Econometrics
Part I: Basic regression.

References on Bruce Hansen's textbook will be abbreviated with H. Say, H 1.1 means chapter 1.1 from Bruce Hansen's textbook.

Stochastic order symbols (H 6.18)
Asymptotic theory for maximum likelihood and M-estimators. (These notes).
Generalized Least Squares (GLS). (H 4.9)
Heteroskedasticity. White heteroscedasticity-robust standard errors (H 2.13, 4.14)

Stochastic order symbols

Recall that for deterministic sequences x_n and a_n, $n = 1, 2, 3, ...$, notations

$$x_n = o(1) \text{ and } x_n = o(a_n)$$

mean that $x_n \to 0$ and $x_n/a_n \to 0$ as $n \to \infty$, respectively. Further, notation

$$x_n = O(1)$$

means that x_n is bounded uniformly in n, that is, there exists $M > 0$ such that $|x_n| < M$ for all n. Notation $x_n = O(a_n)$ means that x_n/a_n is bounded uniformly in n.

For sequences of *random variables* x_n and a_n, $n = 1, 2, 3, ...$, we introduce similar notation. We say that

$$x_n = o_p(1)$$

(x_n is little oh-p-one) if and only if $x_n \xrightarrow{p} 0$ as $n \to \infty$, and we say that $x_n = o_p(a_n)$ if and only if $x_n/a_p \xrightarrow{p} 0$ as $n \to \infty$. For example,

$$\hat{\beta}_{OLS} - \beta = o_p(1)$$

means that $\hat{\beta}_{OLS} - \beta \xrightarrow{p} 0$, that is that $\hat{\beta}_{OLS}$ is consistent for β. We say that

$$x_n = O_p(1)$$

(x_n is large oh-p-one) if and only if for any $\varepsilon > 0$, there exists $M > 0$ such that $\limsup_{n\to\infty} \Pr(|x_n| > M) \leq \varepsilon$. We write $x_n = O_p(a_n)$ if and only if $x_n/a_n = O_p(1)$.

For example,

$$\sqrt{n}\left\|\hat{\beta}_{OLS}-\beta\right\|=O_p(1).$$

Indeed, by the asymptotic normality of $\hat{\beta}_{OLS}$,

$$\sqrt{n}\left(\hat{\beta}_{OLS}-\beta\right)\xrightarrow{d}N\left(0,\sigma^2\left[E\left(x_ix_i'\right)\right]^{-1}\right).$$

Let $a>0$ be such that $aI_k\leq\sigma^{-2}E\left(x_ix_i'\right).$ Then

$$an\left\|\hat{\beta}_{OLS}-\beta\right\|^2\leq n\left(\hat{\beta}_{OLS}-\beta\right)'\sigma^{-2}E\left(x_ix_i'\right)\left(\hat{\beta}_{OLS}-\beta\right)\xrightarrow{d}\chi^2(k)$$

Hence,

$$\limsup_{n\to\infty}\Pr\left(\sqrt{n}\left\|\hat{\beta}_{OLS}-\beta\right\|>M\right)\leq\Pr\left(\frac{\chi^2(k)}{a}>M^2\right).$$

But the latter probability can be made arbitrarily small by choosing M sufficiently large. In other words, for any $\varepsilon>0$, there exists $M>0$ such that

$$\limsup_{n\to\infty}\Pr\left(\sqrt{n}\left\|\hat{\beta}_{OLS}-\beta\right\|>M\right)\leq\varepsilon,$$

which is the same as to say that $\sqrt{n}\left\|\hat{\beta}_{OLS}-\beta\right\|=O_p(1)$. Obviously, we also can write

$$\left\|\hat{\beta}_{OLS}-\beta\right\|=O_p(n^{-1/2}).$$

Often, people simply write $\hat{\beta}_{OLS}-\beta=O_p(n^{-1/2})$ or $\sqrt{n}\left(\hat{\beta}_{OLS}-\beta\right)=O_p(1)$.

Consistency of the maximum likelihood

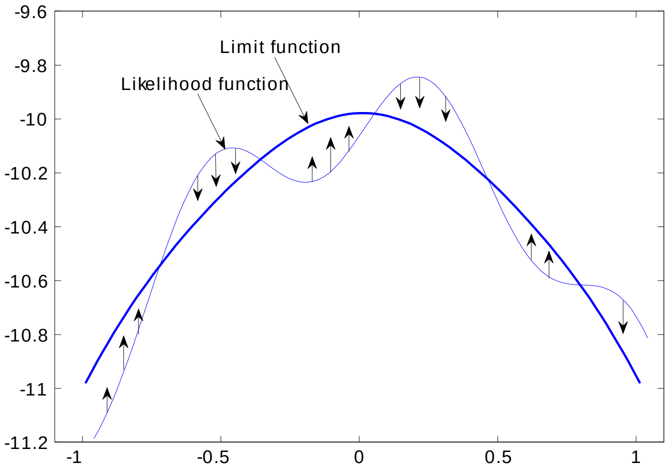
Let z_i be i.i.d. with density $f\left(z,\theta_0\right).$ Recall that the maximum likelihood estimator is

$$\hat{\theta}_{ML}=\arg\max_{\theta}\sum\nolimits_{i=1}^{n}\log f\left(z_i,\theta\right).$$

By Khinchine's LLN, for any θ,

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\log f\left(z_i,\theta\right)\xrightarrow{p}E_{\theta_0}\log f\left(z_i,\theta\right),$$

where E_{θ_0} denotes the expectation when the true value of the parameter θ is θ_0. It seems reasonable to expect that the maximizer $\hat{\theta}_{ML}$ of $\frac{1}{n}\sum\nolimits_{i=1}^{n}\log f\left(z_i,\theta\right)$ converges to the maximizer of $E_{\theta_0}\log f\left(z_i,\theta\right).$



