Notes on Econometrics

Qin Lei

Contents

Preface

This is my compiled version of the notes in reading *Introduction to Mathematical Statistics* by Robert Hogg and Allen Craig, *A Guide to Econometrics* by Peter Kennedy and *Econometric Analysis* by William Greene. This is the version as of March 2nd, 2003.

Part 1

Statistical Theory

CHAPTER 1

Probability and Distributions

1. Conditional Probability and Bayes Rule

(1) $\Pr(A\,|B\,) = \frac{\Pr(A\cap B)}{\Pr(B)}$, where $\Pr(B) > 0$.
(2) $\Pr(A \cap B) = \Pr(A\,|B\,)\Pr(B) = \Pr(B\,|A\,)\Pr(A)$.
(3) $\Pr(\cup_i A_i\,|B\,) = \Sigma P(A_i|B)$, if $\cap A_i = \emptyset$.
(4) $\Pr(B|B) = 1$.
(5) $\Pr(A\,|B \cap C\,) = \frac{\Pr(A\cap B\cap C)}{\Pr(B\cap C)}$.
(6) $\Pr(A\cap B\cap C) = \Pr(A\,|B \cap C\,)\Pr(B\cap C) = \Pr(A\,|B \cap C\,)\Pr(B\,|C\,)\Pr(C)$.
(7) $\Pr(\cap_{i=1}^{k} A_i) = \Pr(A_k\,|A_1 \cap A_2 \cap \cdots \cap A_{k-1}\,)$
$\cdot \Pr(A_{k-1}\,|A_1 \cap A_2 \cap \cdots \cap A_{k-2}\,) \cdots \Pr(A_2\,|A_1\,)\Pr(A_1)$.
(8) If $A_1, ..., A_k$ are nested events, i.e., $A_k \subseteq A_{k-1} \subseteq A_{k-1} \cdots \subseteq A_2 \subseteq A_1$, then
$\Pr(\cap_{i=1}^{k} A_i) = \Pr(A_k) = \Pr(A_k A_{k-1})\Pr(A_{k-1}\,|A_{k-2}\,) \cdots \Pr(A_2\,|A_1\,)\Pr(A_1)$.
(9) Let $\cap_i A_i = \emptyset$ and $B \subseteq A_i$, then we have

$$\Pr(A_i\,|B\,) = \frac{\Pr(B\,|A_i)\Pr(A_i)}{\Pr(B)} = \frac{\Pr(B\,|A_i)\Pr(A_i)}{\Sigma_j \Pr(B\,|A_j\,)\Pr(A_j)}.$$

2. Independence

Events A and B are independent if $\Pr(A\,|B\,) = \Pr(A)$, or $\Pr(B\,|A\,) = \Pr(B)$. In this case, $\Pr(A \cap B) = \Pr(A)\Pr(B)$ holds. Events $A_1, ..., A_k$ are mutually independent if $\Pr(A_j\,|\cap_{i\in I_j} A_i\,) = \Pr(A_j)$, where I_j is any subset of A_i excluding A_j. In the case of mutual independence, $\Pr(A \cap B) = 0$.

3. R.V., P.D.F. and C.D.F.

A random variable X is a function $X : C \to B \subseteq \mathbb{R}$. A realization x of the random variable X is a particular value of the random variable associated with a particular outcome of the experiment. $\Pr(X = x)$ is an induced probability.

Suppose that the set of events B is countable, then X is a discrete random variable. Suppose that we can find a function such that

(1) $f(x) \geqslant 0, \forall x \in B$;
(2) $\sum_{x\in B} f(x) = 1$; and
(3) $\Pr(B_i) = \sum_{x\in B_i} f(x)$,

then $f(x)$ is the probability density function (p.d.f.) associated with the random variable X.

Suppose that the set of events B is uncountable, then X is a continuous random variable. If a function $f(x)$ satisfies the following,

(1) $f(x) \geqslant 0, \forall x \in B$;
(2) $\int_B f(x)dx = 1$; and

(3) $\Pr(B_i) = \int_{B_i} f(x)dx$,
then $f(x)$ is the probability density function (p.d.f.) associated with X.

