Econometrics (I)

Mei-Yuan Chen
Department of Finance
National Chung Hsing University

Sept. 11, 2007
Econometrics

\[ \sigma, \alpha, \beta, \frac{a}{b}, \int_{a}^{b} f(x) dx \]

1. Econometrics is the study of estimation and inference for economic models using economic data.

2. Econometric theory concerns the study and development of tools and methods for applied econometric applications.

3. Applied econometrics concerns the application of these tools to economic data.

4. Financial Econometrics is the study of estimation and inference for economic models using economic data.
Economics Variables

An economic variable is a measure which is well defined by economic theories to evaluate the outcome or performance of an economic activity in a market, a society, a country, or whole the world. For examples, in the national account,$$ Y = C + I + G + (X - M), $$ where $Y$ is the GDP, $C$ is the national private consumption, $I$ is the investment of private sector, $G$ is the government expenditure, $X$ is the export and $M$ is the import. These are all economic variables of a country.
Economics Random Variables

An economic random variable assigns different real values onto the possible outcomes or performances of an economic activity or a market condition via the formal definition from economic theories.
For examples, denote $Y_{(2006, \cdot)}$ as the GDP of all countries in the world on 2006, i.e.,

$$Y_{(2006, \cdot)} : \{US, Japan, UK, Taiwan, China, etc\} \rightarrow \{y_{(2006,i)}, i = 1, 2, 3, \ldots\}.$$  

On the other hand, denote $Y_{(\cdot, 4)}$ as the GDP of Taiwan from 1950 to 2006, i.e.,

$$Y_{(\cdot, 4)} : \{1950, 1951, 1952, etc\} \rightarrow \{y_{(t,4)}, t = 1950, 1951, 1952, \ldots\}.$$

A sample from $Y_{(2006, \cdot)}$ consists of a sample of cross-sectional data and from $Y_{(\cdot, 4)}$ consists of a sample of time series data. Denote $z_{(t,i)}$ as the $(t, i)$th element of $\{[y_{(1950,i)} \ y_{(1951,i)} \ \cdots \ y_{(2006,i)}], i = 1, 2, 3, \ldots\}'$, then $\{z_{(t,i)}, t = 1, \ldots, T, i = 1, \ldots, N\}$ makes a sample of the panel data.
Denote $X_{(2006, \cdot)}$ as the national consumption of all countries in the world on 2006, i.e.,

$$X_{(2006, \cdot)} : \{US, Japan, UK, Taiwan, China, etc\} \rightarrow \{x_{(2006,i)}, i = 1, 2, 3, \ldots \}.\]

On the other hand, denote $X_{(\cdot,4)}$ as the GDP of Taiwan from 1950 to 2006, i.e.,

$$X_{(\cdot,4)} : \{1950, 1951, 1952, etc\} \rightarrow \{x_{(t,4)}, t = 1950, 1951, 1952, \ldots \}.\]

A sample from $X_{(2006, \cdot)}$ consists of a sample of cross-sectional data and from $X_{(\cdot,4)}$ consists of a sample of time series data. Denote $z_{(t,i)}$ as the $(t,i)$th element of $\{[x_{(1950,i)} \ x_{(1951,i)} \ \cdots \ x_{(2006,i)}], i = 1, 2, 3, \ldots]^\prime\}$, then $\{z_{(t,i)}, t = 1, \ldots, T, i = 1, \ldots, N\}$ makes a sample of the panel data.
Measures for a Random Variable

The goal of statistics for a random variable $X$ (univariate random variable) is to explore or demonstrate the distribution of all possible realizations (real values) plotted on a real line via measures for location (expectation, $E(X)$), dispersion (variance, $\text{var}(X)$), skewness (skewness coefficient, $\alpha_3(X)$), and kurtosis (kurtosis coefficient, $\alpha_4(X)$) by a sample of data randomly (drawn independently and identically) from the random variable under study.
Distribution function of a Random Variable

Generally, the distribution of a random variable $X$ can be represented by a cumulative distribution function (CDF) $F(X)$ which is a function of $x, \forall x$. Its definition is

$$F(c) = \int_{-\infty}^{c} dF(x)$$

$$= \int_{-\infty}^{c} f(x) dx, \text{ if } F(x) \text{ is differentiable } \forall x,$$

where $dF(x)$ denote the increment of function $F(\cdot)$ at $x$ and $f(x)$ is called the probability density function (pdf) of $X$. 
Definition of Measures

\[
E(X) = \int_{-\infty}^{\infty} x dF(x)
\]

\[
= \int_{-\infty}^{\infty} x f(x) dx, \quad \text{if } F(x) \text{ is differentiable } \forall x,
\]

\[
\sigma_X^2 = \int_{-\infty}^{\infty} (x - E(X))^2 dF(x)
\]

\[
= \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx, \quad \text{if } F(x) \text{ is differentiable } \forall x,
\]

\[
\alpha_3(X) = \frac{E[(X - E(X))^3]}{\sigma_X^3}
\]

\[
\alpha_4 = \frac{E[(X - E(X))^4]}{\sigma_X^4}.
\]
Given a random sample \( \{x_1, x_2, \ldots, x_n\} \) from \( X \) (the sample is independent and identical, i.e., i.i.d.),

1. \( \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow E(X) \);  
2. \( s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{(n - 1)} \Rightarrow \text{var}(X) \);  
3. \( \hat{\alpha}_3(X) = \left[ \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^3}{(n - 1)} \right] / s_x^3 \Rightarrow \alpha_3(X) \);  
4. \( \hat{\alpha}_4(X) = \left[ \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^4}{(n - 1)} \right] / s_x^4 \Rightarrow \alpha_4(X) \).

Note that the random sample \( \{x_1, x_2, \ldots, x_n\} \) is i.i.d. means

1. the sampling distribution of each observation \( x_i \) is the same as the population distribution of \( X \). That is,  
   \( E(x_i) = E(X), \text{var}(x_i) = \text{var}(X), \alpha_3(x_i) = \alpha_3(X), \) and  
   \( \alpha_4(x_i) = \alpha_4(X) \) for all \( i \);  
2. \( \text{cov}(x_i, x_j) = 0 \), that is \( x_i \) and \( x_j \) are independent, for all \( i \neq j \).
Assume the random variable $X$ is normally distributed (the sampling distribution of each sample observation $x_i, i = 1, \ldots, n$ is same as $X$), then

1. $\bar{x}_n$ has a normal sampling distribution with mean $E(X)$ and variance $\text{var}(X)/n$;

2. $(n - 1)s_x^2/\text{var}(X)$ has the sampling distribution $\chi^2(n - 1)$. 
For the null hypothesis $H_0 : E(X) = \mu_0$, under $X$ is normally distributed, the test statistic

$$t_{\bar{x}_n} = \frac{\bar{x}_n - \mu_0}{s_{\bar{x}_n}} = \frac{\bar{x}_n - \mu_0}{s_x/\sqrt{n}},$$

under the null, has the sampling distribution of $t(n - 1)$. 

