Econometrics
Part I: Basic regression.
Session 4.

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

Model selection. Mallows' C_p. Akaike Information Criterion. (These notes)
Convergence concepts. Law of Large Numbers and Central Limit Theorem. (H 6.1-6.8)
Large Sample Properties of OLS. Hypothesis Testing. Asymptotic confidence intervals. (H 7.1-7.8, 7.11-7.13, 7.16-7.18)

Model selection

Empirical researchers often face a model selection problem: which variables to include in the model and which leave out. To fix ideas, consider two models, unrestricted and restricted,

$$\begin{aligned} Y &= X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon, \\ Y &= X_1\beta_1 + \varepsilon, \end{aligned}$$

where X is an $n \times k$, X_1 is an $n \times p$, and X_2 is an $n \times q$ with $p+q = k$. Suppose that GM1-4 hold for the unrestricted model. Is it possible that we would nevertheless prefer the restricted one? Intuitively, if β_2 can only be estimated very imprecisely, it may pay off to consider the shorter model and ignore the term $X_2\beta_2$ which cannot be well estimated.

Let us formalize this intuition. Suppose we would like to use the long and the short model for prediction. The prediction of Y derived from the long model will be $\hat{Y} = X\hat{\beta}$, whereas that derived from the short model will be $\tilde{Y} = X_1\tilde{\beta}_1$, where estimators with 'hats' are based on the full set of data and those with 'tildes' are based only on X_1. A reasonable measure of accuracy for any predictor $\check{Y}$ of Y is the expected scaled sum of squared deviations of $\check{Y}$ from the best (infeasible) predictor $X\beta$

$$J = E\left[\frac{1}{\sigma^2}\left(\check{Y} - X\beta\right)'\left(\check{Y} - X\beta\right)\right]$$

Here and in what follows in the "model selection" section, the expectations are conditional on X, but we omit the conditioning notation to simplify the formulae's appearance.

For $\check{Y} = \hat{Y}$, we have

$$J_u = E\frac{\left(\hat{\beta}-\beta\right)' X'X \left(\hat{\beta}-\beta\right)}{\sigma^2} = k$$

For $\check{Y} = \tilde{Y}$,

$$J_r = E\frac{\left(\begin{pmatrix} \tilde{\beta}_1 \\ 0 \end{pmatrix} - \beta\right)' X'X \left(\begin{pmatrix} \tilde{\beta}_1 \\ 0 \end{pmatrix} - \beta\right)}{\sigma^2}.$$

To simplify the latter expression, note that $\tilde{\beta}_1$ is a biased estimator of β_1. The omitted variable bias is

$$\begin{aligned} E\tilde{\beta}_1 - \beta_1 &= \left(X_1'X_1\right)^{-1} X_1' \left(X_1\beta_1 + X_2\beta_2\right) - \beta_1 \\ &= \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2. \end{aligned}$$

Note also that (conditional) variance of $\tilde{\beta}_1$ is

$$Var\left(\tilde{\beta}_1\right) = \left(X_1'X_1\right)^{-1} X_1' Var\left(Y\right) X_1 \left(X_1'X_1\right)^{-1} = \sigma^2 \left(X_1'X_1\right)^{-1}.$$

Using the latter two displays, we obtain

$$\begin{aligned} J_r &= E\frac{\left(\tilde{\beta}_1 - E\tilde{\beta}_1\right)' X_1'X_1 \left(\tilde{\beta}_1 - E\tilde{\beta}_1\right)}{\sigma^2} + \frac{\begin{pmatrix} E\tilde{\beta}_1 - \beta_1 \\ -\beta_2 \end{pmatrix}' X'X \begin{pmatrix} E\tilde{\beta}_1 - \beta_1 \\ -\beta_2 \end{pmatrix}}{\sigma^2} \\ &= p + \frac{1}{\sigma^2} \begin{pmatrix} \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2 \\ -\beta_2 \end{pmatrix}' \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2 \\ -\beta_2 \end{pmatrix} \\ &= p + \frac{\beta_2' X_2' M_{X_1} X_2 \beta_2}{\sigma^2} \end{aligned}$$

