Booth Math Camp 2018: Statistical Inference

Jianfei Cao*

September 12, 2018

1 The Basics

1.1 Useful theorems

Definition 1. *Let* $\{X_n\}_{n=1}^{\infty}$ *and* X *be random vectors on* $\mathbb{R}^k$.

(i) $X_n \to_d X$, *if* $\Pr(X_n \leq x) \to \Pr(X \leq x)$ *for all continuous points of* $x \mapsto \Pr(X \leq x)$;

(ii) $X_n \to_p X$, *if* $\Pr(|X_n - X| \geq \epsilon) \to 0$ *for all* $\epsilon > 0$;

(iii) $X_n \to_{as} X$, *if* $\Pr(\lim_{n\to\infty} X_n = X) = 1$.

Remarks: 1. If $X_n \to_{as} X$, then $X_n \to_p X$; If $X_n \to_p X$, then $X_n \to_d X$.

2. If $X_n \to_d X$ and $Y_n \to_d Y$, it is not necessary that $(X_n, Y_n)' \to_d (X, Y)'$. Counter example: Let $X \sim N(0,1)$, $X_n = X$ and $Y_n = -X$ for each n. Then, $X_n \to_d X$ and $Y_n \to_d X$, but $(X_n, Y_n)' \to_d (X, X)'$ does not hold. This is because

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} = \begin{bmatrix} X \\ -X \end{bmatrix} \to_d N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right),$$

but

$$\begin{bmatrix} X \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right).$$

Lemma 1. *(Markov's Inequality) Let* X *be a random variable. Then,* $P(|X| \geq a) \leq \frac{E[|X|]}{a}$ *for* $a > 0$. *More generally,* $P(|X| \geq a) \leq \frac{E[f(|X|)]}{f(a)}$ *for* $a > 0$ *and* f *non-negative increasing function.*

Proof. $P(|X| \geq a) = E[\mathbb{1}\{|X| \geq a\}] \leq E[\frac{|X|}{a}\mathbb{1}\{|X| \geq a\}] = E[|X|/a] = E[|X|]/a.$ □

Corollary 1. *(Chebyshev's Inequality)* $P(|X - \mu| \geq k\sigma) \leq 1/k^2$, *where* $\mu = E[X]$ *and* $\sigma^2 = Var[X]$.

*Questions or comments: jcao0@chicagobooth.com

That is, for any random variable, the probability of being 2 standard deviations away is less than 1/4.

Lemma 2. *(Continuity of probability measure) Let* $P(A) = \Pr(X \in A)$ *be the induced probability measure. Then,*

$$A_1 \subset A_2 \subset \cdots \Rightarrow P(\cup_{n\geq 1} A_n) = \lim_{n\to\infty} P(A_n)$$

and

$$B_1 \supset B_2 \supset \cdots \Rightarrow P(\cap_{n\geq 1} B_n) = \lim_{n\to\infty} P(B_n).$$

Example 1. $X \sim N(0, I_2)$, $A_n = \{(X_1, X_2) : |X_1| < n, |X_2| < n\}$.

Remarks: 1. Even if both X and Y are normally distributed, it does not follow that (X, Y) is jointly normal. Counter example: $X_1 \sim N(0, 1)$ and

$$X_2 = \begin{cases} X_1, \text{ if } U \leq 1/2 \\ -X_1, \text{ if } U > 1/2, \end{cases}$$

where U is uniformly distributed on $[0, 1]$ and independent of X_1. Then $X_2 \sim N(0, 1)$. But $X_2|X_1$ is not normally distributed, violating requirements of joint normality.

2. If (X, Y) is jointly normal and $Cov[X, Y] = 0$, then X is independent of Y.

Theorem 1. *(Continuous Mapping Theorem/CMT) Let* $\{X_n\}_{n=1}^{\infty}$ *and* X *be random vectors on* $\mathbb{R}^k$*, and* $g : \mathbb{R}^k \to \mathbb{R}^d$ *is continuous at a set* $C \subset \mathbb{R}^k$ *where* $\Pr\{X \in C\} = 1$.
(i) If $X_n \to_d X$*, then* $g(X_n) \to_d g(X)$*;*
(ii) If $X_n \to_p X$*, then* $g(X_n) \to_p g(X)$*;*
(iii) If $X_n \to_{as} X$*, then* $g(X_n) \to_{as} g(X)$.

CMT is useful when consistency or asymptotic distribution of $g(X_n)$ are hard to obtain but X_n is easy.

Proof. (ii) Fix $\epsilon > 0$. For each $\delta > 0$, let $B_\delta = \{x \in \mathbb{R}^k : \exists y \in \mathbb{R}^k,\ s.t.\ d(x, y) < \delta \text{ and } d(g(x), g(y)) > \epsilon\}$. Then, for some $x \in \mathbb{R}^k$, if $d(g(x), g(y)) > \epsilon$ and $x \notin B_\delta$, then $d(x, y) \geq \delta$. Thus,

$$\Pr(d(g(X_n), g(X)) > \epsilon) \leq \Pr(X \in B_\delta) + \Pr(d(X_n, X) \geq \delta).$$

Let $A_n = B_{1/n} \cap C$. Pick any $x \in C$. By continuity of g, there exits $\eta > 0$, such that $d(x, y) < \eta$ implies $d(g(x), g(y)) < \epsilon$. That is, for each $n > 1/\eta$, $x \notin B_{1/n}$. So $(\cap_{n\geq 1} A_n) \cap C = \emptyset$. By continuity of probability

measure,

$$\lim_{n\to\infty} \Pr(X \in B_{1/n}) = \lim_{n\to\infty} \Pr(X \in A_n) = \Pr(X \in \cap_{n\geq 1} A_n) \leq 1 - \Pr(X \in C) = 0.$$

This shows $\Pr(X \in B_\delta) \to 0$ as $\delta \to 0$. Then,

$$\limsup_{n\to\infty} \Pr(d(g(X_n), g(X)) > \epsilon) \leq \Pr(X \in B_\delta) + 0.$$

Letting δ go to zero on both sides shows the theorem.