M.-Y. Chen  
Econometrics
For the null hypothesis $H_0 : \text{var}(X) = \sigma_0^2$, under $X$ is normally distributed, the test statistic

$$\phi_{s_x^2} = \frac{(n - 1)s_x^2}{\sigma_0^2},$$

under the null, has the sampling distribution of $\chi^2(n - 1)$. 
Bivariate Random Variables

Put $X_{(2006,\cdot)}$ and $Y_{(2006,\cdot)}$ to define bivariate random variables as

$$(X_{(2006,\cdot)}, Y_{(2006,\cdot)}) : \{US, Japan, UK, Taiwan, China, etc\}$$
$$\rightarrow \{(x_{(2006,i)}, y_{(2006,i)}), i = 1, 2, 3, \ldots\}.$$ 

Similarly, Put $X_{(\cdot,4)}$ and $Y_{(\cdot,4)}$

$$(X_{(\cdot,4)}, Y_{(\cdot,4)}) : \{1950, 1951, 1952, etc\}$$
$$\rightarrow \{(x_{(t,4)}, y_{(t,4)}), t = 1950, 1951, 1952, \ldots\}.$$
The goal of statistics for bivariate random variables is to explore or demonstrate the distribution of all possible realizations (real values), \((x_{(2006,i)}, y_{(2006,i)})\) or \((x_{(t,4)}, y_{(t,4)})\), plotted on a real \(X - Y\) plane via measures for location (expectation), dispersion (variance), skewness (skewness coefficient), and kurtosis (kurtosis coefficient) for each univariate random variable and correlation between these two random variables by a sample of data randomly from these two random variable under study.
Joint Distribution Function and Probability Density Function

Generally, the joint cumulative distribution of $X$ and $Y$ is written as $F(X, Y)$ and is a function of $(x, y), \forall x, y$ and is defined as

$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} dF(x, y)$$

$$= \int_{-\infty}^{a} \int_{-\infty}^{b} df(x, y) dx dy, \quad \text{if } F(x, y) \text{ is differentiable } \forall x, y$$

where $dF(x, y)$ is the increment of function $F(\cdot, \cdot)$ and $f(x, y)$ is the joint probability density function of $X$ and $Y$ at $(x, y)$. 

M.-Y. Chen  
Econometrics
Covariance and Correlation between $X$ and $Y$

The covariance between $X$ and $Y$ is defined as

$$\text{cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))dF(x, y)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))f(x, y)dxdy.$$
The correlation coefficient between $X$ and $Y$ is defined as

$$
\rho_{X,Y} = \text{cov}(Z_X, Z_Y) = E(Z_X Z_Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}},
$$

where $Z_X$ and $Z_Y$ denote the $Z$-score transformations of $X$ and $Y$, i.e.,

$$
Z_X = \frac{X - E(X)}{\sqrt{\text{var}(X)}}, \quad Z_Y = \frac{Y - E(Y)}{\sqrt{\text{var}(Y)}}.
$$
A random sample of pair observations, 
\(\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}\), is drawn from bivariate random variables \((X, Y)\) indicates that

1. each pair of observation \((x_i, y_i)\) has the joint sampling distribution same as \((X, Y)\) distributed, \(x_i\) has the sampling distribution as \(X\) distributed and \(y_i\) has the sampling distribution as \(Y\) distributed.

2. the joint sampling distribution of the pair of observation \((x_i, y_i)\) is independent of the one of \((x_j, y_j)\), the sampling distribution of \(x_i\) is independent of the one of \(x_j\), and the sampling distribution of \(y_i\) is independent of the one \(y_j\) for \(i \neq j\).
The sample covariance of a random sample \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \) is defined as

\[
s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n).
\]

The sample correlation is defined as

\[
\hat{\rho}_{x,y} = \frac{s_{x,y}}{\sqrt{s_x^2} \sqrt{s_y^2}},
\]

where \( s_x^2 \) and \( s_y^2 \) are the sample variance of \( X \) and \( Y \) from the sample.
The marginal cumulative distribution function of $X$ is

$$F_X(x) = \int_{-\infty}^{\infty} dF(x, y)$$

and the marginal probability density function of $X$ is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$ 

Similar definitions is applied to $Y$. 

M.-Y. Chen
Econometrics
Conditional Distribution Functions

The conditional cumulative distribution function of $Y$ at $X = c$ is

$$F_{Y|X=c}(x, y) = \frac{F(c, y)}{\int_{-\infty}^{\infty} dF(c, y)}$$

and the conditional probability function of $Y$ at $X = c$ is

$$f_{Y|X=c}(x, y) = \frac{f(c, y)}{\int_{-\infty}^{\infty} f(c, y)dy}.$$ 

Generally, the conditional distribution of $Y$ on $X = x$ is written as $Y|X = x$ and can be viewed as a univariate random variable and which states the distribution of $Y$ varies with the variables $X$ in the population. Therefore, $Y|X$ is the function of random variables along different realizations of $X$. 

M.-Y. Chen  Econometrics
Conditional Mean and Variance

The conditional mean of $Y$ at $X = c$ is defined as

$$E(Y|X = c) = \int_{-\infty}^{\infty} ydF_{Y|X=c}(x, y) = \int_{-\infty}^{\infty} yf_{Y|X=c}(x, y)dy$$

and the conditional variance of $Y$ at $X = c$ is defined as

$$\text{var}(Y|X = c) = \int_{-\infty}^{\infty} (y - E(Y|X = c))^2 dF_{Y|X=c}(x, y) = \int_{-\infty}^{\infty} (y - E(Y|X = c))^2 f_{Y|X=c}(x, y)dy.$$ 

In general, $E(Y|X = x) = m(x)$ and $\text{var}(Y|X = x) = v(x)$ are functions of $x$.  

M.-Y. Chen  Econometrics
Regression Equation

The conditional regression error $e|X = c$ is defined to be the difference between realizations of $Y|X = c$ (the pair realizations are $(c, y)$) and its conditional mean $E(Y|X = c) = m(c)$:

$$(e|X = c) = (Y|X = c) - m(c),$$

which is also a random variable. Its mean

$$E(e|X = c) = E(Y|X = c) - E(m(c)) = m(c) - m(c) = 0.$$ 

By construction, all realizations of $Y$ conditional $X = c$ (denoted as $y|X = c$) satisfies

$$(y|X = c) = m(x) + (e|X = c).$$
In general, for different values of $x$, all paired realizations of $(x, y)$ can be represented as

$$y = m(x) + e,$$

where $e$ is called the regression error. Properties of regression error:

1. $E(e|x) = 0$ for all $x$;
2. $E(e) = 0$;
3. $E[h(x)e] = 0$ for any function of $h(\cdot)$;
4. $E(xe) = 0$. 

M.-Y. Chen  
Econometrics
Best Predictor of $Y$ on $X = c$

Let $g(c)$ be an arbitrary predictor (demonstration) of $Y$ given $X = c$. The expected squared error using this prediction function is

$$E[(Y|X = c) - g(c)]^2$$

$$= E[m(c) + (e|X = c) - g(c)]^2$$

$$= E[(e|X = c)^2] + 2E[(e|X = c)(m(c) - g(c))] + E[(m(c) - g(c))^2]$$

$$= E[(e|X = c)^2] + E[(m(c) - g(c))^2] \geq E[(e|X = c)^2],$$

by the third property of the regression error. The right-hand side is minimized by setting $m(c) = g(c)$. This concludes the condition mean $m(x)$ is the best predictor of $Y$ condition on $X = x$. 