Now compare J_u and J_r. We see that $J_r < J_u$ if and only if

$$\frac{\beta_2' X_2' M_1 X_2 \beta_2}{\sigma^2} < q.$$

When β_2 is small and/or there is (imperfect) multicollinearity between X_1 and X_2 and/or σ^2 is large, this may well happen. Can we estimate the left hand side? Yes. Consider the restricted sum of squared residuals

$$SSR_r = \left(Y - X_1\tilde{\beta}_1\right)' \left(Y - X_1\tilde{\beta}_1\right) = Y' M_{X_1} Y$$

We have

$$\begin{aligned} E\left(SSR_r\right) &= E\left(\operatorname{tr} Y'M_1Y\right) = \operatorname{tr}\left(E\left(M_1YY'\right)\right) = \operatorname{tr}\left(M_1E\left(YY'\right)\right) \\ &= \operatorname{tr}\left(M_1\left(X\beta\beta'X + \sigma^2 I_n\right)\right) = \beta'XM_1X\beta + \sigma^2 \operatorname{tr} M_1 \\ &= \beta_2'X_2M_1X_2\beta_2 + \sigma^2\left(n-p\right), \end{aligned}$$

$$\frac{\beta_2'X_2M_1X_2\beta_2}{\sigma^2} = \frac{E\left(SSR_r\right)}{\sigma^2} - (n-p)$$

Therefore, the left hand side can be estimated by $\frac{SSR_r}{\hat{\sigma}^2} - (n-p)$ and J_r can be estimated by

$$\hat{J}_r = p + \frac{SSR_r}{\hat{\sigma}^2} - (n-p) = \frac{SSR_r}{\hat{\sigma}^2} - n + 2p.$$

The latter expression is known as Mallows' C_p model selection criterion.

$$C_p = \frac{RSS_r}{\hat{\sigma}^2} - n + 2p.$$

Minimizing it across different sub-models of the "long" model yields a "short" model most adequate for "prediction". Of course, since C_p only estimates J_r, the "most adequate" should be taken with a grain of salt. For a discussion, see Mallows (1973).

Mallows' C_p can be thought of as providing a guidance for the choice of the "optimal" critical value of the F-test for testing the hypothesis that $\beta_2 = 0$. Indeed, according to C_p, we should prefer the "long" model to short only if $C_p > J_u = k$. On the other hand,

$$\begin{aligned} C_p &= \frac{SSR_r}{SSR_u/\left(n-k\right)} - n + 2p = \frac{SSR_r - SSR_u}{SSR_u/\left(n-k\right)} + 2p - k \\ &= \left(k-p\right)\left(\text{F-statistic}\right) + 2p - k \end{aligned}$$

Hence, $C_p > k$ if and only if

$$\begin{aligned} \left(k-p\right)\left(\text{F-statistic}\right) + 2p - 2k &> 0, \text{ that is,} \\ \text{F-statistic} &> 2. \end{aligned}$$

Akaike information criterion

Consider different models of conditional distribution of Y given X, say

$$f(y, \theta|x) \text{ and } g(y, \gamma|x).$$

Both models may be misspecified, so that the true conditional density $h(y|x)$ does not equal either f or g for any parameter values θ and γ. If a researcher must choose between f and g, she may try to see which of them approximate h better. For simplicity of notation, we will suppress conditioning on X, as in the previous section.

Formally, for some distance measure $d(\cdot, \cdot)$ between two distributions, the researcher may try to compare

$$\inf_{\theta} d(f(\cdot, \theta), h(\cdot)) \text{ with } \inf_{\gamma} d(g(\cdot, \gamma), h(\cdot))$$

and choose either f or g depending on which term is smaller.