Let us show that $E_{\theta_0} \log f(z_i, \theta)$ is maximised at the true value of parameter, θ_0. Indeed, consider the KL divergence between $f(z_i, \theta_0)$ and $f(z_i, \theta)$

$$E_{\theta_0} \log \frac{f(z_i, \theta_0)}{f(z_i, \theta)}.$$

Note that the maximizer of $E_{\theta_0} \log f(z_i, \theta)$ minimizes the KL divergence between $f(z_i, \theta_0)$ and $f(z_i, \theta)$. But KL divergence is positive unless $f(z_i, \theta_0)$ and $f(z_i, \theta)$ coincide. Indeed, by Jensen's inequality

$$E_{\theta_0} \log \frac{f(z_i, \theta_0)}{f(z_i, \theta)} = -E_{\theta_0} \log \frac{f(z_i, \theta)}{f(z_i, \theta_0)} \geq -\log E_{\theta_0} \frac{f(z_i, \theta)}{f(z_i, \theta_0)} = -\log 1 = 0.$$

Therefore, θ_0 must be a maximizer of $E_{\theta_0} \log f(z_i, \theta)$.

Remark 1 *If there exists another maximizer θ_1, we must have $f(z_i, \theta_0) = f(z_i, \theta_1)$. In such a case, we say that the parameter is not identified. In the linear regression example, $\theta = (\beta, \sigma^2)$ would be not identified if $X'X$ has rank lower than k (perfect multicollinearity).*

Remark 2 *The convergence $\frac{1}{n}\sum_{i=1}^{n} \log f(z_i, \theta) \xrightarrow{p} E_{\theta_0} \log f(z_i, \theta)$ for any θ is not sufficient for the consistency of $\hat{\theta}_{ML}$. We need a uniform convergence and "enough" curvature of $E_{\theta_0} \log f(z_i, \theta)$ at θ_0. Sufficient conditions for the consistency of the ML estimator are discussed, for example, in chapter 4 of Amemiya (1985) or chapter 5 of van der Vaart (2000). The sufficient conditions are satisfied in great many, but not all, applications.*

Asymptotic normality of the maximum likelihood

Note that $\hat{\theta}_{ML}$ can be obtained as a solution to the likelihood equation $\frac{\partial}{\partial\theta}\frac{1}{n}L(\theta;Z)=0$, where

$$L(\theta;Z)=\sum\nolimits_{i=1}^{n}\log f(z_i,\theta).$$

Let us denote $\frac{\partial}{\partial\theta}\frac{1}{n}L(\theta;Z)$ as $\Psi(\theta)$. Note that $\Psi\left(\hat{\theta}_{ML}\right)=0$. Assuming consistency, $\hat{\theta}_{ML}\xrightarrow{p}\theta_0$, it makes sense to expand $\Psi\left(\hat{\theta}_{ML}\right)$ in a Taylor series around θ_0

$$0=\Psi(\theta_0)+\left(\hat{\theta}_{ML}-\theta_0\right)\Psi'(\theta_0)+\frac{1}{2}\left(\hat{\theta}_{ML}-\theta_0\right)^2\Psi''\left(\tilde{\theta}\right),$$

where $\tilde{\theta}\in\left[\theta_0,\hat{\theta}_{ML}\right]$. Therefore, in cases when θ is scalar,

$$\sqrt{n}\left(\hat{\theta}_{ML}-\theta_0\right)=\frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0)+\left(\hat{\theta}_{ML}-\theta_0\right)\Psi''\left(\tilde{\theta}\right)/2}$$

But under the random sampling assumption

$$\Psi(\theta_0)=\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(z_i,\theta)\bigg|_{\theta=\theta_0}\xrightarrow{p}\frac{\partial}{\partial\theta}E_{\theta_0}\log f(z_i,\theta)\bigg|_{\theta=\theta_0}=0,$$

and by the Lindeberg-Lévy CLT,

$$-\sqrt{n}\Psi(\theta_0)\xrightarrow{d}N\left(0,Var\left(\frac{\partial}{\partial\theta}\log f(z_i,\theta_0)\right)\right)=N\left(0,\frac{1}{n}\mathcal{I}(\theta_0)\right)$$

where $\mathcal{I}(\theta_0)$ is the Fisher information.

Next, by Khinchine's LLN

$$\Psi'(\theta_0)\xrightarrow{p}\frac{\partial^2}{\partial\theta^2}E_{\theta_0}\log f(z_i,\theta)\bigg|_{\theta=\theta_0}$$

Finally, $\left(\hat{\theta}_{ML}-\theta_0\right)\Psi''\left(\tilde{\theta}\right)=o_p(1)$ because $\hat{\theta}_{ML}-\theta_0=o_p(1)$ and $\Psi''\left(\tilde{\theta}\right)$ converges to a finite constant (see Amemiya (1985, Ch. 4) for sufficient conditions for such a convergence). Therefore, by the Continuous Mapping Theorem

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_{ML}-\theta_0\right) &= \frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0)+\left(\hat{\theta}_{ML}-\theta_0\right)\Psi''\left(\tilde{\theta}\right)/2}\\ &\xrightarrow{d}\frac{N\left(0,\frac{1}{n}\mathcal{I}(\theta_0)\right)}{E_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(z_i,\theta_0)}.\end{aligned}$$

Using the information equality (see Session 2), we get

$$\sqrt{n}\left(\hat{\theta}_{ML}-\theta_0\right)\xrightarrow{d}\frac{N\left(0,\frac{1}{n}\mathcal{I}\left(\theta_0\right)\right)}{-\frac{1}{n}\mathcal{I}\left(\theta_0\right)}=N\left(0,n\mathcal{I}\left(\theta_0\right)^{-1}\right).$$

In other words, in large samples, $\hat{\theta}_{ML}$ is approximately normally distributed with mean θ_0 and variance equal to the inverse of the Fisher information. This straightforwardly generalizes to vector θ.

