Topic 1. The linear model and instrumental variables estimators

Consider a linear model,

\[ Y_n = X_n \beta_0 + \varepsilon_n, \quad (1.1) \]

where \( Y_n = (Y_1, \ldots, Y_n)' \) is an observable output vector, \( X_n = (X_i)_{i=1}^n \) is an observable input matrix, \( \beta_0 \in \mathbb{R}^k \) is an unknown vector, and \( \varepsilon_n = (\varepsilon_1, \ldots, \varepsilon_n)' \) is an unobservable random vector with \( E(\varepsilon_n) = 0_n = (0, \ldots, 0)' \) and \( \text{cov}(\varepsilon_n) = E(\varepsilon_n \varepsilon_n') = \Omega \). In the following, the \( n \)'s in \( Y_n, X_n \) and \( \varepsilon_n \) are suppressed in order to simplify the notations. It is a common practice to estimate \( \beta_0 \) using the least squares estimator \( \hat{\beta}_n \), where

\[ \hat{\beta}_n = (X'X)^{-1}X'Y = (\sum_{j=1}^n x_j x_j')^{-1} \sum_{j=1}^n x_j y_j, \quad (1.2) \]

with \( x_j = (x_{j1}, \ldots, x_{jk})' \) being the transpose of the \( j \)th row of \( X \).

Under some "standard" assumptions, Theorem 1.1 investigates the finite-sample properties of \( \hat{\beta}_n \), which hold for any \( n \).
Theorem 1.1. Assume model (1.1). If

(i) $X$ is nonstochastic;
(ii) $XX$ is nonsingular (i.e., $X\alpha \neq 0$, for any $\alpha \neq 0$);
(iii) $\Omega = \sigma^2 I_{nn}$ with $\sigma^2 > 0$ and $I_{nn}$ being the $nn$ identity matrix,

then

(a) $\hat{\beta}_n$ exists and is unique in the sense that

$$\sum_{j=1}^{n} (Y_j - \hat{\beta}'_n X_j)^2 > \sum_{j=1}^{n} (Y_j - \beta_n X_j)^2$$

for any $\beta \neq \hat{\beta}_n$;

(b) $E(\hat{\beta}_n) = \beta_0$ (unbiasedness);

(c) $\hat{\beta}_n$ is the best linear unbiased estimator (BLUE) of $\beta_0$.

Moreover, if

(iv) $\varepsilon \sim N(0, \sigma^2 I_{nn})$

also holds, then

(d) $\hat{\beta}_n \sim N(\beta_0, (XX)^{-1}\sigma^2)$

(e) $\hat{\beta}_n$ is the maximum likelihood estimator (MLE) of $\beta_0$. 
proof.

(a) Since \( XX \) is nonsingular, in view of (1.2), \( \hat{\beta}_n \) obviously exists. To show \( \hat{\beta}_n \) is unique, note that

\[
\sum_{j=1}^{n} (y_j - \hat{\beta}_n x_j)^2 = \sum_{j=1}^{n} (y_j - \hat{\beta}_n x_j + (\hat{\beta}_n - \beta) x_j)^2
\]

\[
= \sum_{j=1}^{n} (y_j - \hat{\beta}_n x_j)^2 + 2(\hat{\beta}_n - \beta) \sum_{j=1}^{n} x_j (y_j - \hat{\beta}_n x_j) + (\hat{\beta}_n - \beta)^2 \sum_{j=1}^{n} x_j^2
\]

(1.3)

Since

\[
\sum_{j=1}^{n} x_j (y_j - \hat{\beta}_n x_j) = 0 \quad \text{(why?)},
\]

(1.4)

and for any \( \alpha \in \mathbb{R}^k \) and \( \alpha \neq 0 \),

\[
\alpha^T \sum_{j=1}^{n} x_j x_j^T \alpha = \alpha^T X^T X \alpha > 0,
\]

by (1.3)–(1.5), for any \( \beta \neq \hat{\beta}_n \),

\[
\sum_{j=1}^{n} (y_j - \beta x_j)^2 > \sum_{j=1}^{n} (y_j - \hat{\beta}_n x_j)^2
\]

which gives the desired result.