(iii) Let Ω be the sample space. Assume for each $\omega \in \Omega_1 \subset \Omega$, we have $X_n(\omega) \to X(\omega)$. Let $\Omega_2 = \{\omega \in \Omega : X(\omega) \in C\}$. Then, for each $\omega \in \Omega_0 = \Omega_1 \cap \Omega_2$, $X_n(\omega) \to X(\omega)$ implies $g(X_n(\omega)) \to g(X(\omega))$ by CMT of non-random sequence. Also, $\Pr(\Omega_0) = \Pr((\Omega_1^c \cup \Omega_2^c)^c) = 1 - \Pr(\Omega_1^c \cup \Omega_2^c) \geq 1 - \Pr(\Omega_1^c) - \Pr(\Omega_2^c) = 1$. □

Theorem 2. *(Weak Law of Large Number/WLLN) If $\{X_i\}_{i=1}^\infty$ is an i.i.d. sequence of random vectors such that $E[|X_i|] < \infty$, then $\bar{X}_n \to_p E[X_i]$ as $n \to \infty$, where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.*

WLLN is often used to show consistency.

Proof. If $Var[X_i] < \infty$, WLLN holds by Chebyshev's Inequality. □

Theorem 3. *(Strong Law of Large Number/SLLN) If $\{X_i\}_{i=1}^\infty$ is an i.i.d. sequence of random vectors such that $E[|X_i|] < \infty$, then $\bar{X}_n \to_{a.s.} E[X_i]$ as $n \to \infty$, where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.*

Theorem 4. *(Central Limit Theorem/CLT) If $\{X_i\}_{i=1}^\infty$ is an i.i.d. sequence of random vectors such that $Var[X_i] < \infty$, then $\sqrt{n}(\bar{X}_n - E[X_i]) \to_d N(0, Var[X_i])$, as $n \to \infty$.*

CLT is often used to show asymptotic normality (often combined with CMT(i)).

Proposition 1. *(Law of Iterated Expectation/LIE) Suppose X and Y are random variables and $E[X]$ exists, then $E[X] = E[E[X|Y]]$. More generally, $E[X|Z] = E[E[X|Y, Z]|Z]$.*

LIE is useful when dealing with mean independence.

Example 2. *Let $f(y) = E[X|Y = y]$. Verify $E[X] = E[f(Y)]$:*

	y_1	y_2
x_1	1/9	2/9
x_2	1/3	1/3

1.2 Goals of statistical inference

The three prototypical tasks of statistical inference are estimation, hypothesis testing, and constructing confidence region.

1.2.1 estimation

Assume the data $\{(W_i, Z_i)\}_{i=1}^n$ are generated by some probability measure $P_{\theta,\eta}$. Suppose the parameter of interest is θ and we observe W_i but not Z_i. Then an estimator of θ is a function $\hat{\theta}_n = \hat{\theta}_n(\{W_i\}_{i=1}^n)$.

Example 3. *$Y_i = X_i + \epsilon_i$, where $X_i \sim N(\mu, \sigma^2)$ and $\epsilon_i \sim N(0, v^2)$. We observe Y_i and only care about μ.*

Definition 2. *$\hat{\theta}_n$ is an **unbiased** estimator of θ if $E[\hat{\theta}_n] = \theta$. It is **consistent** or **asymptotically unbiased** if $\hat{\theta}_n \to_p \theta$. It is **asymptotically normal** if $g(n)(\hat{\theta}_n - \theta) \to_d N(0, \Sigma)$, where $g(n) \to \infty$ as $n \to \infty$. We call $g(n)$ the **rate of convergence**.*

Example 4. *Suppose we have an i.i.d. sample of observations $\{X_i\}_{i=1}^n$ and we know $X_i \sim N(\mu, \sigma^2)$. A natural estimator for μ is the sample analog $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Note that $\bar{X}_n \sim N(\mu, 1/n)$ and is unbiased, consistent, and asymptotically normal. The rate of convergence is $\sqrt{n}$. Note that the rate of convergence is unique.*

1.2.2 hypothesis testing

Suppose we want to perform hypothesis testing where the null hypothesis is H_0, and the significance level is α. A test is a function $\phi_n = \phi_n(\{W_i\}_{i=1}^n)$ that takes values between 0 and 1. Usually, $\phi_n = 1$ means rejecting and $\phi_n = 0$ means not rejecting.

Definition 3. *The **size** of a test is $E_P[\phi_n]$ for some probability measure P that satisfies H_0. The test is **consistent in level** if $\limsup_{n\to\infty} E_P[\phi_n] \leq \alpha$. The **power** of a test is $E_{P'}[\phi_n]$ for some P' that does not satisfies H_0.*

Example 5. $\phi_n = \{|t| > c.v.\}$.

Type-I error refers to rejecting when the null hypothesis is true. The probability of Type-I error is the size. Type-II error refers to not rejecting when the null hypothesis is false. The probability of Type-II error is (1-power), which depends on the true probability measure.

Example 6. *(Trivial test that is consistent in level) Let $\phi_n = \alpha$. Then it is consistent in level. In fact, it has correct size for each n.*

Example 7. *In the previous example, consider the null hypothesis $H_0 : \mu = \mu_0$. Let q_1 and q_2 be the $\alpha/2$- and $(1-\alpha/2)$-quantile of $N(\mu_0, 1/n)$, respectively. Then,*

$$\Pr(q_1 \leq \bar{X}_n \leq q_2) = 1 - \alpha.$$

That is, under the null hypothesis, there is large probability that $\bar{X}_n \in [q_1, q_2]$ when α is small. Hence, we construct a test $\phi_n = \phi_n(X_1, X_2, \ldots, X_n) \in [0,1]$ such that $\phi_n = \mathbf{1}\{\bar{X}_n \notin [q_1, q_2]\}$. In this example, the size is just α. Under some alternative $\mu = \mu_1$ where $\mu_1 \neq \mu_0$, the Type-II error is $F_{\mu_1}(q_2) - F_{\mu_1}(q_1)$, where F_{μ_1} is the c.d.f. of $\bar{X}_n$ under $\mu = \mu_1$.

1.2.3 confidence region

Confidence region is a random set $C_n = C_n(\{W_i\}_{i=1}^n)$ such that $\liminf_{n\to\infty} \Pr(\theta \in C_n) \geq 1 - \alpha$, where θ is the parameter of interest and α is significant level. Note that the probability is taken over C_n.