Akaike (1973) proposed an implementation of this strategy using the Kullback-Leibler divergence as $d(\cdot, \cdot)$:

$$d(\varphi_1, \varphi_2) = \int \varphi_1(z) \log \frac{\varphi_1(z)}{\varphi_2(z)} dz.$$

Kullback (1959, p. 4) motivates the divergence from a Bayesian perspective. Suppose that a researcher assigns prior probabilities $\Pr(\varphi_1)$ and $\Pr(\varphi_2)$ to events that data have density φ_1 and φ_2, respectively. Suppose then one observation is made, and the value of the observation is z. How does the ratio of the posterior probabilities relate to the ratio of the prior probabilities? According to Bayes' theorem

$$\Pr(\varphi_i|z) = \frac{\varphi_i(z) \Pr(\varphi_i)}{\varphi_1(z) \Pr(\varphi_1) + \varphi_2(z) \Pr(\varphi_2)}$$

So that

$$\frac{\Pr(\varphi_1|z)}{\Pr(\varphi_2|z)} = \frac{\varphi_1(z) \Pr(\varphi_1)}{\varphi_2(z) \Pr(\varphi_2)}$$

Taking logarithms, one gets a particularly simple relationship

$$\log \frac{\Pr(\varphi_1|z)}{\Pr(\varphi_2|z)} = \log \frac{\varphi_1(z)}{\varphi_2(z)} + \log \frac{\Pr(\varphi_1)}{\Pr(\varphi_2)}$$

In words, log of posterior odds equal to the log of the prior odds plus log of the ratio of the densities at z. Kullback interprets $\log \frac{\varphi_1(z)}{\varphi_2(z)}$ as the information in $Z = z$

for discrimination in favour of φ_1 against φ_2. The KL divergence is the mean information for discrimination in favour of φ_1 against φ_2 per observation from φ_1 :

$$KL(\varphi_1, \varphi_2) = \int \varphi_1(z) \log \frac{\varphi_1(z)}{\varphi_2(z)} dy = E_{\varphi_1,Z} \log \varphi_1(Z) - E_{\varphi_1,Z} \log \varphi_2(Z)$$

It is not symmetric. Originally, Kullback and Leibler called the symmetrized expression $KL(\varphi_1, \varphi_2) + KL(\varphi_2, \varphi_1)$ the divergence. It does not satisfy the triangle inequality, so it is not formally a distance. It does have information-theoretic justifications in addition to the above bayesian motivation. In the above display, $E_{\varphi_1,Z}$ denotes the expectation with respect to random variable Z having density φ_1.

Akaike proposed to measure the quality of model $f(y, \theta)$ by 2 times the expected KL divergence between $h(y)$ and $f\left(y, \hat{\theta}_{ML}(Y)\right)$. That is

$$\begin{aligned} 2E_{h,Y} \int h(y) \log \frac{h(y)}{f\left(y, \hat{\theta}_{ML}(Y)\right)} dy &= 2E_{h,Y} E_{h,Z} \log \frac{h(Z)}{f\left(Z, \hat{\theta}_{ML}(Y)\right)} \\ &= 2E_{h,Y,Z} \log \frac{h(Z)}{f\left(Z, \hat{\theta}_{ML}(Y)\right)}, \end{aligned}$$

where Y and Z are independent random variables having the same density h, and $\hat{\theta}_{ML}(Y)$ denotes the maximum likelihood estimator of θ based on Y only. The logic for using independent Z and Y in Akaike's measure of quality is the same as in using out of sample prediction error as the measure of the quality of prediction.

Note that

$$2E_{h,Y,Z} \log \frac{h(Z)}{f\left(Z, \hat{\theta}_{ML}(Y)\right)} = 2E_{h,Z} \log h(Z) - 2E_{h,Y,Z} \log f\left(Z, \hat{\theta}_{ML}(Y)\right),$$

so to compare models f and g it is sufficient to evaluate

$$-2E_{h,Y,Z} \log f\left(Z, \hat{\theta}_{ML}(Y)\right), -2E_{h,Y,Z} \log g\left(Z, \hat{\gamma}_{ML}(Y)\right)$$

and choose the model for which such an expression is smaller. But how to evaluate $-2E_{h,Y,Z} \log f\left(Z, \hat{\theta}_{ML}(Y)\right)$ (or $-2E_{h,Y,Z} \log g\left(Z, \hat{\gamma}_{ML}(Y)\right)$)?