M.-Y. Chen Econometrics
Conditional Variance of $Y$ on $X = c$

The conditional variance of $Y$ on $X = c$ is defined as

$$
\sigma^2(c) = \text{var}(Y|X = c) = E[(Y|X = c) - E(Y|X = c)]^2 = E[(e|X = c)^2] = E[e^2|X = c].
$$

Generally, $\sigma^2(x)$ is a non-trivial function of $x$. The conditional standard deviation is

$$
\sigma(x) = \sqrt{\sigma^2(x)}, \forall x.
$$
1. the regression error is **homoskedastic** if 
\[ \sigma^2(x) = E(e^2|X = x) = \sigma^2 \] is constant for all \( x \);

2. the regression error is **heteroskedastic** if 
\[ \sigma^2(x) = E(e^2|X = x) \] is not constant for all \( x \).
The most commonly applied econometric tool is regression. This is used when the goal is to quantify the impact of one set of variables (the regressors, conditioning variable, or covariates) on another variable (the dependent variable). Let $y$ denote the dependent variable and $(x_1, x_2, \ldots, x_k)$ denote the $k$ regressors. It is convenient to write the set of regressors as a vector in $\mathbb{R}^k$. As the conditional mean is the best predictor of $y$ on $(x_1, x_2, \ldots, x_k)$, it is usually studied in econometric analysis.
Linear Conditional Mean Models

Suppose a linear condition mean is considered, i.e.,

\[ E(Y|\mathbf{x}) = m(\mathbf{x}) = \beta_{10}x_1 + \beta_{20}x_2 + \cdots + \beta_{k0}x_k, \]

which is linear in function of \( \mathbf{x} \). The goal of a linear regression analysis is to collect a random sample to get information about \( \beta_{10}, \beta_{20}, \cdots, \beta_{k0} \) by estimations and then to verify the results implied by economic (finance) theory by inference. And then,

\[
\begin{align*}
y|\mathbf{x} & = \beta_{10}x_1 + \beta_{20}x_2 + \cdots + \beta_{k0}x_k + e|\mathbf{x} \\
& = \mathbf{x}'\beta_0 + e|\mathbf{x}.
\end{align*}
\]
Introduce an Intercept Term

Given the conditional mean specified previously,

\[ E(Y|\mathbf{x}) = m(\mathbf{x}) = \beta_{10}x_1 + \beta_{20}x_2 + \cdots + \beta_{k0}x_k, \]

zero unconditional mean is assumed, since

\[ E(Y|\mathbf{X}) = 0 = E(Y) \] given \( \beta_{i0} = 0, \forall i \). To get rid of this restriction, an alternative specification for conditional mean is

\[ E(Y|\mathbf{x}) = m(\mathbf{x}) = \beta_{10} + \beta_{20}x_2 + \cdots + \beta_{k0}x_k. \]

The unconditional mean is set to be \( \beta_{10} \).
Best Linear Predictor

For any $\beta \in \mathbb{R}^k$ a linear predictor for $y$ on $X = x$ is $x'\beta$ which with expected squared prediction error

$$S(\beta) = E[y - x'\beta]^2$$

$$= E(y^2) - 2\beta'E(xy) + \beta'E(xx')\beta$$

which is quadratic in $\beta$. The best linear predictor is obtained by selecting $\beta$ to minimize $S(\beta)$. That is

$$\beta^* = \arg \min_{\beta} S(\beta).$$

The first-order condition for minimization is

$$\nabla_\beta S(\beta) = -2E(xy) + 2E(xx')\beta \overset{set}{=} 0.$$

We have

$$\beta^* = [E(xx')]^{-1}E(xy).$$
Given $\beta^* = [E(xx')]^{-1}E(xy)$, denote $e^* = y - x'\beta^*$ and then we have

\[
E(xe^*) = E[x(y - x'\beta*)] = E\{x[y - x'E(xx')^{-1}E(xy)]\} = E\{E[x[y - x'E(xx')^{-1}E(xy)]|x]\} = E\{xE(y|x) - xx'(xx')^{-1}xE(y|x)\} = 0.
\]

Notice that the error $e^*$ from the linear prediction equation ($x'\beta^*$) is equal to the error from the regression equation, $e = y - m(x)$, when (and only when) the conditional mean is linear in $x$; otherwise they are distinct.
Linear Multiple Regression Models

Given a sample of $T$ observations, \{$(y_t, x_{t2}, \ldots, x_{tk}), t = 1, \ldots, T$\}, this specification can also be expressed as the identification condition:

$$y_t = \beta_{10} + \beta_{20}x_{t2} + \cdots + \beta_{k0}x_{tk} + e_t,$$

for $t = 1, \ldots, T$. This equation is called a linear multiple regression model. In matrix notation, above equation can be written as

$$y = X\beta_0 + e,$$  \hspace{1cm} (1)

where $\beta_0 = (\beta_{10} \beta_{20} \cdots \beta_{k0})'$ is the vector of unknown parameters.
\( y \) and \( X \) contain all the observations of the dependent and explanatory variables, i.e.,

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T2} & \cdots & x_{Tk} \end{bmatrix} = [\ell \quad X_2 \quad \cdots \quad X_k],
\]

where each column vector of \( X \) contains \( T \) observations for an explanatory variable and \( \ell \) is a \( T \times 1 \) vector with all elements 1. \( x_t \) is the column vector denoted as \([1 \quad x_{t2} \quad \cdots \quad x_{tk}]'\).
The basic “identifiability” requirement of this specification is that the number of regressors, $k$, is strictly less than the number of observations, $T$, such that the matrix $X$ is of full column rank $k$. That is, the model does not contain any “redundant” regressor. It is also typical to set the first explanatory variable as the constant one so that the first column vector of $X$ is a $T \times 1$ vector of ones, $\ell$. To summary, the identification condition is ID 1:

$$y_t = \beta_{10} + \beta_{20}x_{t2} + \beta_{30}x_{t3} + \cdots + \beta_{k0}x_{tk} + e_t, \ t = 1, \ldots, T.$$
Estimation for the Best Linear Predictor

Given the sample average is the counterpart of population expectation, that is, the sample moments

\[
\hat{E}(xy) = \frac{1}{T} \sum_{t=1}^{T} x_t y_t, \quad \hat{E}(xx') = \frac{1}{T} \sum_{t=1}^{T} x_t x'_t
\]

are used to estimate the population moments \( E(xy) \) and \( E(xx') \), respectively. Therefore, the estimator for the best linear predictor is, given \( \left[ \sum_{t=1}^{T} x_t x'_t \right]^{-1} \) exists,

\[
\hat{\beta}_T^* = \left[ \hat{E}(xx') \right]^{-1} \hat{E}(xy) = \left[ \sum_{t=1}^{T} x_t x'_t \right]^{-1} \sum_{t=1}^{T} x_t y_t = \left( X'X \right)^{-1} X'y.
\]
Another Derivation for Estimation of the Best Linear Predictor