Asymptotic efficiency of the maximum likelihood.

Recall that according to the Cramér -Rao result, any unbiased estimator of θ_0 has variance no smaller than the inverse of the Fisher information. Since, as we have just seen, $\hat{\theta}_{ML}$ is asymptotically unbiased and achieves the CR lower bound for the variance, we might conclude that $\hat{\theta}_{ML}$ is asymptotically efficient. This turns out to be true, but only if some weird estimators (called irregular) are ruled out.

Hodges' estimator (the weird one). Define the following estimator

$$\hat{\theta}_H=\left\{\begin{array}{cc}\hat{\theta}_{ML} & \text{if } \left|\hat{\theta}_{ML}\right|\geq n^{-1/4}\\ 0 & \text{if } \left|\hat{\theta}_{ML}\right|<n^{-1/4}\end{array}\right\}.$$

Whatever the true value of parameter is, as long as it is not zero, Hodges' estimator is asymptotically equivalent to the ML estimator because the latter would converge to the true value of parameter, and hence, $\left|\hat{\theta}_{ML}\right|\geq n^{-1/4}$ will be true asymptotically, so that $\hat{\theta}_H=\hat{\theta}_{ML}$. On the other hand, if the true value of the parameter is zero, then Hodges' estimator clearly improves over $\hat{\theta}_{ML}$ because $\left|\hat{\theta}_{ML}\right|=O_p\left(n^{-1/2}\right)=o_p(n^{-1/4})$, and $\hat{\theta}_H=0$ for sufficiently large n, delivering the true value of the parameter exactly (with zero variance).

It turns out that in finite samples, Hodges' estimator behaves poorly for θ close to but not exactly equal to zero. Asymptotically, this is reflected in its erratic behavior when the true value of parameter is drifting towards zero so that $\theta=h/\sqrt{n}$ for some fixed h. For such sequences of θ, we have

$$\sqrt{n}\left(\hat{\theta}_H-\theta\right)=\sqrt{n}\left(\hat{\theta}_H-h/\sqrt{n}\right)\rightarrow -h.$$

Regular estimators would have the same asymptotic distribution for any value of $h/\sqrt{n}$ (a small change in the parameter should not change the distribution of the estimator too much). It turns out that the ML estimator is asymptotically efficient (has the smallest asymptotic variance) among all regular estimators. Interested students may consult chapter 8 of van der Vaart (2000) for details.

Asymptotic theory for M-estimators

Maximum likelihood is the leading example of the so-called M-estimators that are obtained by maximizing averages $\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)$ w.r.t. θ, where g are some known functions. Another example is the nonlinear least squares (NLS) estimator of the parameter θ_0 in the model

$$y_i = h(x_i,\theta_0) + \varepsilon_i, \text{ where } E(\varepsilon_i|x_i) = 0$$

and $h(x_i,\theta)$ is a parametric family of nonlinear functions. The NLS maximizes the negative of the sum of squared residuals:

$$-\sum_{i=1}^{n} (y_i - h(x_i,\theta))^2 .$$

Here, we may define z_i as (y_i,x_i) and $g(z_i,\theta)$ as $(y_i - h(x_i,\theta))^2$.

Asymptotic theory for M-estimators can be developed similarly to that of the ML estimator. By LLN, $\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta) \to E_{\theta_0} g(z_i,\theta)$. Typically, $E_{\theta_0} g(z_i,\theta)$ would be maximized at $\theta = \theta_0$ (for NLS, this would follow from the fact that the conditional expectation is minimizing the expected square error). Hence, if the convergence to $E_{\theta_0} g(z_i,\theta)$ is uniform in θ and $E_{\theta_0} g(z_i,\theta)$ has a sufficient amount of curvature at $\theta = \theta_0$, we would expect the M-estimator $\hat{\theta}$ to be consistent for θ_0.

As to the asymptotic normality, note that $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} g\left(z_i,\hat{\theta}\right) = 0$. Expanding $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} g\left(z_i,\hat{\theta}\right)$ at θ_0, we get

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} g(z_i,\theta_0) + \left(\hat{\theta}-\theta_0\right)\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} g(z_i,\theta_0) + O_P\left(\left(\hat{\theta}-\theta_0\right)^2\right). \quad (1)$$

Expansion (1) yields (by the same line of arguments as in the derivation of the asymptotic distribution of the ML)

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} \frac{N\left(0, Var\left(\frac{\partial}{\partial\theta} g(z_i,\theta_0)\right)\right)}{E\frac{\partial^2}{\partial\theta^2} g(z_i,\theta_0)},$$

or more compactly,

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, H^{-1}GH^{-1}\right),$$

where $G = Var\left(\frac{\partial}{\partial\theta} g(z_i, \theta_0)\right)$ and $H = E\frac{\partial^2}{\partial\theta\partial\theta'} g(z_i, \theta_0)$. For NLS, if ε_i is independent from x_i and has variance σ^2, it is not difficult to show that

$$\begin{aligned} G &= 4\sigma^2 E\left[\frac{\partial}{\partial\theta} h(x_i, \theta_0) \frac{\partial}{\partial\theta'} h(x_i, \theta_0)\right] \text{ and} \\ H &= -2E\left[\frac{\partial}{\partial\theta} h(x_i, \theta_0) \frac{\partial}{\partial\theta'} h(x_i, \theta_0)\right] \end{aligned}$$

so that

$$\sqrt{n}\left(\hat{\theta}_{NLS} - \theta_0\right) \xrightarrow{d} N\left(0, \sigma^2 \left(E\left[\frac{\partial}{\partial\theta} h(x_i, \theta_0) \frac{\partial}{\partial\theta'} h(x_i, \theta_0)\right]\right)^{-1}\right).$$

`The following material (up to small sample GLS) is not examinable.`

Importantly, the asymptotic normality of $\hat{\theta}$ can be derived even if $g(z, \theta)$ is not differentiable with respect to θ for some z_i.