Example 8. *In our previous example, let $C_n = [p_1, p_2]$, where p_1 and p_2 are the $\alpha/2$- and $(1-\alpha/2)$-quantile of $N(\bar{X}_n, 1/n)$, respectively. Notice that $\Pr(p_1 \leq \mu \leq p_2) = \Pr(p_1 - \bar{X}_n \leq \mu - \bar{X}_n \leq p_2 - \bar{X}_n)$, implying $\Pr(\mu \in C_n) = 1 - \alpha$.*

Remarks: 1. In the above example, the exact distribution of the estimator can be calculated. In more general cases where distribution of the estimator cannot be obtained, we usually use CLT to derive the asymptotic distribution and perform hypothesis testing (or constructing confidence region). That's why we need $\liminf_{n\to\infty} \Pr(\theta \in C_n) \geq 1 - \alpha$ instead of just $\Pr(\theta \in C_n) \geq 1 - \alpha$.

2. Hypothesis testing and constructing the confidence region are equivalent in the sense that the confidence region can be obtained by inverting the test. Let $\phi_n(t)$ denote the test of $H_0 : \theta = t$ and can only take on 1 or 0 (either rejecting or not rejecting with probability one), then we have $\limsup_{n\to\infty} E_P[\phi_n(\theta)] \leq \alpha$. Construct the confidence region by inverting the test such that $C_n = \{t \in \Theta : \phi_n(t) = 0\}$. Then

$$\liminf_{n\to\infty} \Pr(\theta \in C_n) = \liminf_{n\to\infty} \Pr(\phi_n(\theta) = 0) = 1 - \liminf_{n\to\infty} E[\phi_n(\theta)] \geq 1 - \alpha.$$

2 Linear Model

2.1 Interpretations

Suppose we have a sample of $\{(X_i, Y_i)\}_{i=1}^n$ and assume a linear model

$$Y = X'\beta + U,$$

where Y and U are scalars, and X and β are k-dimensional vectors. There are three interpretations of this linear regression equation.

Interpretation 1. (Linear Conditional Expectation) We assume the conditional expectation of Y is linear in X, i.e. $E[Y|X] = X'\beta$. Then we must have mean independence $E[U|X] = 0$.

Interpretation 2. (Best Linear Predictor) Here $E[Y|X]$ is not necessarily linear in X. Let

$$\beta = \arg\min_{b \in \mathbb{R}^k} E[(Y - X'b)^2],$$

then $X'\beta$ is the best predictor of Y among all functions that is linear in X. Note that we only have $E[XU] = 0$, which is weaker than mean independence.

Interpretation 3. (Linear Causal Model) X is the observed determinant of Y and U is the unobserved determinant. The relationship between X and U is not determined by the model.

2.2 OLS and its properties

The standard procedure of estimating a statistical model is (i) propose an estimator (ii) show consistency (iii) derive asymptotic distribution. Suppose we have i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$, where for each i, Y_i is a scalar and X_i is a k-dimensional vector. Consider the linear model

$$Y = X'\beta + U,$$

where Y and U are scalars, and X and β are k-dimensional column vectors. It can have any of the interpretations, depending on your research question. Also, consider a set of assumptions:

[A1]: $E[XU] = 0$.

[A1']: $E[U|X] = 0$.

[A1"]: $E[U] = 0$ and $X \perp U$.

[A2]: $E[XX'] < \infty$ and is nonsingular.

[A3]: $E[XX'U^2] < \infty$.

[A3']: $E[U^2|X] = \sigma^2 < \infty$.

Note that **[A1"]**⊂**[A1']**⊂**[A1]**. Under **[A2]**, **[A3']**⊂**[A3]**. The error is said to be homoskedastic if $Var[U|X]$ does not vary with X; otherwise, it is heteroskedastic. In terms of our assumptions, we need **[A1']** and **[A3']** for U to be homoskedastic.

Under **[A1]** (or stronger assumptions), the model can be rewritten as

$$E[XY] = E[XX']\beta + E[XU] = E[XX']\beta.$$

Under **[A2]**,

$$\beta = E[XX']^{-1}E[XY].$$

Therefore, we propose the OLS estimator using the sample analog:

$$\hat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right).$$

Note that

$$\hat{\beta}_{OLS} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i U_i\right).$$

Under **[A1']** or **[A1"]**, $\hat{\beta}$ is unbiased.

By WLLN and CMT, $\hat{\beta}_{OLS} \to_p \beta$. Note that to show consistency, we only need **[A1]** and **[A2]**.

Under **[A1]**, **[A2]**, and **[A3]**,

$$\sqrt{n}(\hat{\beta}_{OLS} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i U_i\right) \to_d N(0, E[XX']^{-1}E[XX'U^2]E[XX']^{-1}),$$

by CMT, WLLN, and CLT. Let the asymptotic covariance matrix be

$$V = E[XX']^{-1}E[XX'U^2]E[XX']^{-1}.$$

Case 1: **[A3']** (homoskedasticity)

Then $V = \sigma^2 E[XX']$. A consistent estimator of V is $\hat{V}_{homo} = \hat{\sigma}^2(n^{-1}\sum_i X_i X_i')$, where $\hat{\sigma}^2 = n^{-1}\sum_i \hat{U}_i^2$

and $\hat{U}_i = Y_i - X_i'\hat{\beta}_{OLS}$. Consistency: By WLLN and CMT

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_i (Y_i - X_i'\hat{\beta})^2 \\
&= \frac{1}{n}\sum_i (X_i'(\beta - \hat{\beta}_{OLS}) + U_i)^2 \\
&= (\beta - \hat{\beta}_{OLS})' \left(\frac{1}{n}\sum_i X_i X_i'\right)(\beta - \hat{\beta}_{OLS}) + 2\left(\frac{1}{n}\sum_i U_i X_i'\right)(\beta - \hat{\beta}_{OLS}) + \frac{1}{n}\sum_i U_i^2 \\
&\to_p \sigma^2,
\end{aligned}$$

so by WLLN and CMT

$$\hat{V}_{homo} \to_p \sigma^2 E[X_i X_i'].$$

Case 2: **[A3]** (heteroskedasticity)

A consistent estimator of V is the sample analog:

$$\hat{V}_{hetero} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i X_i' \hat{U}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1},$$

where the residual $\hat{U}_i = Y_i - X_i'\hat{\beta}_{OLS}$. A proof can be found in Azeem's lecture notes.