(b) \( E(\hat{\beta}_n) = (XX^{-1}) X^T EY = (XX^{-1}) X^T \beta_0 = \beta_0 \).
(c) Let \( \hat{\beta}_n \) be a linear unbiased estimator of \( \beta_0 \). Then, 
\[ \hat{\beta}_n = GY, \]
where \( G \) is a known matrix, and 
\[ E(\hat{\beta}_n) = \beta_0, \]
which yields 
\[ E(G(X\beta_0 + \epsilon)) = GX\beta_0 = \beta_0. \]
Since \( \beta_0 \) can be any vector in \( \mathbb{R}^k \), we have 
\[ G^T G = I \]
(1.6)

Now, for any \( a \in \mathbb{R}^k \),
\[ \text{Var}(a^T \hat{\beta}_n) = E((a^T \hat{\beta}_n - a^T \beta_0)^2) \]
\[ = E(a^T (\hat{\beta}_n - \beta_0) + a^T (\hat{\beta}_n - \beta_0)^2) \]
\[ = E(a^T (\hat{\beta}_n - \beta_0)^2) + 2a^T E((\hat{\beta}_n - \beta_0)(\hat{\beta}_n - \beta_0)^T) a + \text{Var}(a^T \hat{\beta}_n). \]

Notice that
\[ \hat{\beta}_n - \beta_0 = (X^T X)^{-1} X^T \epsilon, \]
and
\[ \hat{\beta}_n - \beta_n = (G - (XX^T)X^T)Y. \]

These and (1.6) imply
\[ E((\hat{\beta}_n - \beta_n)(\hat{\beta}_n - \beta_0)^T) = (G - (XX^T)X^T)E(\epsilon \epsilon^T)X(X^T)^{-1} \]
\[ = 0 \cdot (G - (XX^T)X^T)X(X^T)^{-1} = 0_{k \times k}. \]
In view of (1.7) and (1.8),
\[ \text{Var}(a' \beta_n) = \text{Var}(a' \beta_n), \]
which guarantees that \( \beta_n \) is the BLUE.

(d) When \( E \sim N(0_n, \sigma^2 I_n) \) is assumed,
\[ \hat{\beta}_n = \beta_0 + (X'X)^{-1} X' E, \]
as a linear transformation of \( E \), is also normally distributed with mean vector \( \beta_0 \) and covariance matrix \( \text{Cov}(X'(X')^{-1}X E) = \sigma^2 (X'X)^{-1} \). Hence, the claimed result follows.

(e) If \( E \sim N(0_n, \sigma^2 I_n) \), then it is not difficult to see that \( Y_i \)'s are independent and
\[ N( \beta_0 X_i, \sigma^2 ) \] distributed. Therefore, the joint probability density function of \((Y_1, \ldots, Y_n)\) is given by
\[ f_{\beta_0, \sigma^2}(Y_1, \ldots, Y_n) = (2\pi \sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 X_i)^2 \right\}, \]
and hence the corresponding log-likelihood function is
\[
\log L (\beta_0, \sigma_0^2) = \log f_{\beta_0, \sigma_0^2} (Y_1, \ldots, Y_n)
\]

\[
= -\frac{n}{2} \log \sigma_0^2 - \frac{1}{2 \sigma_0^2} \sum_{j=1}^{n} (Y_j - (\beta_0 X_j)^2)
\]

Now,

\[
\frac{\partial}{\partial \beta_0} \log L (\beta_0, \sigma_0^2) = -\frac{1}{2 \sigma_0^2} \sum_{j=1}^{n} Y_j^2 - 2 \left( \sum_{j=1}^{n} X_j Y_j \right) \beta_0
\]

\[
- \left( \frac{1}{2 \sigma_0^2} \sum_{j=1}^{n} X_j^2 \right) \beta_0
\]

(Note that)

\[
\frac{\partial}{\partial (x_1, \ldots, x_k)} \left( \begin{array}{c} f_1(x_1, \ldots, x_k) \\ \vdots \\ f_p(x_1, \ldots, x_k) \end{array} \right) = \left( \begin{array}{c} \frac{\partial f_1}{\partial x_1} \\ \vdots \\ \frac{\partial f_p}{\partial x_k} \end{array} \right)
\]

Therefore,

\[
\frac{\partial A x}{\partial x} = A'.
\]

In addition,

\[
\frac{\partial g(x_1, \ldots, x_k)}{\partial (x_1, \ldots, x_k)} = \left( \begin{array}{c} \frac{\partial g(x_1, \ldots, x_k)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x_1, \ldots, x_k)}{\partial x_k} \end{array} \right)
\]