In the case of multivariate random variables, we have $\mathbf{X} : C \to B \subseteq \mathbb{R}^n$, with $\mathbf{X}$ containing components $X_1, ..., X_n$. The respective p.d.f. $f(x)$ has to satisfy the following,

(1) $f(x_1, ..., x_n) \geqslant 0$;
(2) $\int_B f(x_1, ...,_n)dx_1 \cdots dx_n = 1$; and
(3) $\Pr(B_i) = \int_{B_i} f(x_1, ..., x_n)dx_1 \cdots dx_n$.

$F(x) \equiv \Pr(X \leqslant x)$ is the cumulative distribution function (c.d.f.) of the random variable X if the following conditions hold:

(1) $0 \leqslant F(x) \leqslant 1$;
(2) $F(x)$ is non-decreasing in x;
(3) $\Pr(a < X \leqslant b) = F(b) - F(a)$; and
(4) $F(x)$ is right-continuous.
(The definition for right-continuity: $\forall \varepsilon > 0, \exists \delta > 0, \ni |F(x) - F(c)| < \varepsilon$ for $c \leqslant x \leqslant c + \delta$.)

For discrete random variable X, $F(x) = \sum_{k=-\infty}^{x} f(k)$. For multivariate random variables $X_1, ..., X_n$, $F(x_1, ..., x_n) = \Pr(X_1 \leqslant x_1, ..., X_n \leqslant x_n)$.

4. Distribution of a Function of Random Variable(s)

Given a one-to-one increasing function $u(\cdot)$ and $Y = U(X)$, with given c.d.f. for X, $F(x) = \Pr(X \leqslant x)$. The c.d.f. for Y can be found as follows,

$$G(y) = \Pr(Y \leqslant y) = \Pr(u(X) \leqslant y) = \Pr(X \leqslant u^{-1}(y)) = F[u^{-1}(y)].$$

Therefore, the p.d.f. of Y is obtained as $g(y) = G'(y) = f[u^{-1}(y)]\frac{du^{-1}(y)}{dy}$.

If the function $u(\cdot)$ is one-to-one decreasing, then the c.d.f. for Y can be found as follows,

$$\begin{aligned} G(y) &= \Pr(Y \leqslant y) = \Pr(u(X) \leqslant y) = \Pr(X \geqslant u^{-1}(y)) \\ &= 1 - \Pr(X < u^{-1}(y)) = 1 - F[u^{-1}(y)]. \end{aligned}$$

Finally, the p.d.f. of Y can be found as $g(y) = -f[u^{-1}(y)]\frac{du^{-1}(y)}{dy}$.

If we combine the two scenarios above, we know the p.d.f. of a functional transformation $Y = u(X)$ is generally

$$g(y) = f[u^{-1}(y)] \left| \frac{du^{-1}(y)}{dy} \right|.$$

5. Moments, M.G.F. and C.G.F.

The first moment of a random variable X is its expectation $E(X) = \int xf(x)dx$. The second moment of a random variable X is $E(X^2) = \int x^2 f(x)dx$. The variance of a random variable is the centralized second moments, i.e., $var(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2 = \int [x - E(X)]^2 f(x)dx$. The k^{th} raw moment, as opposed to centralized moments, is given by $E(X^k) = \int x^k f(x)dx$.

For any random variable X, the moment generating function (m.g.f.) $M_X(s)$ is defined to be $M_X(s) \equiv E[\exp(sX)]$, and the characteristic generating function (c.g.f.) $C_X(s)$ is defined to be $C_X(s) = E[\exp(isX)]$, where i is the imaginary unit.

Note that not all random variables have a respective moment generating function but all of them do have a respective characteristic generating function. We focus on the moment generating function here.

Why would we bother with m.g.f.? Let's look at some very desirable properties as follows.

$$\begin{aligned}
M_X(s) &= \int e^{sx} f(x)dx \\
\frac{dM_X(s)}{ds} &= \int x \cdot e^{sx} f(x)dx; \frac{dM_X(s)}{ds}\,|_{s=0} = \int x f(x)dx = E(X). \\
\frac{d^2M_X(s)}{ds^2} &= \int x^2 \cdot e^{sx} f(x)dx; \frac{d^2M_X(s)}{ds^2}\,|_{s=0} = \int x^2 f(x)dx = E(X^2). \\
\frac{d^nM_X(s)}{ds^n} &= \int x^n \cdot e^{sx} f(x)dx; \frac{d^nM_X(s)}{ds^n}\,|_{s=0} = \int x^n f(x)dx = E(X^n).
\end{aligned}$$

The name of m.g.f. arises from the fact that $\frac{d^nM_X(s)}{ds^n}\,|_{s=0}$ delivers the n^{th} raw moment for X. It's very important to note that there is a one-to-one correspondence between m.g.f. and the distribution function of the random variable. As long as we can identify the m.g.f. for a random variable (or a function of random variable), the distribution function is uniquely pinned down.