Theorem 3 *(Theorem 5.23 in van der Vaart (2000)) Suppose that $g(z, \theta)$ is differentiable at θ_0 for $z \in \Omega$ with $\Pr(Z \in \Omega) = 1$, but not necessarily everywhere. Further, suppose that there exists $\bar{g}(z)$ with $E(\bar{g}(Z))^2 < \infty$ such that for every θ_1, θ_2*

$$|g(z, \theta_1) - g(z, \theta_2)| \leq \bar{g}(z) |\theta_1 - \theta_2|. \tag{2}$$

Moreover, assume that $Eg(Z, \theta)$ considered as a function of θ admits a second-order Taylor expansion at a point of maximum θ_0 with second derivative V_0. Then, if $\hat{\theta} \xrightarrow{p} \theta_0$,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = -V_0^{-1} \frac{1}{\sqrt{n}} \sum\nolimits_{i=1}^{n} \frac{\partial}{\partial\theta} g(z_i, \theta_0) + o_p(1),$$

where $o_p(1)$ denotes a random quantity that converges in probability to zero as $n \longrightarrow \infty$.

Example 4 *ML estimator of mean of the double exponential distribution.*

A double exponential distribution has density $f(z, \theta_0) = \frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|z - \theta_0|\right\}$, mean θ_0, and cdf

$$F(z, \theta_0) = \left\{ \begin{array}{cl} \frac{1}{2} e^{\sqrt{2}(\theta_0 - z)} & \text{if } z \leq \theta_0 \\ 1 - \frac{1}{2} e^{\sqrt{2}(\theta_0 - z)} & \text{if } z > \theta_0 \end{array} \right. .$$

Asymmetric versions of this distributions are used, for example, to model sizes of jumps in the stochastic process describing financial asset prices. What is the ML estimator of θ_0 and what is its asymptotic distribution? We have

$$\sum\nolimits_{i=1}^{n} \log f\left(z_i,\theta\right) = -\frac{n}{2}\log 2 - \sum\nolimits_{i=1}^{n} \sqrt{2}\left|z_i-\theta\right|.$$

This is maximized at $\hat{\theta}_{ML}$ which equals the sample median (unique with probability 1 if we assume that n is odd). Unfortunately, $\log f\left(z,\theta\right) = -\frac{1}{2}\log 2 - \sqrt{2}\left|z-\theta\right|$ is not everywhere differentiable with respect to θ. The good news is that the differentiability fails only in one point, namely $\theta = z$, so $\log f\left(z,\theta\right)$ is differentiable with probability one:

$$\frac{1}{\sqrt{2}}\frac{\partial}{\partial\theta}\log f\left(z,\theta\right) = 1\left\{\theta > z\right\} - 1\left\{\theta < z\right\}.$$

The ML estimation of the double exponential mean can be set in the framework of the above theorem by taking $g\left(z,\theta\right) = \log f\left(z,\theta\right) + \sqrt{2}\left|z\right| + \frac{1}{2}\log 2$ (the addition of $\sqrt{2}\left|z\right| + \frac{1}{2}\log 2$ simply shifts $\log f\left(z,\theta\right)$ as a function of θ but makes it bounded as a function of z, which is convenient). Obviously, (2) is satisfied with $\bar{g}\left(z\right) \equiv \sqrt{2}$. Furthermore,

$$E_{\theta_0} g\left(Z,\theta\right) = \sqrt{2}\int\left(\left|z\right| - \left|z-\theta\right|\right) f\left(z,\theta_0\right) dz.$$

Using integration by parts, we get, for $\theta < 0$,

$$\begin{aligned}\frac{1}{\sqrt{2}}E_{\theta_0} g\left(Z,\theta\right) &= \left|\theta\right| F\left(\theta,\theta_0\right) + \left(-\left|\theta\right| F\left(0,\theta_0\right) - \left|\theta\right| F\left(\theta,\theta_0\right)\right) + 2\int_{\theta}^{0} F\left(z,\theta_0\right) dz \\ +\left|\theta\right|\left(F\left(0,\theta_0\right) - 1\right) &= \theta + 2\int_{\theta}^{0} F\left(z,\theta_0\right) dz\end{aligned}$$

Similarly, for $\theta > 0$,

$$\begin{aligned}\frac{1}{\sqrt{2}}E_{\theta_0} g\left(Z,\theta\right) &= -\left|\theta\right| F\left(0,\theta_0\right) + \left(\left|\theta\right| F\left(\theta,\theta_0\right) + \left|\theta\right| F\left(0,\theta_0\right)\right) - 2\int_{0}^{\theta} F\left(z,\theta_0\right) dz \\ -\left|\theta\right|\left(F\left(\theta,\theta_0\right) - 1\right) &= \theta - 2\int_{0}^{\theta} F\left(z,\theta_0\right) dz\end{aligned}$$

Thus overall, we have

$$E_{\theta_0} g\left(Z,\theta\right) = \sqrt{2}\left(\theta - 2\int_{0}^{\theta} F\left(z,\theta_0\right) dz\right)$$

This has a second order Taylor approximation at $\theta = \theta_0$. The first derivative is

$$\sqrt{2}\left(1 - 2F\left(\theta_0, \theta_0\right)\right) = 0$$

and the second is

$$-2\sqrt{2} f\left(\theta_0, \theta_0\right) = -2.$$

Hence,

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta} - \theta_0\right) &= \frac{1}{2}\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{\partial}{\partial\theta} g\left(z_i, \theta_0\right) + o_p(1) \\
&= \frac{1}{\sqrt{2}}\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \left[1\left\{\theta_0 > z_i\right\} - 1\left\{\theta_0 < z_i\right\}\right] + o_p(1).
\end{aligned}$$