2.3 Efficiency

An estimator is efficient when it achieves the lowest variance possible. Or we can say an estimator is more efficient when it has a lower variance than another estimator.

Example 9. *For i.i.d. sample $\{X_i\}_{i=1}^n$ where $E[X_i] = \mu < \infty$ and $Var[X_i] = \sigma^2 < \infty$, we propose two estimators for μ, $\hat{\mu}_1 = (n/2)^{-1}\sum_{i \text{ odd}} X_i$ and $\hat{\mu}_2 = \bar{X}_n$. Both estimators are unbiased and consistent, but $Var[\hat{\mu}_1] = 2\sigma^2/n > \sigma^2/n = Var[\hat{\mu}_2]$. We say $\hat{\mu}_2$ is more efficient than $\hat{\mu}_1$.*

Example 10. *We say a square matrix $M \in \mathbb{R}^{k\times k}$ is positive semi-definite if $\forall x \in \mathbb{R}^k$, $x'Mx \geq 0$. If estimator $\hat{\beta}_1$ and $\hat{\beta}_2$ are k-dimensional vectors, we say $\hat{\beta}_1$ is more efficient than $\hat{\beta}_2$ if $Var[\hat{\beta}_2] - Var[\hat{\beta}_1]$ is positive semi-definite. The intuition is that a linear combination of $\hat{\beta}_1$ will always have a lower variance of the same linear combination of $\hat{\beta}_2$, i.e. $Var[c'\hat{\beta}_2] - Var[c'\hat{\beta}_1] = c'(Var[\hat{\beta}_2] - Var[\hat{\beta}_1])c \geq 0$.*

Theorem 5. *(Gauss-Markov Theorem) Assume i.i.d. sampling, $E[U|X] = 0$ and $E[U^2|X] = \sigma^2$. Then the OLS estimator $\hat{\beta}_{OLS}$ is the best linear unbiased estimator (BLUE), i.e. among all estimators $\tilde{\beta}$ of the form $\tilde{\beta} = \sum_{i=1}^n a_i Y_i$ with $a_i = a_i(\{X_i\}_{i=1}^n)$ being a k-dimensional function of the regressors, such that $E[\tilde{\beta}|X_1, \ldots, X_n] = \beta$, we must have $\hat{\beta}_{OLS} = \arg\min_{\tilde{\beta}} Var[\tilde{\beta}|X_1, \ldots, X_n]$.*

Proof. Let $\mathbb{Y} = (Y_1, \ldots, Y_n)'$ and $\mathbb{X} = (X_1, \ldots, X_n)'$. Let $\tilde{\beta} = A\mathbb{Y}$ where A is a function of $\mathbb{X}$. Write

$$\tilde{\beta} = \hat{\beta}_{OLS} + D\mathbb{Y}$$

for $D = A - (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}$. Then $\tilde{\beta}$ is unbiased only when $E[D\mathbb{Y}|\mathbb{X}] = D\mathbb{X}\beta = 0$ for each β, implying $D\mathbb{X} = 0$.

Thus,

$$Var[\tilde{\beta}|\mathbb{X}] = AVar[\mathbb{Y}|\mathbb{X}]A' = \sigma^2(D + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')(D + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}')' = \sigma^2 DD' + \sigma^2(\mathbb{X}'\mathbb{X})^{-1} \geq Var[\hat{\beta}_{OLS}|\mathbb{X}].$$

□

In more general cases, U is heteroskedastic ($E[U^2|X]$ is a function of X). Suppose $E[U|X] = 0$ and $E[U^2|X = x] = \sigma^2(x)$, then the linear model can be written as

$$\frac{1}{\sigma(X)}Y = \frac{1}{\sigma(X)}X'\beta + \frac{1}{\sigma(X)}U,$$

i.e. a transformed linear model such that

$$Y^* = (X^*)'\beta + U^*,$$

where $Y^* = Y/\sigma(X)$, etc. Note that this is a linear model with homoskedasticity. Namely,

$$\textbf{[A1']}: \quad E[\sigma(X)^{-1}U|\sigma(X)^{-1}X] = E[E[\sigma(X)^{-1}U|X, \sigma(X)^{-1}X]|\sigma(X)^{-1}X] = 0.$$

$$\textbf{[A3']}: \quad E[(\sigma(X)^{-1}U)^2|\sigma(X)^{-1}X] = E[E[\sigma(X)^{-2}U^2|X, \sigma(X)^{-1}X]|\sigma(X)^{-1}X] = 1.$$

By Gauss-Markov Theorem, OLS regression of $\sigma(X)^{-1}Y$ on $\sigma(X)^{-1}X$ yields an efficient estimator. This is called the generalized least square estimator (GLS), which is

$$\hat{\beta}_{GLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2(X_i)}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2(X_i)}X_iY_i\right).$$

We can show $\hat{\beta}_{GLS}$ is unbiased, consistent, and asymptotically normal. Also, by the Gauss-Markov theorem, for each $\tilde{\beta} = A\mathbb{Y}^*$ such that $E[\tilde{\beta}|\mathbb{X}^*] = \beta$, we must have

$$Var[\tilde{\beta}|X_1^*, \ldots, X_n^*] - Var[\hat{\beta}_{GLS}|X_1^*, \ldots, X_n^*] \geq 0,$$

i.e. the difference between the two matrices is positive semi-definite.

Note that in practice, GLS is infeasible, since we don't observe $\sigma^2(X_i)$ from the sample. We now need extra assumptions from our economic model to proceed. For example, your intuitions says the conditional variance is quadratic in X, then you may assume

$$E[U^2|X] = aX^2 + bX + c,$$

for some unknown parameters a, b, c. Note that this is again a conditional expectation relationship, and an OLS estimator of U^2 on X is consistent. Since we do not observe U, we use the OLS regression residuals $\hat{U}_i = Y_i - X_i'\hat{\beta}_{OLS}$ to estimate (a, b, c) and obtain $(\hat{a}, \hat{b}, \hat{c})$. Finally, we replace $\sigma^2(X_i)$ in the formula of GLS by $\hat{\sigma}^2(X_i) = \hat{a}X_i^2 + \hat{b}X_i + \hat{c}$. This is called a feasible generalized least square estimator (FGLS). The general procedure is:

Step 1: Do OLS and form the residuals $\hat{U}_i = Y_i - X_i'\hat{\beta}_{OLS}$.