Akaike argues that a reasonable approximation would be

$$AIC = \underbrace{-2 \log f\left(Y, \hat{\theta}_{ML}(Y)\right)}_{-2\times\text{log likelihood}} + \underbrace{2 \dim(\theta)}_{2\times\text{dimension of } \theta}$$

It is reasonable because AIC is an unbiased estimator of $-2E_{h,Y,Z}\log f\left(Z,\hat{\theta}_{ML}(Y)\right)$. Let us see why in the linear regression context.

Suppose that the true model is

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

with GM1-4 satisfied. In addition, let us assume that $\sigma^2 = 1$ is known. However, the researcher considers a short model

$$Y = X_1\theta + e.$$

They mistakenly believe that the GM1-4 assumptions hold with $\sigma^2 = 1$ for this short model.

Of course, then, $\hat{\theta}_{ML}(Y) = (X_1'X_1)^{-1}X_1'Y$ is an estimate of the coefficient in the BLP of Y (and of Z) given X_1 only. Note that

$$\hat{\theta}_{ML}|X \sim N\left(\bar{\beta}_1, (X_1'X_1)^{-1}\right) \text{ where } \bar{\beta}_1 = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

so $\hat{\theta}_{ML}$ is biased for β_1. Also note that $X_1\bar{\beta}_1$ is the BLP of Y (and of Z) given X_1 only, and

$$E(Z - X_1\bar{\beta}_1)'(Z - X_1\bar{\beta}_1) = \beta_2'X_2'M_1X_2\beta_2 + n.$$

Now

$$-2f\left(Z,\hat{\theta}_{ML}(Y)\right) = n\log(2\pi) + \left(Z - X_1\hat{\theta}_{ML}(Y)\right)'\left(Z - X_1\hat{\theta}_{ML}(Y)\right)$$

Applying $E_{h,Y,Z}$ to both sides, we get

$$\begin{aligned}
-2E_{h,Y,Z}\log f\left(Z,\hat{\theta}_{ML}(Y)\right) &= n\log(2\pi) + E_{h,Z}\left(Z - X_1\bar{\beta}_1\right)'\left(Y - X_1\bar{\beta}_1\right) \\
&\quad + E_{h,Y}\left[\left(\hat{\theta}_{ML}(Y) - \bar{\beta}_1\right)'X_1'X_1\left(\hat{\theta}_{ML}(Y) - \bar{\beta}_1\right)\right] \\
&= n\log(2\pi) + \beta_2'X_2'M_1X_2\beta_2 + n + \dim(\theta).
\end{aligned}$$

Here $n\log(2\pi) + \beta_2'X_2'M_1X_2\beta_2 + n$ can be interpreted as -2 times expected (misspecified) log-likelihood evaluated at the value of the parameter that delivers the BLP. Estimating the BLP parameter (using a separate sample Y), adds to the uncertainty and decreases the expected likelihood value. This effect is captured by $+\dim(\theta)$ term.

Now let us investigate $-2E_{h,Y}\log f\left(Y,\hat{\theta}_{ML}\left(Y\right)\right)$. We have

$$\begin{aligned}
& -2E_{h,Y}\log f\left(Y,\hat{\theta}_{ML}\left(Y\right)\right)\\
= \; & n\log\left(2\pi\right)+E_{h,Y}\left(Y-X_1\hat{\theta}_{ML}\left(Y\right)\right)'\left(Y-X_1\hat{\theta}_{ML}\left(Y\right)\right)\\
= \; & n\log(2\pi)+E_{h,Y}SSR_r\\
= \; & n\log(2\pi)+\beta_2'X_2M_1X_2\beta_2+\left(n-\dim\left(\theta\right)\right).
\end{aligned}$$

This can be interpreted as -2 times expected (misspecified) log-likelihood evaluated at the value of the parameter that delivers the BLP MINUS $\dim\left(\theta\right)$. That is, if we use THE SAME sample to assess the likelihood of the data AND to get the ML estimate of the parameter, the expected likelihood is too large (we overfit). This is similar to the danger of using the same sample for estimation of parameters AND evaluation of the quality of the obtained predictor.