Given the fourth property of regression error, $E(\mathbf{xe}) = \mathbf{0}$ and the linear conditional mean, denote $g(\beta) = E[\mathbf{x}_t(y_t - \mathbf{x}'_t\beta)] = \mathbf{0}$. Therefore, the sample counterpart of $g(\beta)$ is

$$
\hat{g}(\beta) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}'_t\beta)
$$
By setting $\hat{g}(\beta) = 0$ and solving the system of $k$ equations, the solution of the parameters $\beta$, $\tilde{\beta}_T$ satisfy

$$0 = \hat{g}(\tilde{\beta}) = \frac{1}{T} \sum_{t=1}^{T} x_t (y_t - x_t' \tilde{\beta})$$

$$= \frac{1}{T} \sum_{t=1}^{T} x_t y_t - \frac{1}{T} \sum_{t=1}^{T} x_t x_t'.$$

We have the estimator derived by the method of moment as

$$\tilde{\beta}_T = [ \sum_{t=1}^{T} x_t x_t' ]^{-1} \sum_{t=1}^{T} x_t y_t = (X'X)^{-1} X'y$$

which is same as $\hat{\beta}^*_T$. 
OLS Estimator

Our objective now is to find a $k$-dimensional regression hyperplane that “best” fits the data $(y, X)$. For obtaining an OLS estimators, we must minimize the average of the sum of squared errors:

$$Q(\beta) := \frac{1}{T}(y - X\beta)'(y - X\beta).$$

(2)
OLS Estimator: First Order Condition

The first order conditions for the OLS minimization problem are:

\[ \nabla_\beta Q(\beta) = \frac{\nabla_\beta (y'y - 2y'X\beta + \beta'X'X\beta)}{T} \]
\[ = -2X'(y - X\beta)/T \]
\[ \overset{\text{set}}{=} 0, \]

the last equality can also be written as

\[ X'X\beta_T \overset{\text{set}}{=} X'y \]

which is known as the normal equation.
OLS Estimator

To have a unique solution for the system equation for $\beta$, $(X'X)^{-1}$ has to exist. This is first assumption has to be satisfied to have solution for $\beta$ uniquely.

[A2] The $T \times k$ data matrix $X$ is full column rank. Given that $X$ is of full column rank, $X'X$ is p.d. and hence invertible. The solution to the normal equations can then be expressed as

$$\hat{\beta}_T = (X'X)^{-1}X'y.$$  \hspace{1cm} (3)

Note that $\hat{\beta}_T = \tilde{\beta}_T = \hat{\beta}^*$. 

It is easy to see that the second order condition is also satisfied because

\[ \nabla_{\beta}^2 Q(\beta) = 2(X'X)/T \]

is p.d. Hence, \( \hat{\beta}_T \) is the minimizer of the OLS criterion function and known as the OLS estimator for \( \beta \). As the matrix inverse is unique, the OLS estimator is also unique.
The vector of OLS fitted values is

$$\hat{y} = X\hat{\beta}_T,$$

and the vector of OLS residuals is

$$\hat{e} = y - \hat{y}.$$

By the normal equations, $X'\hat{e} = 0$ so that $\hat{y}'\hat{e} = 0$. When the first regressor is the constant one, $X'\hat{e} = 0$ implies that

$$\ell'\hat{e} = \sum_{t=1}^T \hat{e}_t = 0.$$ It follows that $\sum_{t=1}^T y_t = \sum_{t=1}^T \hat{y}_t$, and the sample average of the data $y_t$ is the same as the sample average of the fitted values $\hat{y}_t$. 

M.-Y. Chen

Econometrics
If $\mathbf{X}$ is not of full column rank and then its column vectors satisfies an exact linear relationship, this is also known as the problem of **exact multicollinearity**. In this case, without loss of generality we can write

$$\mathbf{x}_1 = \gamma_2 \mathbf{x}_2 + \cdots + \gamma_k \mathbf{x}_k,$$

where $\mathbf{x}_i$ is the $i$th column of $\mathbf{X}$ and $\gamma_2, \ldots, \gamma_k$ are not all zero. Then, for any number $a \neq 0$,

$$\beta_1 \mathbf{x}_1 = (1 - a) \beta_1 \mathbf{x}_1 + a \beta_1 (\gamma_2 \mathbf{x}_2 + \cdots + \gamma_k \mathbf{x}_k).$$
The linear specification (1) is thus observationally equivalent to

\[ X\beta^* := (1 - a)\beta_1 x_1 + (\beta_2 + a\beta_1 \gamma_2)x_2 + \cdots + (\beta_k + a\beta_1 \gamma_k)x_k, \]

where the elements of \( \beta^* \) vary with \( a \) and therefore could be anything. That is, the parameter vector \( \beta \) is not identified when exact multicollinearity is present. Practically, when \( X \) is not of full column rank, \( X'X \) is not invertible, and there are infinitely many solutions to the normal equations \( X'X\beta = X'y \). Consequently, the OLS estimator \( \hat{\beta}_T \) cannot be computed as (3). Exact multicollinearity usually arises from inappropriate model specifications.
For example, including both total income, total wage income, and total non-wage income as regressors results in exact multicollinearity because total income is, by definition, the sum of wage and non-wage income.

It is also easy to verify that the magnitude of the coefficient estimates $\hat{\beta}_{iT}$ are affected by the measurement units of variables. Thus, a larger coefficient estimate does not necessarily imply that the associated explanatory variable is more important in explaining the behavior of $y$. In fact, the coefficient estimates are not comparable in general.
The OLS estimators are derived without resorting to the knowledge of the “true” relationship between $y$ and $X$. That is, whether $y$ is indeed generated according to our linear specification is irrelevant to the computation of the OLS estimator; it does affect the properties of the OLS estimator, however.
Geometric Interpretations

\[ \hat{e} = (I - P_X)y \]

\[ P_X y = x_1 \hat{\beta}_1 T + x_2 \hat{\beta}_2 T \]
Frisch-Waugh-Lovell

Let $X = [X_1 \ X_2]$, where $X_1$ is $T \times k_1$ and $X_2$ is $T \times k_2$, and $k_1 + k_2 = k$. We can write

$$y = X_1 \beta_1 + X_2 \beta_2 + \text{random error},$$

and $\hat{\beta}_T = (\hat{\beta}_1' \hat{\beta}_2')'$. Consider the following regressions:

$$y = X_1 b_1 + \text{random error},$$
$$y = X_2 b_2 + \text{random error},$$

and their OLS estimators are $\hat{b}_1_T$ and $\hat{b}_2_T$. Are $\hat{\beta}_1_T = \hat{b}_1_T$ and $\hat{\beta}_2_T = \hat{b}_2_T$? Usually not!
Frisch-Waugh-Lovell Theorem

Let $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$ and $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$ denote the orthogonal projection matrices on $\text{span}(X_1)$ and $\text{span}(X_2)$, respectively. Given a vector $y$, $(I - P_{X_2})y$ and $(I - P_{X_1})y$ can be uniquely decomposed into two orthogonal components:

$$(I - P_{X_2})y = (I - P_{X_2})X_1\hat{\beta}_1 + (I - P_X)y,$$

$$(I - P_{X_1})y = (I - P_{X_1})X_2\hat{\beta}_2 + (I - P_X)y.$$
Frisch-Waugh-Lovell Theorem: Illustration

1. Perform a regression of $y$ on $X_1$ ($T \times k_1$) and then the resulted residual vector ($T \times 1$), $\hat{e}_{y-X_1}$, is reserved.