Step 2: Propose a model for conditional variance of the disturbance:

$$E[U^2|X = x] = \sigma^2(x),$$

where $\sigma^2 : \mathbb{R}^k \to \mathbb{R}$. Exploit the relationship between $\hat{U}_i^2$ and X_i to obtain a function estimator $\hat{\sigma}^2(x)$.

Step 3: The FGLS estimator is

$$\hat{\beta}_{FGLS} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{\sigma}^2(X_i)}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{\sigma}^2(X_i)}X_iY_i\right).$$

Note that in Step 2, there is no guarantee that your estimates yields a meaningful variance such that $\hat{\sigma}(X_i) \geq 0$ for each i. Usually, this will not cause problems in the limit, given that the conditional variance is correctly specified.

3 Endogeneity

3.1 Examples of endogeneity

In the above regression equation, a regressor X_i is endogenous if $E[X_iU] \neq 0$; it is exogenous if $E[X_iU] = 0$. Endogeneity is ubiquitous in social scientific research, and it is often the main task of an empirical paper to overcome it.

Omitted Variable Bias: Suppose the true model is $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + U$ and $E[U|X_1, X_2] = 0$, but

one can only observe X_1. The corresponding linear model is $Y = \beta_0^* + X_1\beta_1 + U^*$, where $\beta_0^* = \beta_0 + E[X_2\beta_2]$ and $U^* = (X_2 - E[X_2])\beta_2 + U$. OLS will be estimating

$$\beta_1^* = \frac{Cov[X_1, Y]}{Var[X_1]} = \beta_1 + \frac{Cov[X_1, X_2]}{Var[X_1]}\beta_2.$$

That is, OLS is consistent only when $Cov[X_1, X_2] = 0$ or $\beta_2 = 0$. Another way to see it is to notice that we require $E[XU^*] = 0$ for the OLS estimator to be consistent and that $E[XU^*] = Cov[X_1, X_2]\beta_2$.

Measurement Error: Suppose the true model is $Y = \alpha + X\beta + U$ where $E[U|X] = 0$, but we do not observe X directly. Instead, we observe $X^* = X + V$, where V has mean zero and is independent with anything else. Then the regression equations becomes $Y = \alpha + X^*\beta + U^*$ where $U^* = U - V\beta$. Now we have $E[X^*U^*] = -E[V^2]\beta$, which is not zero when V is non-degenerate. In this case, OLS will estimate

$$\beta^* = \frac{Cov[X^*, Y]}{Var[X^*]} = \frac{Var[X]}{Var[X] + Var[V]}\beta,$$

and this is called attenuation bias.

Simultaneous Equation System: Consider the following system of demand and supply:

$$\begin{aligned} Q^d &= \alpha^d + \beta^d P + \gamma^d X^d + U^d \\ Q^s &= \alpha^s + \beta^s P + \gamma^s X^s + U^s, \end{aligned}$$

where Q and P are quantity and price, X is some demand or supply shifters, and U is demand or supply shock. Note that quantity and price are simultaneously determined by both equations, and we only observe data points where supply and demand meet. Equating two equations gives

$$P = \frac{1}{\beta^d - \beta^s}(\alpha^s - \alpha^d + U^s - U^d + \gamma^d Xd - \gamma^s X^s),$$

which is correlated the error. Simply regressing quantities on prices may even yields an upward sloping demand curve.

Sample Selection: Consider a linear regression equation

$$Y = X'\beta + U$$

where $E[XU] = 0$. Assume Y is observed only when $Y > 0$. Then regression of Y on X will not necessarily yield a consistent estimate, because $E[XU|Y > 0] = 0$ does not need to hold.

3.2 Reduced form solution - IV

If we observe the driving force of unobservables, we can include it in the regression. If not, we often use IV to solve endogeneity. One way to interpret IV is to predict X as if it is not affected by the unobservables, and then regress Y on the predicted vale of X.

3.2.1 IV and 2SLS

Suppose the linear regression equation

$$Y = X'\beta + U$$

with $\beta \in \mathbb{R}^k$ and assume endogeneity, i.e. at least one of the regressors is endogenous. Further assume we have a set of variables $Z \in \mathbb{R}^l$ such that $l \geq k$ and Z includes all exogenous variables in X. Key assumptions of instrumental variables to work:

[EX]: (exclusion) $E[ZU] = 0$.

[RE]: (relevance) $E[ZX']$ has full rank k.

Other regularity assumptions are similar to what we need for OLS, including $E[ZZ'] < \infty$ and is non-singular, and $E[ZX'] < \infty$. Note that the exclusion condition implies

$$E[ZX']\beta = E[ZY],$$

thus if $l = k$, we can write

$$\beta = E[ZX']^{-1}E[ZY]$$

under the relevance condition. This case ($l = k$) is called just-identification and we propose the instrumental variable (IV) estimator

$$\hat{\beta}_{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i,$$

which is the sample analog just as in the OLS case. When $l > k$, we rewrite the exclusion condition as

$$E[XZ']E[ZZ']^{-1}E[ZX']\beta = E[XZ']E[ZZ']^{-1}E[ZY],$$

which implies

$$\beta = (E[XZ']E[ZZ']^{-1}E[ZX'])^{-1}E[XZ']E[ZZ']^{-1}E[ZY].$$

This is called over-identification and we propose the two stage least square (2SLS) estimator

$$\hat{\beta}_{2SLS} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right).$$

Another way to look at it is to note $\pi = E[ZZ']^{-1}E[ZX']$ is the coefficient for the best linear predictor of X on Z. Also, we have $E[ZX'] = E[ZZ']\pi$. Therefore, the exclusion condition can be written as

$$E[\pi' ZZ' \pi]\beta = E[\pi' ZY],$$

and thus

$$\beta = E[\pi' ZZ' \pi]^{-1} E[\pi' ZY].$$

This suggests an estimator such that

$$\tilde{\beta}_{2SLS} = \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i Y_i,$$

where

$$\tilde{X}_i = \hat{\pi}' Z_i = \left(\left(\frac{1}{n}\sum_{j=1}^{n} Z_j Z_j'\right)^{-1} \frac{1}{n}\sum_{j=1}^{n} Z_j X_j'\right)' Z_i$$

is the best linear predictor. That corresponds to the name of 2SLS: the first stage is to regress X on Z, and the second stage is to regress Y on the fitted values from the first stage. Note that $\hat{\beta}_{2SLS}$ and $\tilde{\beta}_{2SLS}$ are mathematically the same.