6. Multivariate M.G.F. and Independence

Two random variables X and Y are said to be independent if $f(x,y) = f_X(x)f_Y(y)$ and there is no confounding ranges between X and Y (i.e., the domain of X is independent with the domain of Y). Despite that the independence of X and Y implies $f(x|y) = f_X(x)$ and $f(y|x) = f_Y(y)$, $f(x|y) = f_X(x)$ doesn't necessarily imply the independence of X and Y. The joint m.g.f. for a bivariate case can be defined as $M_{X,Y}(s,t) = E[\exp(sX + tY)]$, and for the multivariate case $M_{X_1,...,X_n}(s_1,...,s_n) = E[\exp(s_1X_1 + \cdots + s_nX_n)]$.

It can be shown easily that

$$\begin{aligned}
\frac{\partial^2 M_{X,Y}(s,t)}{\partial s^2}\,|_{s=0,t=0} &= E(X^2); \\
\frac{\partial^2 M_{X,Y}(s,t)}{\partial t^2}\,|_{s=0,t=0} &= E(Y^2); \\
\frac{\partial^2 M_{X,Y}(s,t)}{\partial s \partial t}\,|_{s=0,t=0} &= E(XY).
\end{aligned}$$

Note the following properties around independence.

(1) $X_1,...,X_n$ are mutually independent iff $f(x_1,...,x_n) = g_1(x_1)\cdots g_n(x_n)$ plus no confounding ranges.
(2) If $X_1,...,X_n$ are mutually independent, then $\Pr(x_1 \in A_1,...,x_n \in A_n) = \Pr(x_1 \in A_1)\cdots\Pr(x_n \in A_n)$.
(3) If $X_1,...,X_n$ are mutually independent, then $E[u_1(X_1)u_2(X_2)\cdots u_n(X_n)] = E[u_1(X_1)]\cdot E[u_2(X_2)]\cdots E[u_n(X_n)]$.
(4) The mutual independence of $X_1,...,X_n$ is equivalent to $M_{X_1,...X_n}(s_1,...,s_n) = M_{X_1}(s_1)M_{X_2}(s_2)\cdots M_{X_n}(s_n)$.

7. Weak Law of Large Numbers

The weak law of large numbers indicates that the sample mean of any random variable converges to its true population mean in probability. In math notations, it says $\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - \mu\right| \geqslant \varepsilon\right] = 0, \forall \varepsilon > 0$.

To prove the weak law of large numbers, we need something known as Chebyshev's Inequality, which is just a special case of Markov's Inequality. Covered below are both theorems.

Markov's Inequality says that for $u(X) \geqslant 0$, $\Pr[u(X) \geqslant c] \leqslant \frac{E[u(X)]}{c}$ holds. Why is it so? Let $A \equiv \{x : u(X) \geqslant c\}$, then $E[u(X)] = \int u(x)f(x)dx \geqslant \int_A u(x)f(x)dx \geqslant \int_A cf(x)dx = c\Pr(u(X) \geqslant c)$ obviates the aforementioned inequality.

Chebyshev's Inequality says the following. If X is a random variable with finite mean $E(x) = \mu$ and finite variance $var(X) = \sigma^2 < \infty$, then $\Pr(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}$ holds. Why is it so? Note first that $E[(X-\mu)^2] = \sigma^2$ and $\Pr(|X - \mu| \geqslant k\sigma) = \Pr[(X-\mu)^2 \geqslant k^2\sigma^2]$. Using Markov's Inequality, we have $\Pr[(X-\mu)^2 \geqslant k^2\sigma^2] \leqslant \frac{E(X-\mu)^2}{k^2\sigma^2} = \frac{1}{k^2}$. Hence Chebyshev's Inequality follows.

