}^{n}\frac{q(z_t,\hat{\theta},\hat{\gamma})}{\partial\theta} = 0$. For notational convenience let $h(z,\theta,\gamma) = \frac{\partial q(z_t,\theta,\gamma)}{\partial\theta}$. Let also $H(z,\theta,\gamma) = \frac{\partial h(z,\theta,\gamma)}{\partial\theta}$ and $\Gamma(z,\theta,\gamma) = \frac{\partial h(z,\theta,\gamma)}{\partial\gamma}$. Also let $H = EH(z_t,\theta_0,\gamma_0)$, $\Gamma = E\Gamma(z,\theta_0,\gamma_0)$, $h = h(\theta_0,\gamma_0)$. Then just taylor expand in the usual way:

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sum h\left(z_t,\hat{\theta},\hat{\gamma}\right) = 0 \Longleftrightarrow \frac{1}{\sqrt{n}}\sum h(\theta_0,\hat{\gamma}) + \frac{1}{n}\sum H(\theta^*,\hat{\gamma})\sqrt{n}\left(\hat{\theta}-\theta_0\right) = 0. \Longrightarrow \\ &\sqrt{n}\left(\hat{\theta}-\theta_0\right) = \left[\frac{1}{n}\sum H(\theta^*,\hat{\gamma})\right]^{-1}\frac{1}{\sqrt{n}}\sum h(\theta_0,\hat{\gamma}) \overset{LD}{=} H^{-1}\left[\frac{1}{\sqrt{n}}\sum h(\theta_0,\gamma_0) + \frac{1}{n}\sum\Gamma(\theta_0,\gamma^*)\sqrt{n}(\hat{\gamma}-\gamma_0)\right] \\ &\overset{LD}{=} H^{-1}\left[\frac{1}{\sqrt{n}}\sum h + \Gamma\left(\frac{1}{\sqrt{n}}\phi(z_t) + o_p(1)\right)\right] \overset{LD}{=} H^{-1}\left[\frac{1}{\sqrt{n}}\sum h + \Gamma\frac{1}{\sqrt{n}}\phi(z_t)\right] \end{aligned}$$

So that $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N(0,V)$ for

$$V = H^{-1}E(h+\Gamma\phi)(h'+\phi'\Gamma')H^{-1'} = H^{-1}(Ehh' + Eh\phi'\Gamma + \Gamma E\phi h' + \Gamma E\phi\phi'\Gamma')H^{-1'}$$

2. GMM both first stage $\hat{\gamma}$ and second stage $\hat{\theta}$: Now $\phi = -M^{-1}m(z)$, for some moment condition $m(z,\gamma)$. $h(\theta,\hat{\gamma}) = G'Wg(z,\theta,\hat{\gamma})$ so that $H = G'WG$, $\Gamma = G'W\frac{\partial g}{\partial\gamma} \equiv G'WG_\gamma$ for $G_\gamma \equiv \frac{\partial g}{\partial\gamma}$. Plug these into the above general case:

$$V = (G'WG)^{-1}\left[G'W\Omega WG + G'W(Eg\phi')G_\gamma'WG + G'WG_\gamma E(\phi g')WG' + G'WG_\gamma(E\phi\phi')G_\gamma'WG\right](G'WG)^{-1}$$

If $W = I$, and G is invertible, then this simplies to

$$V = G^{-1}\left[\Omega + (Eg\phi')G_\gamma' + G_\gamma(E\phi g') + G_\gamma(E\phi\phi')G_\gamma'\right]G^{-1'}.$$

The same formula can be obtained by Newey's "stacking moment" approach, for details read sec 6 of Newey McFadden chapter.

3. Again if you have trouble differentiating $\frac{\partial g(\theta,\gamma)}{\partial \theta}$ or $\frac{\partial g(\theta,\gamma)}{\partial \gamma}$, then simply take expectation before differentiation, just replace H and Γ by $\frac{\partial E g(\theta,\gamma)}{\partial \theta}$ and $\frac{\partial g(\theta,\gamma)}{\partial \gamma}$. A good exercise to try a γ and/or θ that results from a quantile regression estimate.