3.2.2 properties of 2SLS

Since the IV estimator is a special case of 2SLS, we only discuss the latter.

Consistency: Write

$$\hat{\beta}_{2SLS} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Z_i' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Z_i U_i\right).$$

Then under i.i.d. sampling, $\hat{\beta}_{2SLS}$ consistently estimates β when $E[ZZ']$ exists and is non-singular, $E[ZX']$ exists and is full rank, and $E[ZU] = 0$.

Asymptotic normality: Assume further that $Var[ZU] = E[ZZ'U^2] < \infty$. Then we have

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \left(\hat{\pi}'\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right)\hat{\pi}\right)^{-1} \hat{\pi}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i U_i\right) \to_d N(0, \Omega),$$

where

$$\Omega = E[\pi' ZZ'\pi]^{-1}(\pi' E[ZZ'U^2]\pi)E[\pi' ZZ'\pi]^{-1}.$$

A consistent estimator for Ω is again the sample analog

$$\hat{\Omega} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\pi} Z_i Z_i' \hat{\pi}\right)^{-1} \hat{\pi}'\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \hat{U}_i^2\right)\hat{\pi}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\pi} Z_i Z_i' \hat{\pi}\right)^{-1},$$

where $\hat{U}_i = Y_i - X_i'\hat{\beta}_{2SLS}$. Note that the residual is not the one from the second stage of 2SLS.

3.3 Structural modeling solution - Heckman correction

The IV estimator is a reduced-form approach because it does not impose distribution assumption or other structural assumption. If one has extra institutional knowledge about the model, or is willing to introduce more identification assumptions, structural modeling approach can be used to solve the endogeneity problem. A famous example to solve self-selection is Heckman correction, which is now generalized by control function approach. We will use the example of self-selection in the job market to illustrate its usage.

3.3.1 unconditional case

Suppose a sample $\{Y_i^*\}_{i=1}^n$ is collected on incomes of individuals. However, we can only observe the incomes from those who choose to work. That is, the data is censored (distinguish from truncated data). We want to learn the distribution of their incomes if they choose to work.

Suppose the (potential) income of an individual i is Y_i and the c.d.f. of Y is F. The observed income $Y_i^* = Y_i$ is $Y_i > c$, where c is the threshold for entering the job market. Let $Y_i^* = c$ if $Y_i \leq c$. Let $D = \mathbf{1}\{Y > c\}$. Then, the conditional distribution of Y^* on D for some $y > c$ is

$$F^*(y|D=1) = \Pr(Y \leq y | Y > c) = \frac{\Pr(Y^* \leq y, Y > c)}{\Pr(Y > c)} = \frac{F(y) - F(c)}{1 - F(c)}$$

and

$$F^*(y|D=0) = 1.$$

Thus, the joint distribution can be written as

$$F_{Y^*,D}(y,d) = F^*(y|D=d)\Pr(D=d) = \left(\frac{F(y)-F(c)}{1-F(c)}\right)^d (1-F(x))^d 1^{1-d} F(c)^{1-d} = (F(y)-F(c))^d F(c)^{1-d}.$$

We can now use maximum likelihood to recover the distribution of Y, if we impose some functional form of F. The most common assumption is that Y is normal.

3.3.2 conditional case

We extent the previous example and assume we have a sample of $\{(Y_i^*, X_i)\}_{i=1}^n$, where X includes a set of regressors that may contribute to income determination. Assume a linear model

$$Y = X'\beta + U,$$

where X and U are independent. Again, we assume Y_i is only observed when it is higher than some threshold c. In this case, simply doing regression with observed sample may severely distort statistical inference, because the coefficients in the best linear predictor may be different for each subsample. For example, one more year of eduction may be more effective on the population with low income than the high-income subgroup. Therefore, regressing income on education with only the observed data may underestimate the effect of education.

To account for the selection, we introduce the control function method. The idea is to take out the endogenous part in the disturbance and control for it. This is generally infeasible if we don't impose extra distributional assumptions. Let $D = \mathbf{1}\{Y > c\}$ and U follow some distribution F_U. Then,

$$\begin{aligned}
\Pr(D=1|X=x) &= \Pr(Y > c|X=x) \\
&= \Pr(U > c - X'\beta|X=x) \\
&= \Pr(U > c - x'\beta|X=x) \\
&= \Pr(U > c - x'\beta) \\
&= 1 - F_U(c - x'\beta).
\end{aligned}$$

Conditional distribution of U is

$$F_U(z|D=1,X=x) = \frac{\Pr(U \le z, U > c - x'\beta)}{\Pr(D=1|X=x)} = \begin{cases} \frac{F_U(z)-F_U(c-x'\beta)}{1-F_U(c-x'\beta)}, & \text{if } z > c - x'\beta \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$E[Y|D=1,X=x] = x'\beta + E[U|D=1,X=x] = x'\beta + \frac{1}{1-F_U(c-x'\beta)}\int_{c-x'\beta}^{\infty} z\, dF_U(z).$$

Now we can impose functional form on F_U and the RHS of this regression function is a function of x. In principle, we can run OLS of the observed Y on X and the control, since this is a conditional expectation equation.