On the other hand,

$$Var\left(1\left\{\theta_0 > z_i\right\} - 1\left\{\theta_0 < z_i\right\}\right) = E\left(1\left\{\theta_0 > z_i\right\} - 1\left\{\theta_0 < z_i\right\}\right)^2 = 1$$

Therefore, by CLT,

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \left[1\left\{\theta_0 > z_i\right\} - 1\left\{\theta_0 < z_i\right\}\right] \xrightarrow{d} N\left(0, 1\right)$$

and

$$\frac{1}{\sqrt{2}}\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \left[1\left\{\theta_0 > z_i\right\} - 1\left\{\theta_0 < z_i\right\}\right] \xrightarrow{d} \frac{1}{\sqrt{2}} N\left(0, 1\right) = N\left(0, \frac{1}{2}\right).$$

Thus, finally, by Slutsky's lemma,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{2}\right).$$

Interestingly and importantly, the sample median $\hat{\theta}$ has asymptotic variance $1/2$, whereas the sample mean has asymptotic variance 1. Here the sample mean is less efficient than the sample median, the ML estimator. This assumes that the data comes from the double exponential distribution. If it does not, then the median may even be an inconsistent estimator of the mean, for example, when the distribution is asymmetric. This highlights an important observation that, although ML is an efficient estimator, it is not robust with respect to model misspecifications.

Small Sample GLS

Recall GM3: $Var\left(Y|X\right) = E\left(\varepsilon\varepsilon'|X\right) = \sigma^2 I$. This assumption was not necessary for unbiasedness, but it was crucial to OLS being 'best'. We now replace this by a more general GM3' assumption, which just defines the variance

(GM3') $Var\left(Y|X\right) = \Omega$ positive definite.

Under (GM3'), the diagonal elements are not necessarily equal, nor are the covariances necessarily zero.

We now have the following,

$$Var\left(\hat{\beta}_{OLS}|X\right) = \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}.$$

Under (GM3'), OLS may no longer provide the minimum variance linear unbiased estimator (BLUE).

There exists for any positive definite symmetric matrix Ω, a matrix , P, found using the spectral decomposition of Ω, such that

$$\begin{aligned} P\Omega P' &= I \text{ and} \\ P'P &= \Omega^{-1} \end{aligned}$$

P is often denoted $\Omega^{-1/2}$. The matrix P is used to rotate the variables as follows:

$$PY = PX\beta + P\varepsilon$$

which can be compactly written as

$$\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}$$

With the variables transformed by P, the original GM conditions, particularly with regard to variance, apply

(GM3) $Var\left(\tilde{Y}|\tilde{X}\right) = Var\left(\tilde{\varepsilon}|\tilde{X}\right) = I.$

Indeed,

$$Var\left(\tilde{\varepsilon}|\tilde{X}\right) = E\left(\tilde{\varepsilon}\tilde{\varepsilon}'|\tilde{X}\right) - E\left(\tilde{\varepsilon}|\tilde{X}\right)\left(E\left(\tilde{\varepsilon}|\tilde{X}\right)\right)'.$$

But

$$E\left(\tilde{\varepsilon}|\tilde{X}\right) = E\left(E\left(\tilde{\varepsilon}|\tilde{X}, X\right)|\tilde{X}\right) = E\left(E\left(\tilde{\varepsilon}|X\right)|\tilde{X}\right) = E\left(PE\left(\varepsilon|X\right)|\tilde{X}\right) = E\left(P \times 0|\tilde{X}\right) = 0.$$

Similarly,

$$\begin{aligned}
E\left(\tilde{\varepsilon}\tilde{\varepsilon}'|\tilde{X}\right) &= E\left(E\left(\tilde{\varepsilon}\tilde{\varepsilon}'|\tilde{X},X\right)|\tilde{X}\right) = E\left(E\left(\tilde{\varepsilon}\tilde{\varepsilon}'|X\right)|\tilde{X}\right) = E\left(PE\left(\varepsilon\varepsilon'|X\right)P'|\tilde{X}\right) \\
&= E\left(P\Omega P'|\tilde{X}\right) = E\left(I|\tilde{X}\right) = I.
\end{aligned}$$

Now, the estimator is as follows:

$$\hat{\beta}_{GLS} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{Y} = \left(X'P'PX\right)^{-1}X'P'PY = \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y.$$

Theorem 5 *Under GM1-GM2-GM3', GLS is the Best Linear Unbiased Estimator (BLUE).*

The variance of $\hat{\beta}_{GLS}$ has the following simple form:

$$\begin{aligned}
Var\left(\hat{\beta}_{GLS}|X\right) &= Var\left(\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y|X\right) \\
&= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Var\left(Y|X\right)\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1} \\
&= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1} \\
&= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1} \\
&= \left(X'\Omega^{-1}X\right)^{-1}.
\end{aligned}$$

Now, consider the special case of $\Omega = \sigma^2 I$.

$$\begin{aligned}
\hat{\beta}_{GLS} &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y = \left(X'\sigma^{-2}IX\right)^{-1}X'\sigma^{-2}IY = \sigma^2\left(X'X\right)^{-1}\sigma^{-2}X'Y \\
&= \left(X'X\right)^{-1}X'Y = \hat{\beta}_{OLS}
\end{aligned}$$

$$Var\left(\hat{\beta}_{GLS}|X\right) = \left(X'\sigma^{-2}IX\right)^{-1} = \sigma^2\left(X'X\right)^{-1}$$

so we see that OLS is just a special case of GLS.