Therefore,

\[
\frac{\partial x' A x}{\partial x} = \begin{cases} 2 A x & \text{if } A \text{ is symmetric}, \\ (A + A') x & \text{if } A \text{ is not symmetric}. \end{cases}
\]

Let \( \frac{\partial}{\partial \beta_0} \log L (\beta_0, \sigma_0^2) = 0 \). We get

\[
\beta_0 = \hat{\beta}_n. \quad (1.9)
\]
Moreover, \[ \frac{d}{d \sigma^2} \log L (\beta_0, \sigma^2) = -\frac{1}{2} \frac{1}{\sigma^2} + \frac{1}{2} \left( \frac{1}{\sigma^2} \right) \sum_{j=1}^{n} (y_j - (\beta_0 x_j)^2) = 0 \]

implies
\[ \sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - (\beta_0 x_j)^2) \quad \text{(1.10)} \]

Equations (1.9) and (1.10) show that the solution for
\[ \frac{d}{d (\beta_0, \sigma^2)} \log L (\beta_0, \sigma^2) = 0 \]

is
\[ (\beta_0, \sigma^2) = (\hat{\beta}_n, \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{\beta}_n x_j)^2) \]

To show \((\hat{\beta}_n, \hat{\sigma}^2)\) really maximizes \(\log L (\beta_0, \sigma^2)\),
one need to check whether
\[ \frac{d^2}{d (\beta_0, \sigma^2)} \log L (\beta_0, \sigma^2) \]
is negative definite, that is, for any \(a \in \mathbb{R}^{k+1}\)
\[ a^T \frac{d^2}{d (\beta_0, \sigma^2)} \log L (\beta_0, \sigma^2) a < 0 \quad \text{(1.11)} \]
The proof of (1.11) is left as an exercise.
Remark 1. In fact, it can be shown that

\( \hat{\beta}_n \) is the best unbiased estimator in the sense that for any unbiased \( \tilde{\beta}_n \), \( \text{Cov}(\hat{\beta}_n) - \text{Cov}(\tilde{\beta}_n) \) is negative definite. To illustrate this result, consider the simplest case where \( \beta_0 \) is 1-dimensional and \( \sigma_0^2 \) is known. Let \( \tilde{\beta}_n \) be any unbiased estimator of \( \beta_0 \). Then, \( E(\tilde{\beta}_n) = \beta_0 \). Now,

\[
1 = \frac{d}{d \beta_0} \int \tilde{\beta}_n f(\beta_0, \ldots, \beta_n) \, d\beta_1 \ldots d\beta_n
\]

\[
= \int \tilde{\beta}_n \frac{\partial}{\partial \beta_0} f_\beta_0(\beta_1, \ldots, \beta_n) \, d\beta_1 \ldots d\beta_n
\]

\[
= \int \tilde{\beta}_n \frac{1}{\partial \beta_0} \frac{\log f_\beta_0(\beta_1, \ldots, \beta_n)}{\partial \beta_0} f_\beta_0(\beta_1, \ldots, \beta_n) \, d\beta_1 \ldots d\beta_n
\]

\[
= E(\tilde{\beta}_n \frac{\partial}{\partial \beta_0} \log f_\beta_0(\beta_1, \ldots, \beta_n))
\]

Why? \[
= E(\tilde{\beta}_n - \beta_0) \frac{\partial}{\partial \beta_0} \log f_\beta_0(\beta_1, \ldots, \beta_n)
\]

By the Cauchy-Schwarz inequality,

\[
1 \geq \left( E(\tilde{\beta}_n - \beta_0) \frac{\partial}{\partial \beta_0} \log f_\beta_0(\beta_1, \ldots, \beta_n) \right)^2
\]

\[
= \text{Var}(\tilde{\beta}_n) E\left\{ \left( \frac{\partial}{\partial \beta_0} \log f_\beta_0(\beta_1, \ldots, \beta_n) \right)^2 \right\}
\]
Therefore,

$$\text{Var}(\tilde{\beta}_n) \leq \frac{1}{E\left\{ \left( \frac{d}{d\beta_0} \log f_0(x_i - x_n) \right)^2 \right\}}$$

since

$$E\left\{ \left( \frac{d}{d\beta_0} \log f_0(x_i - x_n) \right)^2 \right\} = \frac{1}{\delta_0^2} \sum_{j=1}^{n} \xi_j^2 = \text{Var}(\tilde{\beta}_n)$$

(why?), we obtain

$$\text{Var}(\tilde{\beta}_n) \geq \text{Var}(\hat{\beta}_n).$$

Suppose that $\Sigma$ exhibits heteroskedasticity or serial correlation, so that $\Sigma \neq \delta_0^2 I$. The following theorem shows that when $\Sigma$ is known, the generalized least squares (GLS) estimator, $\hat{\beta}_n, w = (X^{\prime}X)^{-1}X^{\prime}w^{\prime}Y$, has some nice properties.