4 Extremum Estimator

4.1 Maximum likelihood

4.1.1 unconditional likelihood

Suppose we observed a sample $\{X_i\}_{i=1}^n$, where the joint density is $p_X(x_1, x_2, \ldots, x_n; \theta_0)$ and $\theta \in \Theta$. Assume for each $\theta \in \Theta$, the density function $p_X(x_1, x_2, \ldots, x_n; \theta)$ exists. The likelihood function of this sample is defined by $\ell_n(\theta) = p_X(X_1, \ldots, X_n; \theta)$. The log-likelihood is defined by $L_n(\theta) = n^{-1}\log \ell_n(\theta)$. The Maximum Likelihood estimator is defined by

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \ell_n(\theta) = \arg\max_{\theta\in\Theta} L_n(\theta).$$

Example 11. *Suppose we have i.i.d. sampling and for each i, X_i has density $p(x;\theta_0)$. Then the likelihood is $\ell_n(\theta) = \prod_{i=1}^n p(X_i;\theta)$. The log-likelihood is $L_n(\theta) = n^{-1}\sum_{i=1}^n \log p(X_i;\theta)$.*

Example 12. *Suppose the sample is not i.i.d. and follows AR(1) such that $X_{i+1} = \rho X_i + U_i$, U_i is $N(0,\sigma^2)$, and X_1 is fixed.. Then, the density can be written as*

$$p_X(x_1,\ldots,x_n;\theta) = p_{X_n|X_1,\ldots,X_n}(x_n|x_1,\ldots,x_{n-1};\theta) \times \cdots \times p_{X_1}(x_1|\theta) = \prod_{i=2}^n p_{X_i|X_{i-1}}(x_i|x_{i-1};\theta),$$

where $p_{X_i|X_{i-1}}$ is the density of $N(x_{i-1},\sigma^2)$.

Example 13. *Let X_i follow i.i.d. uniform distribution on $[0,\theta_0]$ with $0 < \theta_0 < \infty$. Then, $p(x;\theta) = \theta^{-1}\mathbf{1}\{0 \le x \le \theta\}$ and $\ell_n(\theta) = \theta^{-n}\mathbf{1}\{0 \le X_1,\ldots,X_n \le \theta\}$. The ML estimator for θ_0 is $\hat{\theta}_{ML} = \max_i X_i$.*

4.1.2 conditional likelihood

Suppose now we have data on both X and Y with the sample $\{(Y_i, X_i)\}_{i=1}^n$. The condition distribution of all Y_i's on all X_i's is $p_{Y|X}(y_1, \ldots, y_n|x_1, \ldots, x_n; \theta_0)$. The conditional likelihood is defined by

$$\ell_n(\theta) = p_{Y|X}(Y_1, \ldots, Y_n|X_1, \ldots, X_n; \theta).$$

The ML estimator is defined accordingly. Similarly, if the sample is i.i.d. with conditional density $p_{Y|X}(y|x;\theta)$, the likelihood function is just $\ell_n(\theta) = \prod_{i=1}^n p(Y_i|X_i;\theta)$ and the log-likelihood is $L_n(\theta) = n^{-1}\sum_{i=1}^n \log p(Y_i|X_i;\theta)$.

Example 14. *We revisit the linear model $Y = X'\beta + U$ and assume $X \perp U$ and U follows i.i.d. $N(0, \sigma^2)$. Let the parameter of interest be $\theta = (\beta, \sigma^2)$. The conditional density is*

$$p(y|x;\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y - x'\beta)^2}{2\sigma^2}\right)$$

and the log-likelihood is

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^n -\sqrt{2\pi}\sigma - \frac{(y - x'\beta)^2}{2\sigma^2}.$$

Note that the first order condition w.r.t. β requires $\sum_{i=1}^n X_i(Y_i - X_i'\beta) = 0$, which implies

$$\hat{\beta}_{ML} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_i Y_i = \hat{\beta}_{OLS}.$$

Note that this is generally not true, since to perform ML require extra distributional assumptions, so ML estimator is expected to outperform estimators that do not require parametric assumptions.

Example 15. *Suppose*

$$Y_i = \mathbf{1}\{X'\theta_0 + U \geq 0\}$$

where U follows some distribution F. Then the conditional density is

$$p(y|x;\theta) = (1 - F(-x'\theta))^y F(-x'\theta)^{1-y}$$

and the log-likelihood is

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^n Y_i \log(1 - F(-X_i'\theta)) + (1 - Y_i)\log F(-X_i'\theta).$$

In general, there is no analytical solution to the maximization problem. If F is standard normal, this model

is called a Probit model; it is called a Logit model if

$$F(u) = \exp(u)/(1 + \exp(u)).$$

The ML estimator can be solved efficiently in these two models.

4.2 Generalized method of moments

Method of moments stands for a class of estimators that are solved for by equating the sample analog of the moments to the populations ones. Examples:

[OLS]: $E[X(Y - X_i'\beta)] = 0$

[IV]: $E[Z(Y - X_i'\beta)] = 0$

[2SLS]: $\pi' E[Z(Y - X_i'\beta)] = 0$, where $\pi = E[ZZ']^{-1}E[ZX]$.

[ML]: $E[\frac{\partial}{\partial\theta}\log p(Y|X;\theta)] = 0$

In general cases, the solution to the sample analog equations does not necessarily exist. Suppose we have J moments where $E[m_j(X, Y; \theta)] = 0$ for $j = 1, \ldots J$. Let $m = (m_1, \ldots, m_J)'$. The GMM estimator is defined by

$$\hat{\theta}_{GMM} = \arg\min_{\theta\in\Theta} \left(\frac{1}{n}\sum_{i=1}^{n} m(X_i, Y_i; \theta)\right)' W \left(\frac{1}{n}\sum_{i=1}^{n} m(X_i, Y_i; \theta)\right),$$

for some weighting matrix W.

Consistency and asymptotic normality can be established for each W, but how can we choose W? It can be shown that efficiency bound is achieved when $W = E[m(X, Y; \theta_0)m(X, Y; \theta_0)']^{-1}$, where θ_0 is the true underlying parameter. Notice the similarity to GLS (generalized least square). The optimal weighting matrix is apparently infeasible. We can use a feasible 2-step GMM estimator:

Step 1: Perform GMM estimation using identity weighting matrix and obtain $\hat{\theta}_1$.

Step 2: Use the sample analog of the efficient weighting matrix $\hat{W} = n^{-1}\sum_{i=1}^{n} m(X_i, Y_i; \hat{\theta}_1)m(X_i, Y_i; \hat{\theta}_1)'$ and perform GMM estimation again. The resulting estimator $\hat{\theta}_2$ is the 2-step GMM estimator.