Notes on GLS

(1) So far we have assumed that Ω is known, when typically it is not. However, knowing Ω up to scale is sufficient to estimate. For example, let's use $\alpha\Omega$ for Ω :

$$\begin{aligned}
\tilde{\beta}_{GLS} &= \left(X'\left(\alpha\Omega\right)^{-1}X\right)^{-1}X'\left(\alpha\Omega\right)^{-1}Y \\
&= \alpha\left(X'\Omega^{-1}X\right)^{-1}\alpha^{-1}X'\Omega^{-1}Y \\
&= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y = \hat{\beta}_{GLS}
\end{aligned}$$

Still, estimation of $Var\left(\hat{\beta}_{GLS}|X\right)$ will be a problem if Ω is only known up to scale.

(2) Note also that the R^2 of the original equation loses meaning after the transformation to GLS

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\left(Y - \bar{Y}\right)'\left(Y - \bar{Y}\right)}$$

$$\begin{aligned}
\hat{\varepsilon}_{OLS} &= Y - X\hat{\beta}_{OLS} \text{ where } \hat{\beta}_{OLS} = \arg\min_{\hat{\beta}} \hat{\varepsilon}'\hat{\varepsilon} \rightarrow \text{ smallest } \hat{\varepsilon}'\hat{\varepsilon} \\
\hat{\varepsilon}_{GLS} &= Y - X\hat{\beta}_{GLS} \text{ where } \hat{\beta}_{GLS} = \arg\min_{\hat{\beta}} \hat{\varepsilon}'\Omega^{-1}\hat{\varepsilon} \rightarrow \text{ smallest } \hat{\varepsilon}'\Omega^{-1}\hat{\varepsilon}
\end{aligned}$$

This means that $RSS_{OLS} > RSS_{GLS}$, even though the GLS estimator is ‘best’.

(3) Under GM 1-2-3’, OLS remains unbiased BUT it’s usual variance estimator cannot be used for inference. Usually, we need to know Ω to even perform GLS, while without it, we can still perform OLS and get unbiased, if inefficient, estimators. However, without Ω, we are unable to directly calculate the variance of OLS using the formula

$$Var\left(\hat{\beta}_{OLS}|X\right) = \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}.$$

With some work, we will be able to get around this problem in particular leading cases.

So we are left with two directions to pursue:

- FGLS (Feasible GLS) If we have an estimate $\hat{\Omega}$ of Ω, then we can estimate β by
$$\hat{\beta}_{FGLS} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}Y.$$
  Then, we can explore the properties of $\hat{\beta}_{FGLS}$, which will rely on having a ‘good’ estimate $\hat{\Omega}$ of Ω.

- Perform OLS and ‘fix up’ the standard error estimates. The key here is to estimate $X'\Omega X$ for unknown Ω which leads to an estimate of $Var\left(\hat{\beta}_{OLS}|X\right)$.

Heteroskedasticity, GLS

Now we consider the case of conditional heteroskedasticity. While the off-diagonal terms of the variance-covariance matrix Ω are still zero, the diagonal terms are not necessarily equal. As before, if Ω is known, we can use it to rotate, or 'reweight' the variables, and perform GLS.

$$\begin{aligned}
\Omega &= \begin{bmatrix} E\left(\varepsilon_1^2|X\right) & 0 & \cdots & 0 \\ 0 & E\left(\varepsilon_2^2|X\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E\left(\varepsilon_n^2|X\right) \end{bmatrix}, \\
\Omega^{-1} &= \begin{bmatrix} \frac{1}{E\left(\varepsilon_1^2|X\right)} & 0 & \cdots & 0 \\ 0 & \frac{1}{E\left(\varepsilon_2^2|X\right)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{E(\varepsilon_n^2|X)} \end{bmatrix}, \\
\Omega^{-1/2} &= \begin{bmatrix} \frac{1}{\sqrt{E\left(\varepsilon_1^2|X\right)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{E\left(\varepsilon_2^2|X\right)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{E(\varepsilon_n^2|X)}} \end{bmatrix}
\end{aligned}$$

The GLS with heteroskedasticity would proceed by running OLS on the $\Omega^{-1/2}$ transformed variables. Let $\tilde{Y} = \Omega^{-1/2}Y$ and $\tilde{X} = \Omega^{-1/2}X$, i.e.

$$\tilde{Y}_i = \frac{Y_i}{\sqrt{E\left(\varepsilon_i^2|X\right)}} \text{ and } \tilde{x}_i = \frac{x_i}{\sqrt{E\left(\varepsilon_i^2|X\right)}}.$$

For a diagonal matrix Ω, the transformation essentially reweights the contribution of each data point.

Let us check the asymptotic distribution of $\hat{\beta}_{GLS}$. We will assume that (Y_i, x_i) are i.i.d. draws from their joint distribution. In such a case,

$$E\left(\varepsilon_i^2|X\right) = E\left(\varepsilon_i^2|x_i\right).$$

We have

$$\sqrt{n}\left(\hat{\beta}_{GLS} - \beta\right) = \left(\frac{1}{n}\sum \frac{x_i x_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)^{-1} \frac{1}{\sqrt{n}}\sum \frac{x_i \varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}.$$

By LLN,

$$\frac{1}{n}\sum \frac{x_i x_i'}{E\left(\varepsilon_i^2|x_i\right)} \xrightarrow{p} E\frac{x_i x_i'}{E\left(\varepsilon_i^2|x_i\right)}$$

and, as long as the latter matrix is not degenerate, by CMT,

$$\left(\frac{1}{n}\sum\frac{x_ix_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)^{-1}\xrightarrow{p}\left(E\frac{x_ix_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)^{-1}$$