Theorem 1.2. Assume model (1.1). If (i) and (ii) hold and (iii) is replaced with (iii$'$) $\Sigma$ is finite and positive definite,
then \( (a) \hat{\beta}_{n,w} \) is unbiased and the BLUE.

if we further assume

\[
(iv) \quad \varepsilon \sim N(0_n , \Omega ) ,
\]

then

\[
(b) \quad \hat{\beta}_{n,w} \text{ is the MLE} .
\]

proof. (a) can be shown by an argument similar to that used to verify (b) and (c) of Theorem 1.1 and the proof is left as an exercise. To show (b), notice that the joint density function of \((Y_1, \ldots, Y_n)\) is given by

\[
f_{\beta_0}(Y_1, \ldots, Y_n) = (2\pi)^{-d/2} \exp \left\{ -\frac{1}{2} (Y - \beta_0 \Lambda)^{-1} (Y - \beta_0 \Lambda)/2 \right\} ,
\]

and hence the corresponding log likelihood function is

\[
\log L(\beta_0) = \log f_{\beta_0}(Y_1, \ldots, Y_n)
\]

\[= - \frac{n}{2} \log 2\pi - \frac{1}{2} \log \det \Lambda - \frac{1}{2} (Y - \beta_0 \Lambda)^{-1} (Y - \beta_0 \Lambda).
\]

Let

\[
\frac{d}{\beta_0} \log L(\beta_0) = 0_k . \tag{1.12}
\]
Then, one gets
\[(X^T \tilde{x} X) \beta_0 = X^T \tilde{x} Y. \quad \text{(why?)}\]

Therefore, \( \hat{\beta}_{\text{OLSW}} \) is the solution to (1.12). By showing that
\[\frac{\partial^2 \log L(\beta_0)}{\partial \beta_0^2} \bigg|_{\beta_0 = \hat{\beta}_{\text{OLSW}}} \]
is negative definite, the desired result (6) follows.

Remark 2. Assume model (1.1). Then, under (i), (ii), (iii'), and (iv'), \( \hat{\beta}_{\text{OLSW}} \) is also the best unbiased estimator.

The rest of this chapter focuses on situations where Assumption (i) in Theorem (1.1) does not hold, that is, \( X \) can be a random matrix. We begin with several examples showing that the OLS estimator fails to retain some desired properties.

Example 1. Suppose the data are generated as
\[(\text{Measurement error models}) \quad y_k = w_k \beta_0 + \epsilon_k, \; k = 1, \ldots, n, \quad (1.13)\]
but we measure \( \mathbf{w} \) subject to errors \( \mathbf{\varepsilon} \), as

\[
X_t = W_t + \mathbf{\varepsilon}_t, \quad t = 1, \ldots, n, \tag{1.11}
\]

where \( \beta_0 > 0 \) and \((W_t, \mathbf{v}_t, \mathbf{\varepsilon}_t) \sim N \left( (\mathbf{0}), \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma_v^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right) \)

with \( \mathbf{u} \in \mathbb{R} \) and \( \sigma^2, \sigma_v^2, \sigma^2 > 0 \). In view of (1.13) and (1.14), we have

\[
Y_t = X_t \beta_0 + \mathbf{\varepsilon}_t, \tag{1.15}
\]

where \( \mathbf{\varepsilon}_t = \mathbf{v}_t - X_t \beta_0 \).