4.3 Extremum estimator

An M-estimator is defined by minimizing a sum over a function of the sample, i.e.

$$\hat{\theta}_M = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} f(X_i, Y_i; \theta).$$

Examples are OLS and ML.

A much more general version of this estimator is the Extremum estimator, which includes many popular estimators, including OLS, ML, GMM, etc. It is defined by minimizing some criterion function, not necessarily a sum:

$$\hat{\theta}_{EE} = \arg\min_{\theta \in \Theta} \hat{Q}(\theta),$$

where $\hat{Q}$ is a function of the data. Consistency and asymptotic normality can be established under regularity conditions. See Newey and McFadden (1994) for details.

5 Panel data

A set of panel data is a set of observations $\{(X_{i,t}, Y_{i,t})\}_{(i,t)\in I}$ indexed by both i and t, where $X_{i,t}$ has k dimensions. A balanced panel means that there are N and T such that the index set $I = \{1, \dots, N\} \times \{1, \dots, T\}$. We will focus on the balanced panel case. For the three following sections, we will use the same statistical model:

$$Y_{it} = \alpha_i + X_{it}'\beta + U_{it},$$

which we call the linear fixed effects model.

5.1 Fixed-effect estimator

For each i, let $\bar{z}_i$ be the within-individual average for $z \in \{Y, X, U\}$, i.e. $\bar{z}_i = T^{-1}\sum_{t=1}^{T} z_{it}$. Let $\dot{z}_{it}$ be the demeaned value, i.e. $\dot{z}_{it} = z_{it} - \bar{z}_i$. Then, the linear fixed effects model becomes

$$\dot{Y}_{it} = \dot{X}_{it}'\beta + \dot{U}_{it}.$$

By pooling data regardless of i and t and then applying OLS, we propose the fixed-effect estimator

$$\hat{\beta}_{FE} = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\dot{X}_{it}\dot{Y}_{it}\right).$$

There is a numerically equivalent representation of FE. For each i and t, let D_{it} be an indicator vector of N dimensions such that D_{it} has one on its i-th entry and zero everywhere else. Let $Z_{it} = (X_{it}', D_{it}')'$. Now we regress Y_{it} on Z_{it} with the pooled data, and obtain

$$\hat{\theta}_{FE} = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Z_{it}Z_{it}'\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Z_{it}Y_{it}\right).$$

The first k entries of $\hat{\theta}_{FE}$ will be numerically equivalent to $\hat{\beta}_{FE}$.

For either method, make sure X does not include a constant, otherwise you will have multicollinearity.

The statistical properties of the fixed-effect estimator vary according to the assumptions you are willing to make.

Case 1: Assume (Y_{it}, X_{it}, U_{it}) is i.i.d. for each i and t, then we can treat $\hat{\theta}_{FE}$ as an OLS estimator and derive standard OLS properties under regularity conditions.

Case 2: Assume (Y_{it}, X_{it}, U_{it}) is independent across i, but is correlated within i. That is, we assume $E[U_{it}U_{js}] = 0$ for $i \neq j$, but we might have $E[U_{it}U_{is}] \neq 0$ for some t and s. Also, assume $N \to \infty$ and T is fixed. In this case, the standard OLS mean-independence condition $E[U_{it}|X_{it}] = 0$ is not sufficient for consistency. We need some form of "strict exogeneity". An example is:

[SE]: (strict exogeneity) $E[U_{it}|\alpha_i, X_{i1}, X_{i2}, \ldots, X_{iT}] = 0$ for each i and t.

The consistency of the fixed-effect estimator can be established by

$$\begin{aligned}
\hat{\beta}_{FE} &= \left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{Y}_{it}\right)\right) \\
&= \beta + \left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{U}_{it}\right)\right) \\
&\to_p \beta + E\left[\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{X}_{it}'\right]^{-1} E\left[\frac{1}{T}\sum_{t=1}^{T}\dot{X}_{it}\dot{U}_{it}\right] \\
&= \beta.
\end{aligned}$$

To justify the convergence in probability, we need to check conditions of WLLN and invertibility. The last equality is due to **[SE]**.

We can follow the similar procedure to obtain the asymptotic distribution of $\hat{\beta}_{FE}$.

Note that there are other popular sets of assumptions in addition to the above two examples.

5.2 First-difference estimator

Another way to cancel out the individual effect is to take first difference. Namely, let $\tilde{z}_{i,t} = z_{i,t} - z_{i,t-1}$ for $z \in \{Y, X, U\}$. Then the linear fixed effects model is

$$\tilde{Y}_{it} = \tilde{X}_{it}'\beta + \tilde{U}_{it}.$$

Pooled OLS can be used now to estimate β, i.e.

$$\hat{\beta}_{FD} = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{X}_{it}' \right)^{-1} \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{Y}_{it} \right),$$

where we assume data of $t = 0$ is also available for notation simplicity. Again, different sets of assumptions yields different statistical properties.

5.3 Fama-MacBeth estimator

The Fama-MacBeth estimator (FM) is especially popular in financial research. It can take various forms, but the idea is to first divide observations into G groups, then estimate the parameter in each subgroup, and take the average of the G estimators as the final estimator. A typical FM estimator in the context of the linear fixed effect model is given as follows:

For each i, first do OLS of $\{Y_{it}\}_{t=1}^{T}$ on $\{X_{it}\}_{t=1}^{T}$ and a constant. Ignore the estimator for the constant and let the coefficients of X be $\hat{\beta}_i$. Then, the Fama-MacBeth estimator is given by

$$\hat{\beta}_{FM} = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}_i.$$

FM estimator was proposed in 1973 and the theory behind it was established relatively recently. It is conceptually simple and easy to implement. Surprisingly, it has very nice statistical properties. The intuition behind this is that we can treat each $\hat{\beta}_i$ as being drawn from some distribution. As long as they are approximately independent, their average will have nice properties.

References

[1] van der Vaart, A. W. (2000), *Asymptotic Statistics*, Chapter 2.

[2] Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.*

[3] Newey, W. K., & McFadden, D. (1994). *Large sample estimation and hypothesis testing. Handbook of econometrics*, 4, 2111-2245.