Combining the above results, we see that

$$E_{h,Y}AIC=-2E_{h,Y,Z}\log f\left(Z,\hat{\theta}_{ML}\left(Y\right)\right)$$

as claimed.

Choosing a model that minimizes AIC delivers the "best" approximation to the true model in terms of the expected KL divergence. In the context of the linear regression, we may imagine that AIC will be minimized not at the true model but at that which is sufficiently elaborate to approximate the truth well, but also sufficiently parsimonious so that its estimate is not too noisy. For a book-long treatment of model selection see Claeskens and Hjort (2008). A good paper trying to explain Akaike (1973) is deLeeuw (1992). The 2022 version of Hansen's book (available commercially) covers the above material in chapter 28.4.

Convergence concepts.

Convergence in probability. We say that a sequence of random variables $\{z_n\}_{n=1}^{\infty}$ converges in probability to a random variable z (which may be a constant) if and only if $\Pr(|z_n-z|>\varepsilon)\to 0$ as $n\to\infty$ for any positive ε. The convergence in probability is denoted as $z_n\xrightarrow{p}z$ or $\operatorname{plim} z_n=z$. If z_n are random vectors and $\|z_n\|=(z_n'z_n)^{1/2}$ is their Euclidean norm, then we say that $z_n\xrightarrow{p}z$ if $\Pr(\|z_n-z\|>\varepsilon)\to 0$ as $n\to\infty$ for any positive ε. The convergence in probability of vectors is equivalent to the element-by-element convergence in probability.

Almost sure convergence. We say that z_n almost surely converges to z if

$$\Pr\left(\lim \|z_n - z\| = 0\right) = 1.$$

The almost sure convergence is denoted as $z_n \xrightarrow{a.s.} z$. The a.s. convergence implies the convergence in probability but not vice versa.

Convergence in distribution. Random variables z_n converge in distribution to a random variable z if and only if $F_{z_n}(x) \to F_z(x)$ at all points of continuity of $F_z(x)$, where F_{z_n} and F_z are the cdf's. The definition is the same for the random vectors, with cdf's defined as

$$F_\xi(x) = \Pr\left(\xi_1 \leq x_1, ..., \xi_k \leq x_k\right).$$

Alternative names for the convergence in distribution are weak convergence or the convergence in law. We will denote the convergence in distribution as $z_n \xrightarrow{d} z$. Sometimes, notation $z_n \Rightarrow z$ is used. A useful characterization of the convergence in distribution is that it is equivalent to convergence of $Ef(z_n)$ to $Ef(z)$ for all bounded continuous functions f. A comprehensive discussion of the different convergence concepts can be found in Serfling (1980).

The component-wise convergence in distribution is not equivalent to the vector convergence in distribution. An important result linking the vector and scalar convergence in distribution is as follows.

Cramér-Wold device. $z_n \xrightarrow{d} z$, where z_n and z are k by 1, if and only if $c'z_n \xrightarrow{d} c'z$ for all $c \in \mathbb{R}^k$.

Convergence in probability implies convergence in distribution but not vice versa. Important theorem for both for convergence in probability and convergence in distribution is:

Continuous Mapping Theorem. Let g be continuous at every point $a \in C$, where $\Pr(z \in C) = 1$. Then

1) If $z_n \xrightarrow{p} z$, then $g(z_n) \xrightarrow{p} g(z)$

2) If $z_n \xrightarrow{d} z$, then $g(z_n) \xrightarrow{d} g(z)$

A useful related result is

Slutsky lemma. Let z_n, z and x_n be random variables. If $z_n \xrightarrow{d} z$ and $x_n \xrightarrow{p} c$ for a constant c, then

1) $z_n + x_n \xrightarrow{d} z + c$

2) $x_n z_n \xrightarrow{d} cz$

3) $x_n^{-1} z_n \xrightarrow{d} c^{-1} z$ provided $c \neq 0$.