To prove the weak law of large numbers, note that

$$\Pr\left[\left|\frac{X_n}{n} - \mu\right| \geqslant \varepsilon\right] = \Pr\left(|X_n - n\mu| \geqslant n\varepsilon\right) = \Pr\left[|X_n - n\mu| \geqslant \left(\frac{n\varepsilon}{\sigma}\right)\sigma\right] \leqslant \frac{\sigma^2}{n^2\varepsilon^2},$$

and thus $\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - \mu\right| \geqslant \varepsilon\right] = 0$.

8. List of Common Distributions

(1) Bernoulli: one experiment, two possible outcomes.
$f(x) = p^x(1-p)^{1-x}$, x(success) $= 1$, x(failure) $= 0$.

(2) Binomial: exactly x successes out of n trials of Bernoulli experiments.
$f(x) = \binom{n}{x}p^x(1-p)^{n-x}$, $\mu = np$, $\sigma^2 = np(1-p)$,
$M(t) = (1-p+pe^t)^n$.

(3) Negative Binomial: exactly x failures before the k^{th} success.
$f(x) = \binom{x+k-1}{k-1}p^k(1-p)^x$.

(4) Geometric: exactly x failures before the first success.
$f(x) = p(1-p)^x$, $\mu = \frac{1-p}{p}$, $\sigma^2 = \frac{1-p}{p^2}$, $M(t) = \frac{pe^s}{1-e^s(1-p)}$.

(5) Hyper-geometric: N balls with R red ones, drawing exactly x red out of n draws without replacement.
$f(x) = \binom{R}{x}\binom{N-R}{n-x}/\binom{N}{n}$.

(6) Pareto:
$f(x,\theta) = \theta \cdot x_0^\theta \cdot x^{-(\theta+1)}$, where $x \geqslant x_0$, $\mu = \frac{x_0\theta}{\theta-1}$.

(7) Trinomial: one experiment, three possible outcomes, repeat n times.
$f(x,y) = \frac{n!}{x!y!(n-x-y)!}p_1^x p_2^y p_3^{n-x-y}$, where $p_1 + p_2 + p_3 = 1$.
$M(t_1,t_2) = (p_1e^{t_1} + p_2e^{t_2} + p_3)^n$.

(8) Multinomial:
$f(x_1,...,x_n) = \frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$, where $\Sigma p_k = 1$ and $\Sigma x_k = n$.
$M(t_1,...,t_n) = (p_1e^{t_1} + \cdots + p_{k-1}e^{t_{k-1}} + p_k)^n$.

(9) Poisson: (special case of Binomial distribution, $n \to \infty$, $p \to 0$, $np = \lambda$)
$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, ...$ $\mu = \sigma^2 = \lambda$.
$M(t) = \exp[\lambda(e^t - 1)]$.

(10) Gamma:
$f(x,\alpha,\beta) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\Gamma(\alpha)\beta^{\alpha}}$, where $\alpha \geqslant 0, \alpha, \beta > 0$.
$\mu = \alpha\beta$, $\sigma^2 = \alpha\beta^2$, $M(t) = (1-\beta t)^{-\alpha}$, where $t < \frac{1}{\beta}$.

(11) Chi-square: (special case of Gamma distribution, $\alpha = \frac{r}{2}$ and $\beta = 2$)
$f(x) = \frac{1}{\Gamma(\frac{r}{2})\sqrt{2}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}$, $M(t) = (1-2t)^{-\frac{r}{2}}$, where $t < \frac{1}{2}$.

(12) Exponential: (special case of Gamma distribution, $\alpha = 1$ and $\beta = \frac{1}{\lambda}$)
$f(x) = \lambda \exp(-\lambda x)$, where $x > 0$, $\mu = \frac{1}{\lambda}$, $\sigma^2 = \frac{1}{\lambda^2}$,
$M(t) = \frac{\lambda}{\lambda - t}$, $(t < \lambda)$.