2. Perform $k_2$ regressions of each component in $X_2$ on $X_1$ and then $k_2$ resulted residual vectors, $\hat{E}_{X_2-X_1}$ ($T \times k_2$), are reserved.

3. Perform a regression of $\hat{e}_{y-X_1}$ on $\hat{E}_{X_2-X_1}$ and then the estimator $\tilde{\beta}_{2T}$ is same as $\hat{\beta}_{2T}$ in the regression $\hat{y} = X_1\hat{\beta}_{1T} + X_2\hat{\beta}_{2T}$.
Measures of Goodness of Fit: Absolute Measure

A natural goodness-of-fit measure is the regression variance $\hat{\sigma}^2_T = \hat{e}'\hat{e}/(T - k)$. This measure, however, is not invariant with respect to measurement units of the dependent variable.
Measures of Goodness of Fit: Relative Measure

Recall that

$$\sum_{t=1}^{T} y_t^2 = \sum_{t=1}^{T} \hat{y}_t^2 + \sum_{t=1}^{T} \hat{e}_t^2.$$ 

where TSS, RSS, and ESS denote total, regression, and error sum of squares, respectively. The non-centered coefficient of determination (or non-centered $R^2$) is defined to be the proportion of TSS that can be explained by the regression hyperplane:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}.$$  

(4)
Clearly, $0 \leq R^2 \leq 1$, and the larger the $R^2$, the better the model fits the data. In particular, a model has a perfect fit if $R^2 = 1$, and it does not account for any variation of $y$ if $R^2 = 0$. Note that $R^2$ is non-decreasing in the number of variables in the model.

As $\hat{y}'\hat{y} = \hat{y}'y$, we can also write

$$R^2 = \frac{\hat{y}'\hat{y}}{y'y} = \frac{(\hat{y}'y)^2}{(y'y)(\hat{y}'\hat{y})} = \cos^2 \theta,$$

where $\theta$ is the angle between $y$ and $\hat{y}$. That is, $R^2$ is a measure of the linear association between these two vectors.
Centered $R^2$

It is also easily verified that, when the model contains a constant term,

$$\sum_{t=1}^{T} (y_t - \bar{y})^2 = \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T} \hat{e}_t^2,$$

where $\bar{y} = \tilde{y} = \frac{\sum_{t=1}^{T} y_t}{T}$. Analogous to (4), the centered coefficient of determination (or centered $R^2$) is defined as

$$\text{Centered } R^2 = \frac{\text{Centered RSS}}{\text{Centered TSS}} = 1 - \frac{\text{ESS}}{\text{Centered TSS}}.$$  (5)

This measure also takes on values between 0 and 1 and is non-decreasing in the number of variables in the model.
If the model does not contain a constant term, the centered $R^2$ may be negative. As

$$\sum_{t=1}^{T} (y_t - \bar{y})(\hat{y}_t - \bar{y}) = \sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2,$$

we immediately get

$$\frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} = \frac{\left[\sum_{t=1}^{T} (y_t - \bar{y})(\hat{y}_t - \bar{y})\right]^2}{\left[\sum_{t=1}^{T} (y_t - \bar{y})^2\right]\left[\sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2\right]}.$$

That is, the centered $R^2$ is also the squared sample correlation coefficient of $y$ and $\hat{y}$. 
Adjusted $R^2$

The adjusted $R^2$, $\bar{R}^2$, is the centered $R^2$ adjusted for the degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\hat{e}'\hat{e}/(T - k)}{(y'y - T\bar{y}^2)/(T - 1)}.$$ 

It can also be shown that

$$\bar{R}^2 = 1 - \frac{T - 1}{T - k} (1 - R^2) = R^2 - \frac{k - 1}{T - k} (1 - R^2).$$

That is, $\bar{R}^2$ is the centered $R^2$ with a penalty term depending on model complexity and explanatory ability. Clearly, $\bar{R}^2 < R^2$ except for $k = 1$ or $R^2 = 1$. Note also that $\bar{R}^2$ need not be increasing with the number of explanatory variables; in fact, $\bar{R}^2 < 0$ when $R^2 < (k - 1)/(T - 1)$. 

M.-Y. Chen Econometrics
Properties of the OLS Estimators: Unbias

As

\[ \hat{\beta}_T = (X'X)^{-1}X'y \]
\[ = (X'X)^{-1}X'(X\beta_0 + e) \]
\[ = (X'X)^{-1}X'X\beta_0 + (X'X)^{-1}X'e \]
\[ = \beta_0 + (X'X)^{-1}X'e, \]

then to have \( \hat{\beta}_T \) to be unbiased for \( \beta_0 \), the following assumption has to be imposed:

**A3**: \( E(e|X) = 0 \).

That is, given **ID1**, **A2** and **A3**, \( E(\hat{\beta}_T - \beta_0|X) = 0 \) and \( E(\hat{\beta}_T) = \beta_0 \).
Variance-Covariance Matrix of Regression Error

The conditional variance-covariance matrix of the regression error vector $e$ is

$$D = \text{var}(e|X) = E(ee'|X)$$

$$= \begin{bmatrix}
E(e_1^2|x_1) & E(e_1e_2|x_1) & E(e_1e_3|x_1) & \cdots & E(e_1e_T|x_1) \\
E(e_2e_1|x_2) & E(e_2^2|x_2) & E(e_2e_3|x_2) & \cdots & E(e_2e_T|x_2) \\
E(e_3e_1|x_3) & E(e_3e_2|x_3) & E(e_3^2|x_3) & \cdots & E(e_3e_T|x_3) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
E(e_Te_1|x_T) & E(e_Te_2|x_T) & E(e_Te_3|x_T) & \cdots & E(e_T^2|x_T)
\end{bmatrix}$$

when the data are random sample then $(x_t, e_t)$ is independent of $(x_s, e_s)$ for $t \neq s$, thus

$$E(e_t^2|X) = E(e_t^2|x_t) = \sigma_t^2$$

$$E(e_te_s|X) = E(e_te_s|x_t) = E(e_t|x_t)E(e_s|x_t) = 0.$$
Thus in general

\[ D = \text{var}(e|X) \]  

\[ = \begin{bmatrix}
\sigma_1^2 & 0 & 0 & \cdots & 0 \\
0 & \sigma_2^2 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sigma_T^2
\end{bmatrix} \]  

when the data are random. Under the homoskedasticity restriction, 

\[ E(e_t^2|x_t) = \sigma_0^2 \text{ for all } t, \] 

then

\[ D = \begin{bmatrix}
\sigma_0^2 & 0 & 0 & \cdots & 0 \\
0 & \sigma_0^2 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sigma_0^2
\end{bmatrix} = \sigma_0^2 I_T, \]  

which is the classical assumption for the linear regression models. That is

**A4**: \( \text{var}(e|X) = \sigma_0^2 I_T \).
Variance-Covariance Matrix of OLS Estimator

The conditional variance-covariance matrix for $\hat{\beta}_T$ is

$$V_T = E \left[ (\hat{\beta}_T - \beta_0)(\hat{\beta}_T - \beta_0)' | X \right]$$