Law of Large Numbers (LLN) and Central Limit Theorem (CLT)

Khinchine's LLN: If Y_i are i.i.d with finite mean $EY_i = m < \infty$ then $\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} m$.

Chebychev's LLN: If Y_i are uncorrelated, and such that $EY_i = m_i < \infty$, $Var(Y_i) = \sigma_i^2 < \infty$ and $\frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \to 0$, then $\frac{1}{n}\sum_{i=1}^{n} (Y_i - m_i) \xrightarrow{p} 0$.

The Chebyshev LLN can be easily proven using the so-called Markov inequality:

$$\Pr(|\xi| > \varepsilon) \leq \frac{E|\xi|^p}{\varepsilon^p}.$$

A proof of this inequality is simple. Let $F_\xi(x)$ be the cumulative distribution function of ξ. We have

$$\Pr(|\xi| > \varepsilon) = \int_{|x|>\varepsilon} dF_\xi(x) \leq \int_{|x|>\varepsilon} \frac{|x|^p}{\varepsilon^p} dF_\xi(x) \leq \int \frac{|x|^p}{\varepsilon^p} dF_\xi(x) = \frac{E|\xi|^p}{\varepsilon^p}.$$

In the special case when $\xi = \eta - E\eta$ and $p = 2$, the Markov inequality becomes the Chebyshev inequality

$$\Pr(|\eta - E\eta| > \varepsilon) \leq \frac{Var(\eta)}{\varepsilon^2}.$$

Lindeberg-Levy CLT. If Y_i are i.i.d. with finite mean m and variance σ^2, then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - m) \xrightarrow{d} N(0, \sigma^2)$$

The multivariate version of the theorem says: If Y_i are i.i.d. with mean m and variance-covariance Σ, then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - m) \xrightarrow{d} N(0, \Sigma). \quad (1)$$

This can be proven from the univariate version with the help of the Cramer-Wold device. Indeed, consider $c'\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - m) = \sqrt{n}(c'\frac{1}{n}\sum_{i=1}^{n} Y_i - c'm)$. By univariate classical CLT this converges to $N(0, c'\Sigma c)$, but this is the distribution of $c'N(0, \Sigma)$. Since $c'\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - m) \xrightarrow{d} c'N(0, \Sigma)$ for any c, we have (1).

OLS in large samples

The following large sample model assumptions correspond directly to small sample, GM assumptions as noted:

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

(OLS1) $E(x_i x_i')$ is finite nonsingular (GM1) $\operatorname{rank} X = k$

(OLS2) $E(y_i|x_i) = x_i'\beta$ (GM2) $E(Y|X) = X'\beta$

(OLS3) $Var(y_i|x_i) = \sigma^2$ (GM3) $Var(Y|X) = \sigma^2 I$

Remark 1 *Throughout, we will maintain an assumption that the first four moments of ε_i and x_i exist, but will not mention this explicitly. Such an existence implies, for example the existence of $E(\varepsilon_i^2 x_i x_i')$ via Cauchy-Schwarz inequality. The existence of $E(\varepsilon_i^2 x_i x_i')$ is required to apply Lindeberg-Levy CLT to $\frac{1}{\sqrt{n}}\sum x_i\varepsilon_i$.*

Theorem 2 *Under OLS0-3:*
(i) $\hat{\beta}_{OLS} \xrightarrow{p} \beta$
(ii) $\sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2 \left[E(x_i x_i')\right]^{-1}\right)$

Proof: (i)

$$\begin{aligned}\hat{\beta}_{OLS} &= (X'X)^{-1}X'Y = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'Y\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)\end{aligned}$$

By Khinchine's LLN,

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} E(x_i x_i') \text{ and } \frac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{p} E(x_i y_i)$$

By Continuous Mapping Theorem (CMT)