(13) Normal:
$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$, $M(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

9. Basic Rules on First Two Moments

(1) $E(x) = \int x f_x(x) dx = \int_x \int_y x f(x,y) dy dx$;
(2) $E[g(x,y)] = \int_x \int_y g(x,y) f(x,y) dy dx$;
(3) $E(ax+by+c) = aE(x) + bE(y) + c$;
(4) $Var(ax+by+c) = a^2 Var(x) + b^2 Var(y) + 2ab Cov(x,y) = Var(ax+by)$;
(5) $Cov(ax+by, cx+dy) = acVar(x) + bdVar(y) + (ad+bc)Cov(x,y)$
(bi-linearity of covariance)

10. Bi-variate Normal Distribution

Random variables X and Y follow a bi-variate normal distribution with parameters μ_X, μ_Y, σ_X^2, σ_Y^2, ρ, if

$$f(x,y) = \frac{1}{2\pi}[\sigma_X^2 \sigma_Y^2 (1-\rho^2)]^{-\frac{1}{2}}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\frac{\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

where $\rho^2 < 1$, $-\infty < x < \infty$, $-\infty < y < \infty$.

The above expression in matrix notation is much simpler.

$$\mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix}, \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$
$$f(\mathbf{z}) = \frac{1}{2\pi}|\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mu)'\Sigma^{-1}(\mathbf{z}-\mu)\right].$$

Here are some most important properties of bi-variate and multi-variate normal.

(1) the marginal density of a bi-variate (or multi-variate) normal distribution is still normally distributed. That is, $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.
(2) the conditional density of a bi-variate (or multi-variate) normal distribution is still normally distributed. That is,

$$X|y \sim N\left[\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),\ \sigma_X^2(1-\rho^2)\right],$$
$$Y|x \sim N\left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\ \sigma_Y^2(1-\rho^2)\right].$$

(3) The regression among jointly normally distributed variables is linear!
(4) The moment generating function is $M_{\mathbf{z}}(\mathbf{s}) = \exp\left\{\mu'\mathbf{s} + \frac{1}{2}\mathbf{s}'\Sigma\mathbf{s}\right\}$.

11. Conditioning in a Bi-variate Distribution

(1) The conditional mean $E(y|x)$ is called the regression of y on x. A random variable may always be written as $y = E(y|x) + [y - E(y|x)] = E(y|x) + \varepsilon$.
(2) Conditional variance: $Var(y|x) = E(y^2|x) - [E(y|x)]^2$. If $Var(y|x)$ doesn't change with x, it is called homoskedasticity, i.e., $Var(\varepsilon|x) = \sigma^2$.
(3) Law of iterative expectations:
$E(y) = E_x[E(y|x)]$;
$Cov(x, y) = Cov[x, E(y|x)] = \int_x [x - E(x)]E(y|x)f_x(x)dx$.
   (a) To prove the first result, start from $\int_y yf(x, y)dy = E(y|x)f_x(x)$, then $E(y) = \int_x \int_y yf(x, y)dydx = \int_x E(y|x)f_x(x)dx = E_x[E(y|x)]$.
   (b) For the second result, start from $\int_y yf(x, y)dy = \int_y E(y|x)f(x, y)dy$ and $E(y) = E_x[E(y|x)]$, then $Cov(x, y) = \int_x \int_y [x - E(x)][y - E(y)]$ $f(x, y)dydx = \int_x \int_y [x - E(x)]\{E(y|x) - E_x[E(y|x)]\}f(x, y)dydx$ $= Cov[x, E(y|x)]$.
(4) Variance decomposition: $Var(y) = Var_x[E(y|x)] + E_x[Var(y|x)]$.
The variance of y can be decomposed as the variance of the conditional mean and the expected variance of y around the conditional mean. The first term on the RHS is the regression variance, similar to SSR, and the second term on the RHS is the residual variance, similar to SSE.

Note that $\int_y yf(x, y)dy = \int_y yf(y|x)f_x(x)dy = \left[\int_y yf(y|x)dy\right] f_x(x)$ $= E(y|x)f_x(x) = E(y|x)\int_y f(x, y)dy = \int_y E(y|x)f(x, y)dy$.

(1) (a) To prove $Var(y) = Var_x[E(y|x)] + E_x[Var(y|x)]$, we have

$$\begin{aligned}[y - E(y)]^2 &= \{[y - E(y|x)] + [E(y|x) - E(y)]\}^2 \\ &= [y - E(y|x)]^2 + [E(y|x) - E(y)]^2 \\ &\quad + 2[y - E(y|x)][E(y|x) - E(y)].\end{aligned}$$

Therefore, we have

$$\begin{aligned}Var(y) &= E_x Var(y|x) + Var_x[E(y|x)] \\ &\quad + 2\{E_x[E(y|x)]^2 - E_x[E(y|x)]^2 - [E(y)]^2 + [E(y)]^2\} \\ &= E_x Var(y|x) + Var_x[E(y|x)].\end{aligned}$$