Since $\hat{\beta}_T - \beta_0 = (X'X)^{-1}X'e$,

$$V_T = E \left[ (X'X)^{-1}X'e e'X (X'X)^{-1} | X \right]$$

$$= (X'X)^{-1}X'E[ee' | X]X(X'X)^{-1}$$

$$= (X'X)^{-1}X'DX(X'X)^{-1},$$

where $D$ is defined in (7).
It may be helpful to observe that

\[ X' DX = \sum_{t=1}^{T} x_t x'_t \sigma_t^2. \]

If \( E(e_t^2|x_t) = \sigma_0^2 \) holds, then \( \sigma_t^2 = \sigma_0^2 \), \( D = \sigma_0^2 I_T \) and \( X' DX = X' X \sigma_0^2 \). Thus, \( V_T \) simplifies to

\[
V_T = (X' X)^{-1} X' X \sigma_0^2 (X' X)^{-1}
\]

\[ = \sigma_0^2 (X' X)^{-1}. \]
In the linear regression model,

\[ V_T = (X'X)^{-1}X'DX(X'X)^{-1}. \]  \hspace{2cm} (9)

If \( E(e_t^2|x_t) = \sigma_0^2 \) holds,

\[ V_T = \sigma_0^2(X'X)^{-1}. \]  \hspace{2cm} (10)

The expression \( V_T = (X'X)^{-1}X'DX(X'X)^{-1} \) is often called a “sandwich formula”, because the central variance matrix \( X'DX \) is “sandwiched” between the moment matrices \( (X'X)^{-1} \).
Gauss-Markov Theorem

(Gauss-Markov) In the linear regression model, $\hat{\beta}_T$ is the best linear unbiased estimator for $\beta_0$. Consider an arbitrary linear estimator

$$\tilde{\beta}_T = A y = \left( (X' X)^{-1} X' + C \right) y,$$

where $C$ is an arbitrary non-zero matrix. $\tilde{\beta}_T$ is unbiased if and only if $CX = 0$ since

$$\tilde{\beta}_T = \left( (X' X)^{-1} X' + C \right) y = \left( (X' X)^{-1} X' + C \right) (X \beta_0 + e) = \beta_0 + (X' X)^{-1} X' e + CX \beta_0 + Ce.$$ 

and

$$E[\tilde{\beta}_T | X] = \beta_0 + E[(X' X)^{-1} X' e | X] + E[CX \beta_0 | X] + E[Ce | X] = \beta_0 + CX \beta_0.$$
It follows that when $\hat{\beta}_T$ is unbiased

\[
\text{var}(\hat{\beta}_T|X) = E[(\hat{\beta}_T - \beta_0)(\hat{\beta}_T - \beta_0)'|X]
\]

\[
= E\{[(X'X)^{-1}X'e + Ce][(X'X)^{-1}X'e + Ce]'|X\}
\]

\[
= (X'X)^{-1}X'\sigma_0^2I_TX(X'X)^{-1} + (X'X)^{-1}X'C'
\]

\[
+ CX(X'X)^{-1} + C\sigma_0^2I_T C'
\]

\[
= \sigma_0^2(X'X)^{-1} + \sigma_0^2 CC',
\]

where the first term on the right-hand side is $\text{var}(\hat{\beta}_T)$ and the second term is clearly p.s.d. Thus, for any linear unbiased estimator $\hat{\beta}_T$, $\text{var}(\hat{\beta}_T) - \text{var}(\hat{\beta}_T)$ is a p.s.d. matrix.
OLS Estimator of Error Variance

Under the restriction of homoscedastic errors, $E(e_t^2 | x_t) = \sigma_0^2$ is another parameter under estimation. The OLS estimator for $\sigma_0^2$ is

$$\hat{\sigma}_T^2 = \frac{\hat{e}'\hat{e}}{T - k} = \frac{1}{T - k} \sum_{t=1}^{T} \hat{e}_t^2,$$

where $k$ is the number of regressors. It is clear that $\hat{\sigma}_T^2$ is not linear in $y$.

In the homoskedastic regression model, $\hat{\sigma}_T^2$ is an unbiased estimator for $\sigma_0^2$. 
Proof: Recall that $I - P_X$ is orthogonal to span($X$). Then,

$$\hat{e} = (I_T - P_X)y = (I_T - P_X)(X\beta_0 + e) = (I_T - P_X)e,$$

and

$$E(\hat{e}'\hat{e}|X) = E[e'(I_T - P_X)e|X] = E[\text{trace}(ee'(I_T - P_X))|X].$$

As the trace and expectation operators can be interchanged, we have that

$$E(\hat{e}'\hat{e}|X) = \text{trace}(E[ee'(I_T - P_X)|X]) = \text{trace}[D(I_T - P_X)].$$

By the fact that, $\text{trace}(I_T - P_X) = \text{rank}(I_T - P_X) = T - k$ and $D = \sigma^2 I_T$, it follows that $E(\hat{e}'\hat{e}) = (T - k)\sigma^2_0$ and that

$$E(\hat{\sigma}^2_T) = E(\hat{e}'\hat{e})/(T - k) = \sigma^2_0,$$

proving the unbiasedness of $\hat{\sigma}^2_T$. 
The OLS estimation for variance-covariance matrix of $\hat{\beta}_T$ in the homoskedastic regression model becomes

$$\text{var}(\hat{\beta}_T) = \hat{\sigma}_T^2 (X'X)^{-1}$$

which is unbiased for $\text{var}(\hat{\beta}_T) = \sigma_0^2 (X'X)^{-1}$ provided $\hat{\sigma}_T^2$ is unbiased for $\sigma_0^2$. 