Now, the expectation of the OLS estimator \( \hat{\beta}_n \) satisfies

\[
E(\hat{\beta}_n) = \beta_0 + E \left( \left( \sum_{i=1}^{n} x_i^2 \right)^{-1} \sum_{i=1}^{n} x_i \varepsilon_i \right)
\]

\[
= \beta_0 + 0 + E \left( (\sum_{i=1}^{n} x_i^2)^{-1} \sum_{i=1}^{n} x_i \varepsilon_i \right)
\]

\[
= \beta_0 + 0,
\]

where \( \theta = \frac{E X_t \varepsilon_t}{E X_t^2} \), \( S_t = \varepsilon_t - \theta X_t \), and the second equality is ensured by observing that

\[
\varepsilon_t = 0 X_t + (\varepsilon_t - \theta X_t) = \theta X_t + \delta_t,
\]

and the last one follows from the fact that \( E \left( \sum_{i=1}^{n} x_i^2 \right)^{-1} \sum_{i=1}^{n} x_i \varepsilon_i \) = 0 (why?)
Since \( \theta = \frac{\sum X_k^2 \epsilon_k}{\sum X_k^2} = \frac{\beta_0}{\sigma^2 + \omega^2 + \epsilon_k^2} < 0 \),

\( \hat{\beta}_0 \) is not unbiased in this case.

Example 2. (Time Series models).

\[
y_t = \beta_0 y_{t-1} + \delta_t + \theta_0 \delta_{t-1}, \quad t = 1, \ldots, n, \tag{1.16}
\]

where \( |\beta_0| < 1, |\theta_0| < 1, \theta_0 \neq 0, \delta_t \sim \text{IID } N(0, \sigma^2) \), and the initial condition is given by \( y_0 = 0 \). Note that (1.16) is known as the ARMA(1,1) model. If we rewrite (1.16) as

\[
y_t = \theta_0 y_{t-1} + \epsilon_t,
\]

where \( \epsilon_t = \delta_t + \theta_0 \delta_{t-1} \) and use the OLS estimator

\[
\hat{\theta}_0 = \frac{\sum y_{k-1} y_k}{\sum y_{k-1}^2},
\]

\[
\hat{\beta}_0 = \beta_0 + \frac{\sum y_{k-1} \delta_k - \delta_{t-1}}{\sum y_{k-1}^2}, \tag{1.17}
\]
By an argument that will be introduced in the next two chapters, the first term on the RHS of (1.17) converges in probability to zero, whereas the second term on the RHS of (1.17) converges in probability to a nonzero constant. Therefore, \( \hat{\beta}_n \) is not a consistent estimator of \( \beta_0 \).

Example 3. (Simultaneous equations)

Consider

\[
Y_{x1} = X_{x2} \lambda_0 + W_{x1} \delta_0 + \epsilon_{x1}, \quad \lambda = 1, \ldots, n, (1.18)
\]

where \((W_{x1}, W_{x2}, \epsilon_{x1}, \epsilon_{x2}) \sim i.i.d.\)

\[
N \left( \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \\ 0 \\ 0 \end{pmatrix} \right), \left( \begin{pmatrix} \delta_{w1}^2 & \delta_{w1} \delta_{w2} & 0 & 0 \\ \delta_{w1} \delta_{w2} & \delta_{w2}^2 & 0 & 0 \\ 0 & 0 & \delta_{\epsilon 1}^2 & 0 \\ 0 & 0 & 0 & \delta_{\epsilon 2}^2 \end{pmatrix} \right)
\]

Let \( X_n = (Y_{x2}, W_{x1})' \) and \( \beta_0 = (\lambda_0, \delta_0) \). Then, the OLS estimator of \( \beta_0 \) based on the first equation in (1.18) is given by

\[
\hat{\alpha}_n = \left( \frac{1}{n} \sum_{k=1}^{n} X_k X_k' \right)^{-1} \frac{1}{n} \sum_{k=1}^{n} X_k Y_{x1},
\]

and hence

\[
E(\hat{\alpha}_n) = \beta_0 + E\left( \left( \frac{1}{n} \sum_{k=1}^{n} X_k X_k' \right)^{-1} \frac{1}{n} \sum_{k=1}^{n} X_k \epsilon_{x1} \right). \ (1.19)
\]
Since

\[ E(\mathbf{y}_k \mathbf{x}_1) = (\mathbf{c}_k, 0) \]

by an argument similar to that used in Example, the expectation on the RHS of (1.19) is non-zero, which yields \( \hat{\beta}_n \) is not unbiased.

**Remedy:** Instrumental Variables (Variables that are correlated with \( x_k \), but uncorrelated with \( \mathbf{e}_k \)).