Further, by CLT

$$\frac{1}{\sqrt{n}}\sum\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\xrightarrow{d}N\left(E\left(\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\right),Var\left(\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\right)\right),$$

where

$$E\left(\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\right)=E\left(E\left(\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}|x_i\right)\right)=E\left(\frac{x_iE\left(\varepsilon_i|x_i\right)}{E\left(\varepsilon_i^2|x_i\right)}\right)=0$$

and

$$\begin{aligned}Var\left(\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\right) &= E\left(\frac{x_ix_i'\varepsilon_i^2}{\left(E\left(\varepsilon_i^2|x_i\right)\right)^2}\right)=E\left(E\left(\frac{x_ix_i'\varepsilon_i^2}{\left(E\left(\varepsilon_i^2|x_i\right)\right)^2}|x_i\right)\right)\\ &= E\left(\frac{x_ix_i'E\left(\varepsilon_i^2|x_i\right)}{\left(E\left(\varepsilon_i^2|x_i\right)\right)^2}\right)=E\left(\frac{x_ix_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)\end{aligned}$$

so that

$$\frac{1}{\sqrt{n}}\sum\frac{x_i\varepsilon_i}{E\left(\varepsilon_i^2|x_i\right)}\xrightarrow{d}N\left(0,E\left(\frac{x_ix_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)\right)$$

and by Slutsky's lemma,

$$\sqrt{n}\left(\hat{\beta}_{GLS}-\beta\right)\xrightarrow{d}N\left(0,\left(E\frac{x_ix_i'}{E\left(\varepsilon_i^2|x_i\right)}\right)^{-1}\right)$$

Heteroskedasticity, FGLS

When Ω is not known, we might be able to perform FGLS. To do this, we assume some parametric form (hopefully based on economic theory):

$$E\left(\varepsilon_i^2|x_i\right)=h\left(x_i\right)'\alpha.$$

Perform the following calculations to find $\hat{\beta}_{FGLS}$:

(1) Perform OLS and determine $\hat{\varepsilon}_i=Y_i-x_i'\hat{\beta}_{OLS}$

(2) Regress $\hat{\varepsilon}_i^2$ on constant and $h(x_i)$ to estimate α and subsequently Ω

(3) Compute $\hat{\beta}_{FGLS}$

$$\begin{aligned}\hat{\beta}_{FGLS} &= \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1}Y \\ &= \left(\sum\nolimits_{i=1}^{n} \frac{x_i x_i'}{h\left(x_i\right)'\hat{\alpha}}\right)^{-1} \left(\sum\nolimits_{i=1}^{n} \frac{x_i Y_i}{h\left(x_i\right)'\hat{\alpha}}\right).\end{aligned}$$

What's the limiting distribution of $\hat{\beta}_{FGLS}$?

$$\sqrt{n}\left(\hat{\beta}_{FGLS} - \beta\right) = \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{x_i x_i'}{h\left(x_i\right)'\hat{\alpha}}\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i}{h\left(x_i\right)'\hat{\alpha}}\right) \tag{3}$$

Note that

$$\frac{1}{h\left(x_i\right)'\hat{\alpha}} = \frac{1}{h\left(x_i\right)'\alpha} - \frac{h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)}{\left(h\left(x_i\right)'\alpha\right)^2} + \frac{\left(h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)\right)^2}{\left(h\left(x_i\right)'\bar{\alpha}\right)^3}$$

where $\bar{\alpha}$ is a vector in between $\hat{\alpha}$ and α. Using this, we have for the second term in (3)

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i}{h\left(x_i\right)'\hat{\alpha}} &= \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i}{h\left(x_i\right)'\alpha} - \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)}{\left(h\left(x_i\right)'\alpha\right)^2} \\ &\quad + \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i \left(h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)\right)^2}{\left(h\left(x_i\right)'\bar{\alpha}\right)^3}\end{aligned} \tag{4}$$

For the first component on the right hand side, by CLT

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i}{h\left(x_i\right)'\alpha} \xrightarrow{d} N\left(0, E\frac{x_i x_i'}{h\left(x_i\right)'\alpha}\right). \tag{5}$$

For the second component,

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)}{\left(h\left(x_i\right)'\alpha\right)^2} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i h\left(x_i\right)'}{\left(h\left(x_i\right)'\alpha\right)^2}\sqrt{n}\left(\hat{\alpha}-\alpha\right).$$

But, by LLN,

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{x_i \varepsilon_i h\left(x_i\right)'}{\left(h\left(x_i\right)'\alpha\right)^2} \xrightarrow{p} E\left(\frac{x_i \varepsilon_i h\left(x_i\right)'}{\left(h\left(x_i\right)'\alpha\right)^2}\right) = 0$$

Assuming that $\sqrt{n}\left(\hat{\alpha}-\alpha\right)$ converges to some distribution, we have by Slutsky's lemma

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{x_i\varepsilon_i h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)}{\left(h\left(x_i\right)'\alpha\right)^2}\xrightarrow{p}0 \tag{6}$$

Finally, for the last term in (4),

$$\begin{aligned}\left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{x_i\varepsilon_i\left(h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)\right)^2}{\left(h\left(x_i\right)'\bar{\alpha}\right)^3}\right| &= \left|\frac{1}{\sqrt{n}}\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{x_i\varepsilon_i\left(h\left(x_i\right)'\sqrt{n}\left(\hat{\alpha}-\alpha\right)\right)^2}{\left(h\left(x_i\right)'\bar{\alpha}\right)^3}\right| \\ &\leq \frac{1}{\sqrt{n}}\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\left|x_i\varepsilon_i\right|\left\|h\left(x_i\right)\right\|^2}{\left|h\left(x_i\right)'\bar{\alpha}\right|^3}\left\|\sqrt{n}\left(\hat{\alpha}-\alpha\right)\right\|^2\end{aligned}$$

Assuming that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\left|x_i\varepsilon_i\right|\left\|h\left(x_i\right)\right\|^2}{\left|h\left(x_i\right)'\bar{\alpha}\right|^3}$$

converges in probability to a finite vector, and $\sqrt{n}\left(\hat{\alpha}-\alpha\right)$ converges to some distribution, we see that

$$\frac{1}{\sqrt{n}}\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\left|x_i\varepsilon_i\right|\left\|h\left(x_i\right)\right\|^2}{\left|h\left(x_i\right)'\bar{\alpha}\right|^3}\left\|\sqrt{n}\left(\hat{\alpha}-\alpha\right)\right\|^2=\frac{1}{\sqrt{n}}O_P\left(1\right)\xrightarrow{p}0.$$