Gaussian Quasi-MLE

In normal regression, $e_t | x_t \sim N(0, \sigma^2)$ and then the likelihood for a single observation is

$$L_t(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left( -\frac{1}{2\sigma^2} (y_t - x'_t \beta)^2 \right).$$

Then the log-likelihood function for the full sample $y^T = (y_1, \ldots, y_T)$ is

$$L_T(y^T; \beta, \sigma^2) = \sum_{t=1}^{T} \log L_t(y_t; \beta, \sigma^2)$$

$$= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta).$$
The MLE \((\tilde{\beta}_T, \tilde{\sigma}_T^2)\) maximize \(L_T\). The FOCs of the maximization problem are

\[
\frac{\nabla L_T}{\nabla \beta} = \frac{X'(y - X\beta)}{\sigma^2} \equiv 0,
\]

\[
\frac{\partial L_T}{\partial (\sigma^2)} = -\frac{T}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4} \equiv 0,
\]

which yield the MLEs of \(\beta_0\) and \(\sigma^2\):

\[
\tilde{\beta}_T = (X'X)^{-1}X'y,
\]

\[
\tilde{\sigma}_T^2 = \frac{(y - X\tilde{\beta}_T)'(y - X\tilde{\beta}_T)}{T} = \hat{e}'\hat{e}/T.
\]

Clearly, the MLE \(\tilde{\beta}_T\) is the same as the OLS estimator \(\hat{\beta}_T\), but the MLE \(\tilde{\sigma}_T^2\) is different from \(\hat{\sigma}_T^2\). In fact, \(\tilde{\sigma}_T^2\) is biased estimator because 

\[
E(\tilde{\sigma}_T^2) = \sigma_0^2(T - k)/T \neq \sigma_0^2.
\]
In normal regression models, the OLS estimators $\hat{\beta}_T$ and $\hat{\sigma}^2_T$ are the minimum variance unbiased estimator (MVUE).
Consider a collection of independent random variables $z^T = (z_1, \ldots, z_T)$, where $z_t$ has the density function $f_t(z_t, \theta)$ with $\theta$ a $r \times 1$ vector of parameters. Let the joint log-likelihood function of $z^T$ be 

$$L_T(z^T; \theta) = \log f^T(z^T; \theta).$$

Then the score function

$$s^T(z^T; \theta) := \nabla \log f^T(z^T; \theta)$$

$$= \frac{1}{f^T(z^T; \theta)} \nabla f^T(z^T; \theta)$$

is the $r \times 1$ vector of the first order derivatives of $\log f^T$ with respect to $\theta$. Under regularity conditions, differentiation and integration can be interchanged. When the postulated density function $f^T$ is the true density function of $z^T$, we have

$$E[s^T(z^T; \theta)] = \int \frac{1}{f^T(z^T; \theta)} \nabla f^T(z^T; \theta) f^T(z^T; \theta) \, dz^T$$

$$= \nabla \left( \int f^T(z^T; \theta) \, dz^T \right)$$

$$= 0.$$
That is, \( s^T(z^T; \theta) \) has mean zero. The variance of \( s^T \) is the Fisher’s information matrix:

\[
B_T(\theta) := \text{var}[s^T(z^T; \theta)] = E[s^T(z^T; \theta) s^T(z^T; \theta)'].
\]

Consider the \( r \times r \) Hessian matrix of the second order derivatives of \( \log f^T \):

\[
H_T(z^T; \theta) = \nabla^2 \log f^T(z^T; \theta) = \nabla \left( \frac{1}{f^T(z^T; \theta)} [\nabla f^T(z^T; \theta)]' \right) = \frac{1}{f^T(z^T; \theta)} \nabla^2 f^T(z^T; \theta) - \frac{1}{f^T(z^T; \theta)^2} [\nabla f^T(z^T; \theta)][\nabla f^T(z^T; \theta)]',
\]

where \( \nabla^2 f = \nabla(\nabla f)' \).
As
\[
\int \frac{1}{f^T(z^T; \theta)} \nabla^2 f^T(z^T; \theta) f^T(z^T; \theta) dz^T = \nabla^2 \left( \int f^T(z^T; \theta) dz^T \right) = 0,
\]
the expected value of the Hessian matrix becomes
\[
E[H_T(z^T; \theta)] = -\int \left( \frac{1}{f^T(z^T; \theta)^2} [\nabla f^T(z^T; \theta)][\nabla f^T(z^T; \theta)]' \right) f^T(z^T; \theta) dz^T
\]
\[
= E[s^T(z^T; \theta) s^T(z^T; \theta)]
\]
\[
= -B_t(\theta).
\]
This established the information matrix equality:

\[ B_T(\theta) + E[H_T(z^T; \theta)] = 0. \]

Suppose now \( r = 1 \) for simplicity so that both \( s^T \) and \( B_T \) are scalar. Let \( \hat{\theta}_T \) denote an unbiased estimator. Then

\[
\text{cov}[s^T(z^T; \theta), \hat{\theta}] = \frac{\partial}{\partial \theta} \int \hat{\theta}_T f^T(z^T; \theta) \, dz^T = \frac{\partial}{\partial \theta} E(\hat{\theta}_T) = 1,
\]

by unbiasedness. By the celebrated Cauchy-Schwartz inequality:

\[
\frac{\text{cov}[s^T(z^T; \theta), \hat{\theta}_T]}{\text{var}[s^T(z^T; \theta)] \text{var}(\hat{\theta}_T)} = \frac{1}{\text{var}[s^T(z^T; \theta)] \text{var}(\hat{\theta}_T)} \leq 1.
\]

It follows that \( \text{var}(\hat{\theta}_T) \geq 1/B_T(\theta) \). The RHS, \( 1/B_T(\theta) \), is also known as the Cramér-Rao lower bound. Thus, all unbiased estimators must have variance greater than or equal to the inverse of information. When \( \theta \) is multi-dimensional, we have that \( \text{var}(\hat{\theta}_T) - B_T(\theta)^{-1} \) is a positive semi-definite matrix.
In our application, the inverse of the information matrix evaluated at the true parameters $\beta_0$ and $\sigma^2_0$ can be easily calculated as

$$\begin{bmatrix}
\sigma^2_0(X'X)^{-1} & 0 \\
0 & 2\sigma^4_0/T \\
\end{bmatrix}.$$

As $\hat{\beta}_T$ achieves the Cramér-Rao lower bound, it is efficient within the class of all unbiased estimators for $\beta_0$, i.e., it is the MVUE. It can be shown that any other unbiased estimator of $\sigma^2_0$ has variance greater than or equal to that of $\hat{\sigma}^2_T$; hence $\hat{\sigma}^2_T$ is also the MVUE.
Consider the log-likelihood function of \( \mathbf{y} = (y_1, \ldots, y_T)' \) to be

\[
L_T(\mathbf{y}; \beta, \sigma^2) = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).
\]

The corresponding Hessian matrix is

\[
H_T(\beta, \sigma^2) = \begin{bmatrix}
-\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X}) & -\frac{1}{\sigma^4}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta) \\
-\frac{1}{\sigma^4}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta) & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)
\end{bmatrix}.
\]
The information matrix is defined as $B_T(\beta, \sigma^2) = -E[H_T(\beta, \sigma^2)]$.