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \xrightarrow{p} \left[E(x_i x_i')\right]^{-1} \text{ and}$$
$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) \xrightarrow{p} \left[E(x_i x_i')\right]^{-1} E(x_i y_i)$$

Further,

$$\begin{aligned}
\left[E\left(x_i x_i'\right)\right]^{-1} E\left(x_i y_i\right) &= \left[E\left(x_i x_i'\right)\right]^{-1} E\left(E\left(x_i y_i | x_i\right)\right) \\
&= \left[E\left(x_i x_i'\right)\right]^{-1} E\left(x_i E\left(y_i | x_i\right)\right) \\
&= \left[E\left(x_i x_i'\right)\right]^{-1} E\left(x_i x_i'\right) \beta \\
&= \beta.
\end{aligned}$$

In summary,

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta.$$

Proof of (ii).

$$\hat{\beta}_{OLS} - \beta = \left(X'X\right)^{-1} X'\varepsilon = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

$$\sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) = \left(X'X\right)^{-1} X'\varepsilon = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

By the CMT, the Cramér-Wold, and Slutsky's lemma

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i\right) &\xrightarrow{d} \left[E\left(x_i x_i'\right)\right]^{-1} N\left(0, \sigma^2 E\left(x_i x_i'\right)\right) \\
&\sim N\left(0, \left[E\left(x_i x_i'\right)\right]^{-1} \sigma^2 E\left(x_i x_i'\right) \left[E\left(x_i x_i'\right)\right]^{-1}\right)
\end{aligned}$$

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

Note the following calculation within the above proof:

$$\begin{aligned}
Var\left(x_i \varepsilon_i\right) &= E\left[x_i \varepsilon_i \varepsilon_i' x_i'\right] - E\left[x_i \varepsilon_i\right] E\left[x_i \varepsilon_i\right]' \\
&= E\left[\varepsilon_i^2 x_i x_i'\right] - E\left[x_i \varepsilon_i\right] E\left[x_i \varepsilon_i\right]' \\
&= E\left[E\left(\varepsilon_i^2 x_i x_i' | x_i\right)\right] - E\left[E\left(x_i \varepsilon_i | x_i\right)\right] E\left[x_i \varepsilon_i\right]' \\
&= E\left[E\left(\varepsilon_i^2 | x_i\right) x_i x_i'\right] - E\left[x_i E\left(\varepsilon_i | x_i\right)\right] E\left[x_i \varepsilon_i\right]' \\
&= \sigma^2 E\left[x_i x_i'\right].
\end{aligned}$$

Part (i) of the above theorem holds even if the condition of a linear CEF, (OLS2), is replaced with a weaker condition:

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

Theorem 3 *Under OLS0-1-2'-3,*
(i) $\hat{\beta}_{OLS} \xrightarrow{p} \beta$

We also have

Theorem 4 *(continued) Under OLS0-1-2-3,*
(iii) $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$
(iv) $W \xrightarrow{d} \chi^2(p)$
(v) $t \xrightarrow{d} N(0,1)$.

Proof: The proof of (i) and (ii) remains essentially the same. For (iii), we have

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-k}\varepsilon' M_X \varepsilon = \frac{1}{n-k}\varepsilon'\varepsilon - \frac{1}{n-k}\varepsilon' X \left(X'X\right)^{-1} X'\varepsilon \\ &= \frac{n}{n-k}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \frac{n}{n-k}\frac{1}{n}\sum_{i=1}^{n} x_i'\varepsilon_i \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\end{aligned}$$

By Khinchine's LLN and Slutsky's theorem, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. For (iv), write

$$W = \frac{\sqrt{n}\left(R\hat{\beta} - q\right)'\left(\sigma^2 R\left(\frac{1}{n}X'X\right)^{-1}R'\right)^{-1}\sqrt{n}\left(R\hat{\beta} - q\right)}{\hat{\sigma}^2/\sigma^2}$$