So

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{x_i\varepsilon_i\left(h\left(x_i\right)'\left(\hat{\alpha}-\alpha\right)\right)^2}{\left(h\left(x_i\right)'\bar{\alpha}\right)^3}\xrightarrow{p}0 \tag{7}$$

Using (5), (6), and (7) in (4) and applying Slutsky's lemma, we get

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{x_i\varepsilon_i}{h\left(x_i\right)'\hat{\alpha}}\xrightarrow{d}N\left(0,E\frac{x_ix_i'}{h\left(x_i\right)'\alpha}\right).$$

Similarly, we can show that

$$\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{x_ix_i'}{h\left(x_i\right)'\hat{\alpha}}\right)^{-1}\xrightarrow{p}\left(E\frac{x_ix_i'}{h\left(x_i\right)'\alpha}\right)^{-1}.$$

Hence, using Slutsky's lemma once more, this time in (3),

$$\sqrt{n}\left(\hat{\beta}_{FGLS}-\beta\right)\xrightarrow{d}\left(E\frac{x_ix_i'}{h\left(x_i\right)'\alpha}\right)^{-1}N\left(0,E\frac{x_ix_i'}{h\left(x_i\right)'\alpha}\right)=N\left(0,\left(E\frac{x_ix_i'}{h\left(x_i\right)'\alpha}\right)^{-1}\right),$$

which is the same asymptotic distribution as that of GLS. Moreover, we can make a stronger claim that $\hat{\beta}_{FGLS}$ and $\hat{\beta}_{GLS}$ are asymptotically equivalent in the sense that

$$\sqrt{n}\left(\hat{\beta}_{FGLS}-\hat{\beta}_{GLS}\right)\xrightarrow{p}0.$$

Heteroskedasticity, OLS

Recall our large sample assumptions:

(OLS0) (y_i, x_i) is an i.i.d. sequence

(OLS1) $E\left(x_i x_i'\right)<\infty$ and is non-singular

(OLS2) $E\left(\varepsilon_i|x_i\right)=0$

(OLS3) $Var\left(Y_i|x_i\right)=Var\left(\varepsilon_i|x_i\right)=\sigma^2$

Consider weaker assumptions

(OLS2') $E\left(\varepsilon_i x_i\right)=0$

(OLS3') $Var\left(\varepsilon_i x_i\right)=V<\infty$ and is non-singular

Theorem 6 *Under OLS0-1-2'-3'*
(i) $\hat{\beta}_{OLS}\xrightarrow{p}\beta$ *(OLS is consistent)*
(ii) $\sqrt{n}\left(\hat{\beta}_{OLS}-\beta\right)\xrightarrow{d}N\left(0,\left(E\left(x_i x_i'\right)\right)^{-1}V\left(E\left(x_i x_i'\right)\right)^{-1}\right)$. *If* $V=\sigma^2 E\left(x_i x_i'\right)$ *as in the homoskedasticity case, this reduces to* $N\left(0,\sigma^2\left(E\left(x_i x_i'\right)\right)^{-1}\right)$

Proof of the theorem is straightforward, and is omitted.

White (1980) proposed a method for estimating the asymptotic variance of OLS. Note first that we can estimate the expectations by the sample averages:

$$\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}x_i x_i'\right)^{-1}\xrightarrow{p}\left(E\left(x_i x_i'\right)\right)^{-1}.$$

If ε_i were known, we could have estimated V as follows

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\varepsilon_i^2 x_i x_i'\xrightarrow{p}V$$

As usual, we do not know $\varepsilon_i=Y_i-x_i'\beta$, as β is unknown. However, since $\hat{\beta}_{OLS}$ remains consistent, so perhaps we can use $\hat{\varepsilon}_i=Y_i-x_i'\hat{\beta}_{OLS}$. Thus

$$\hat{V}=\frac{1}{n}\sum\nolimits_{i=1}^{n}\hat{\varepsilon}_i^2 x_i x_i'.$$

Let us show that $\hat{V}\xrightarrow{p}V$. We have

$$\hat{\varepsilon}_i^2=\left(\varepsilon_i-x_i'\left(\hat{\beta}_{OLS}-\beta\right)\right)^2$$

Therefore,

$$\hat{V} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \varepsilon_i^2 x_i x_i' - 2\frac{1}{n}\sum\nolimits_{i=1}^{n} \varepsilon_i \left(x_i x_i'\right) x_i' \left(\hat{\beta}_{OLS} - \beta\right) + \frac{1}{n}\sum\nolimits_{i=1}^{n} \left(x_i x_i'\right) \left(x_i' \left(\hat{\beta}_{OLS} - \beta\right)\right)^2$$

The last two terms converge in probability to zero (this can be seen using arguments similar to those that we used in the derivation of the asymptotic equivalence of FGLS and GLS). The first term converges in probability to V by the LLN.

The White heteroskedasticity robust standard errors are computed from

$$\widehat{Var}\left(\hat{\beta}_{OLS}\right) = \frac{1}{n}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i x_i'\right)^{-1} \hat{V} \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i x_i'\right)^{-1}.$$

Note that this is a biased estimate of the covariance matrix of $\hat{\beta}$ (see Angrist and Pischke, 8.1 for a discussion).

References

[1] Amemiya, T. (1985) *Advanced Econometrics*, Basil Blackwell.

[2] Angrist, J. D., and Pischke, J-S. (2009) *Mostly harmless Econometrics. An Empiricist's Companion*, Princeton University Press.

[3] van der Vaart, A. W. (2000) *Asymptotic Statistics*, Cambridge University Press.

[4] White, H. (1980) "A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817–838