**Example 4.** Consider (1.16) again. If we use

\[
\hat{\beta}_n = \left( \sum_{k=2}^{n-1} y_{k-2} y_{k-1} \right) \left( \sum_{k=2}^{n-1} y_{k-2} y_k \right)^{-1} \sum_{k=2}^{n-1} y_{k-2} \mathbf{e}_k
\]

to estimate \( \beta_0 \),

then

\[
\hat{\beta}_n = \beta_0 + \left( \sum_{k=2}^{n-1} y_{k-2} y_{k-1} \right)^{-1} \sum_{k=2}^{n-1} y_{k-2} \mathbf{e}_k
\]

(1.20)

where \( \mathbf{e}_k = \delta_k + 0.5 \delta_{k-1} \). As will be shown in the next two chapters, the second term on the RHS of (1.20) converges to zero in probability. Therefore, unlike \( \hat{\beta}_n \), \( \hat{\beta}_n \) is now a consistent estimator of \( \beta_0 \).
Remark 3. Note that $\tilde{\beta}_n$ is not necessarily an unbiased estimator of $\beta_0$. But it can be shown that $\tilde{\beta}_n$ is "asymptotically unbiased" in the sense that

$$\lim_{n \to \infty} E(\tilde{\beta}_n) = \beta_0.$$

In this case, $y_{k-1}$ is correlated with $y_{k-1}$, but uncorrelated with $
abla x = \delta x + \beta_0 \delta x - 1$. Hence, it is an instrumental variable. In general, let $z_k = (z_{k1}, \ldots, z_{ke})$ be a $k$-dimensional instrumental vector. Then,

$$Q_k = E( z_k \nabla x) = E( z_k (x_k - x_k \beta_0) )$$

which implies

$$E(z_k y_k) = E(z_k x_k) \beta_0 \quad (1.21)$$

Since the expectations in (1.21) are usually unknown, we replace expectations with sample averages, and consider finding a solution $\tilde{\beta}_0$.
\[ \frac{1}{n} \sum_{k=1}^{n} z_k (y_k - x_k' \beta_0) = \frac{Z (y - X \beta_0)}{n} = 0_k, \]  
(1.22)

where \( Z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \). Note that (1.22) is a system of \( n \) equations in \( k \) unknowns. If \( n \leq k \), then

\[ (Z'X)^{-1} Z'Y \]

is a solution to (1.22), where \((Z'X)^{-1}\) denotes a generalized inverse of \( Z'X \). Note that a generalized inverse of a matrix \( A \) is any matrix \( G \) such that \( AGA = A \).

In particular, if \( n = k \) and \( Z'X \) is nonsingular, then

\[ (Z'X)^{-1} Z'Y \]

is the unique solution to (1.22). If \( n < k \), however, (1.22) can have no solution. To alleviate the problem, we can estimate \( \beta_0 \) by finding that value of \( \beta \) that minimizes the quadratic distance

\[ d_n(\beta) = (Y - X \beta)' Z P_n Z (Y - X \beta), \]  
(1.23)
where \( \hat{\beta}_n \) is a symmetric \( n \times n \) positive definite matrix which may be stochastic. Now, a solution to

\[
\frac{\partial \ell_n(\beta)}{\partial \beta} = -X'Z\hat{\beta}_n Z(Y - X\beta) = 0
\]

is given by

\[
\hat{\beta}_n = \left( X'Z\hat{\beta}_n ZX \right)^{-1} X'Z\hat{\beta}_n ZY
\]

provided the inverse of \( X'Z\hat{\beta}_n ZX \) exists. After examining the second order derivatives, it can be shown that \( \hat{\beta}_n \) is the minimizer of (1.23) and is called "instrumental variables (IV) estimator."

Example 4. When \( Z = X \) and \( \hat{\beta}_n = \left( \frac{XX}{n} \right)^{-1} \)

\[
\hat{\beta}_n = \hat{\beta}_n.
\]

Given any \( Z \), choosing \( \hat{\beta}_n = \left( \frac{ZZ}{n} \right)^{-1} \) gives an estimator known as two-stage least squares (2SLS).
Remark 3. Under certain regularity conditions, it can be shown that $\hat{\beta}_n$ is a consistent estimator of $\beta$. The details will be given in the next two chapters.

Exercises:

1. Show (1.11).

2. Prove (a) of Theorem 1.2.

3. Show Remark 2 in the special case where $\beta_0$ is 1-dimensional and $6_0$ is known.