The $(1, 2)$-th element of information matrix evaluated at true values $\beta_0$ and $\sigma_0$ becomes

$$E\left[\frac{1}{\sigma_0^4}(X'y - X'X\beta_0)\right] = \frac{1}{\sigma_0^4}E(X'y - X'X\beta) = \frac{1}{\sigma_0^4}(X'X)E(e) = 0.$$ 

And, the $(2, 2)$-th element of $B_T(\beta, \sigma^2)$ evaluated at $\beta_0$ and $\sigma_0^2$ is

$$-E\left(\frac{T}{2\sigma_0^4} - \frac{1}{\sigma_0^6}(y - X\beta_0)'(y - X\beta_0)\right)$$

$$= -\frac{T}{2\sigma_0^4} + \frac{1}{\sigma_0^6}E(e'e)$$

$$= -\frac{T}{2\sigma_0^4} + \frac{T\sigma_0^2}{\sigma_0^6}$$

$$= \frac{-T + 2T}{2\sigma_0^4} = \frac{T}{2\sigma_0^4}. $$
Sampling Distribution of $\hat{\beta}_T$ in Normal Regression Models

In normal regression models,

$$y_t = \mathbf{x}_t' \beta_0 + e_t$$

$$e_t \mid x_t \sim N(0, \sigma_0^2).$$

In normal regression models,

(a) $\hat{\beta}_T \sim N(\beta_0, \sigma_0^2 (X'X)^{-1}).$

(b) $(T - k)\hat{\sigma}_T^2 / \sigma_0^2 \sim \chi^2(T - k).$

(c) $\hat{\sigma}_T^2$ has mean $\sigma_0^2$ and variance $2\sigma_0^4 / (T - k).$
Consistency and Asymptotic Normality of OLS Estimators

Under the linear projection model:

\[ y_t = x_t \beta_0 + e_t \]

\[ E(x_t e_t) = 0, \]

that is, \( x \beta_0 \) is the projection of \( y \) on the linear space of \( x \). Given \( E(x_t e_t) = 0 \), \( x_t \) may be stochastic, but it must be uncorrelated with \( e_t \) and \( e_t \) is not required being heteroskedastic.
Consider an AR\((p)\) model for \(y_t\):
\[
y_t = c + \psi_1 y_{t-1} + \cdots + \psi_p y_{t-p} + e_t. \tag{11}
\]

It can be seen that by recursive substitution, \(y_t\) is a linear function of the current and past \(e_t\). If \(\{e_t\}\) is a white noise, then for \(x_t = (1 \ y_{t-1} \ldots y_{t-p})'\), we have
\[
E(x_t e_t) = 0,
\]
by the white noise property of \(\{e_t\}\). Thus, an AR\((p)\) model with \(\{e_t\}\) being a white noise satisfies \(E(x_t e_t) = 0\).
On the other hand, suppose that \( \{e_t\} \) is an MA\((q)\) process:

\[
e_t = u_t - \phi_1 u_{t-1} - \cdots - \phi_q u_{t-q},
\]

where \( \{u_t\} \) is a white noise with mean zero and variance \( \sigma_u^2 \). In this case, \( y_t \) is known as an ARMA\((p,q)\) process. Note that

\[
E(e_t e_{t-i}) = -(\phi_i - \phi_{i+1} \phi_{i-1} - \cdots - \phi_q \phi_{q-i}) \sigma_u^2, \quad i = 1, \ldots, q,
\]

with \( \phi_0 = 1 \) and \( \phi_j = 0 \) if \( j < 0 \). That is, \( e_t \) and \( e_{t-i} \) are correlated for \( i = 1, \ldots, q \). It follows that

\[
E(x_t e_t) \neq 0.
\]
Consistency and Asymptotic Normality of OLS Estimators

Suppose that the following conditions hold:

[B1] \( y_t = x_t' \beta_0 + e_t \) such that \( E(x_t e_t) = 0 \) for \( t = 1, \ldots, T \),

[B2] \{x_t x_t'\} obeys a SLLN (WLLN) such that \( M_T \) are p.d. and for some \( \delta > 0 \), \( \text{det}(M_T) > \delta \) for all \( T \) sufficiently large,

[B3] \{x_t e_t\} obeys a SLLN (WLLN).

Then \( \hat{\beta}_T \) exists a.s. (in probability) for all \( T \) sufficiently large and \( \hat{\beta}_T \rightarrow \beta_0 \) a.s. (in probability).
As $\hat{\beta}_T = (X'X)^{-1}X'y$, $(X'X)$ is a stochastic matrix so that existence of $(X'X)^{-1}$ can not be assumed. It is known that

$$\hat{\beta}_T = (X'X)^{-1}X'y$$

$$= (X'X)^{-1}X'(X\beta_0 + e)$$

$$= (X'X)^{-1}X'X\beta_0 + \sqrt{T}(X'X)^{-1}X'e$$

$$= \left(\frac{X'X}{T}\right)^{-1} \left(\frac{X'X}{T}\right) \beta_0 + \left(\frac{X'X}{T}\right)^{-1} \left(\frac{X'e}{T}\right).$$
\[
\frac{X' e}{\sqrt{T}} = \begin{bmatrix}
\frac{\sum_{t=1}^{T} x_{t1} e_t}{T} \\
\frac{\sum_{t=1}^{T} x_{t2} e_t}{T} \\
\vdots \\
\frac{\sum_{t=1}^{T} x_{tk} e_t}{T}
\end{bmatrix} \rightarrow \begin{bmatrix}
\frac{\sum_{t=1}^{T} E(x_{t1} e_t)}{T} \\
\frac{\sum_{t=1}^{T} E(x_{t2} e_t)}{T} \\
\vdots \\
\frac{\sum_{t=1}^{T} E(x_{tk} e_t)}{T}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
\vdots \\
0
\end{bmatrix}.
\]
Consistency and Asymptotic Normality of OLS Estimators

Given Assumptions [B1], suppose that the following conditions hold:

[B2'] \{x_t x_t'\} obeys a WLLN such that \(M_T\) are p.d. and for some \(\delta > 0\), \(\text{det}(M_T) > \delta\) for all \(T\) sufficiently large, and

[B3'] \{x_t e_t\} obeys a CLT such that \(\Xi_T = O(1)\) is p.d. and for some \(\delta > 0\), \(\text{det}(\Xi_T) > \delta\) for all \(T\) sufficiently large.

Then, \(\Sigma_T^{-1/2} \sqrt{T}(\hat{\beta}_T - \beta_0) \xrightarrow{D} N(0, I_k)\), where \(\Sigma_T = M_T^{-1} \Xi_T M_T^{-1}\). If, in addition, there exists a \(\hat{\Xi}_T\) p.s.d. and symmetric such that \(\hat{\Xi}_T - \Xi_T \xrightarrow{P} 0\), then \(\hat{\Sigma}_T - \Sigma_T \xrightarrow{P} 0\), and

\[\hat{\Sigma}_T^{-1/2} \sqrt{T}(\hat{\beta}_T - \beta_0) \xrightarrow{D} N(0, I_k),\]

where \(\hat{\Sigma}_T = (\sum_{t=1}^T x_t x_t' / T)^{-1} \hat{\Xi}_T (\sum_{t=1}^T x_t x_t' / T)^{-1}\).
\[ \sqrt{T} \hat{\beta}_T = (X'X)^{-1} X'y \]
\[ = \sqrt{T}(X'X)^{-1} X'(X\beta_0 + e) \]
\[ = \sqrt{T}(X'X)^{-1} X'X\beta_0 + \sqrt{T}(X'X)^{-1} X'e \]
\[ = \sqrt{T} \left( \frac{X'X}{T} \right)^{-1} \left( \frac{X'X}{T} \right) \beta_0 + \left( \frac{X'X}{T} \right)^{-1} \left( \frac{X'e}{\sqrt{T}} \right). \]