As we have seen, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, and

$$\begin{aligned}\sqrt{n}\left(R\hat{\beta} - q\right) &= \sqrt{n}R\left(\hat{\beta} - \beta\right) \xrightarrow{d} N\left(0, \sigma^2 R\left[E\left(x_i x_i'\right)\right]^{-1}R'\right) \\ &= \left(\sigma^2 R\left[E\left(x_i x_i'\right)\right]^{-1}R'\right)^{1/2} N\left(0, I_p\right).\end{aligned}$$

Furthermore, since $\frac{1}{n}X'X \xrightarrow{p} E\left(x_i x_i'\right)$, by the Continuous Mapping Theorem,

$$\left(\sigma^2 R\left(\frac{1}{n}X'X\right)^{-1}R'\right)^{-1} \xrightarrow{p} \left(\sigma^2 R\left[E\left(x_i x_i'\right)\right]^{-1}R'\right)^{-1}.$$

Finally, by Slutsky's lemma and by Continuous Mapping Theorem

$$W \xrightarrow{d} N\left(0, I_p\right)' N\left(0, I_p\right) \sim \chi_p^2.$$

Part (v) can be shown similarly.

Recall the small sample result: $W/p|X \sim F(p, n-k)$ and thus, $W|X \sim pF(p, n-k)$ and note the following for compatibility to the large sample result: $pF(p, n-k) \xrightarrow{d} \chi^2(p)$.

Asymptotic confidence intervals

We can use the fact that $t \xrightarrow{d} N(0,1)$ to build asymptotic confidence intervals for β_j. Indeed, in large samples,

$$\begin{aligned} \Pr\left(\left|\frac{\sqrt{n}\left(\hat{\beta}_j - \beta_j\right)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}X'X\right)^{-1}_{jj}}}\right| \leq 1.96\right) &\approx 0.95, \text{ and} \\ \Pr\left(\left|\hat{\beta}_j - \beta_j\right| \leq 1.96\sqrt{\hat{\sigma}^2\left(X'X\right)^{-1}_{jj}}\right) &\approx 0.95, \end{aligned}$$

so that

$$\Pr\left(\hat{\beta}_j - 1.96\sqrt{\hat{\sigma}^2\left(X'X\right)^{-1}_{jj}} \leq \beta_j \leq \hat{\beta}_j + 1.96\sqrt{\hat{\sigma}^2\left(X'X\right)^{-1}_{jj}}\right) \approx 0.95.$$

What sample is large?

Often students ask what size of the sample would qualify as large. Some light on this issue can be shed by the following result, known as the Berry-Esseen theorem

Theorem 5 *Let z_i be i.i.d. and such that*

$$Ez_i = 0, \ Ez_i^2 = \sigma^2 > 0, \ \text{and } E\left(|z_i|^3\right) = \rho < \infty.$$

Let F_n be the cdf of $\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n} z_i$, and let Φ be the cdf of $N(0,1)$. Then, for all n and x,

$$|F_n(x) - \Phi(x)| \leq \frac{\rho}{\sigma^3\sqrt{n}}.$$

The theorem tells us that the quality of the normal approximation depends not only on the sample size, but also on the size of the absolute third moment relative to the standard deviation. If it is large, it may take a very large sample to accurately approximate the distribution of the average of z_i by the normal one. On the other hand, in the extreme case where z_i are themselves normal, the approximation is exact for any sample size.

References

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

[2] Akaike, H. (1973) "Information theory and the maximum likelihood principle," in *2nd International Symposium on Information Theory* (B.N. Petrov and F. Csaki, eds.) Akademiai Kiado, Budapest.

[3] Claeskens, G., and Hjort, N. L. (2008) *Model Selection and Model Averaging*, Cambridge University Press.

[4] deLeeuw, J. (1992) "Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle," in Kotz, S., and Johnson, N.L. eds. *Breakthroughs in Statistics: vol1 Foundations and Basic Theory*, Springer-Verlag, 599–609.

[5] Kullback, S. (1959) *Information Theory and Statistics*, John Wiley & Sons.

[6] Mallows, C.L. (1973) "Some Comments on C_p," Technometrics 42, 87-94.

[7] Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, John Wiley&Sons.