(2) If $E(y|x) = \alpha + \beta x$, then $Cov(x, y) = Cov[x, E(y|x)] = Cov(x, \alpha + \beta x) = \beta Var(x)$, so $\beta = Cov(x, y)/Var(x)$ and $Var_x[E(y|x)] = \beta^2 Var(x) = \rho_{xy}^2 Var(y)$.
(3) $E_x[Var(y|x)] = Var(y) - Var_x[E(y|x)]$
On average, conditioning reduces the variance of the variable subject to the conditioning.
(4) If $E(y|x) = \alpha + \beta x$ and $Var(y|x)$ is a constant, then $Var(y|x) = \sigma_y^2(1 - \rho_{xy}^2)$.
(5) The coefficient of determination (COD) is equal to the ratio of regression variance to the total variance. If $E(y|x) = \alpha + \beta x$, then $COD(= R^2) = \rho^2$.

12. Multivariate Distribution

(1) $E(\mathbf{x}) = \mu, \quad Var(\mathbf{x}) = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)'] = \Sigma$ (variance-covariance matrix)
By dividing σ_{ij} by $\sigma_i\sigma_j$, we obtain the correlation matrix $\mathbf{R}$.

(2) $E(\mathbf{Ax}) = \mathbf{A}E(\mathbf{x}) = \mathbf{A}\mu; \quad Var(\mathbf{Ax}) = \mathbf{A}\Sigma\mathbf{A}'$
Particularly, if $(n-K)\frac{S^2}{\sigma^2} \sim \chi^2(n-K)$, then $E[(n-K)\frac{S^2}{\sigma^2}] = n-K$, from $E[\chi^2(\lambda)] = \lambda$, implies $E(S^2) = \sigma^2$; and $Var[(n-K)\frac{S^2}{\sigma^2}] = 2(n-K)$, from $Var[\chi^2(\lambda)] = 2\lambda$, implies $Var(S^2) = \frac{2\sigma^4}{n-K}$.

13. Multivariate Normal Distribution

(1) If $\mathbf{x} \sim N(\mu, \Sigma)$, then $\mathbf{Ax} + \mathbf{b} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}')$.

(2) Quadratic forms in a standard normal vector.
   (a) If $\mathbf{x} \sim N(\mathbf{0}, \mathrm{I})$, $\mathbf{C}$ is square, and $\mathbf{C}'\mathbf{C} = \mathbf{I}$(i.e., $\mathbf{C}$ is an orthogonal matrix), then $\mathbf{C}'\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$.
   (b) If $\mathbf{x} \sim N(\mathbf{0}, \mathrm{I})$ and $\mathbf{A}$ is idempotent, then $\mathbf{x}'\mathbf{Ax} \sim \chi^2(J)$, where J is the rank of $\mathbf{A}$.
   (c) $\Sigma(x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x}$, and $rank(M^0) = tr(M^0) = n(1 - \frac{1}{n}) = n - 1$, thus $\Sigma(x_i - \bar{x})^2 \sim \chi^2(n-1)$.
   (d) $n\bar{x}^2 = \mathbf{x}'(\mathbf{jj}')\mathbf{x}$, where $\mathbf{j} = \frac{1}{\sqrt{n}}\mathbf{i}$. It could be verified that $\mathbf{jj}'$ is idempotent with rank of 1, so $n\bar{x}^2 \sim \chi^2(1)$.
   (e) $\Sigma x_i^2 = \Sigma(x_i - \bar{x})^2 + n\bar{x}^2 \Leftrightarrow \mathbf{x}'\mathbf{x} = \mathbf{x}'\mathbf{M}^0\mathbf{x} + \mathbf{x}'(\mathbf{I} - \mathbf{M}^0)\mathbf{x} \Leftrightarrow \chi^2(n) = \chi^2(n-1) + \chi^2(1)$.
   (f) If $\mathbf{x} \sim N(\mathbf{0}, \mathrm{I})$, $\mathbf{A}$ and $\mathbf{B}$ are idempotent, then $\mathbf{x}'\mathbf{Ax}$ and $\mathbf{x}'\mathbf{Bx}$ are independent if $\mathbf{AB} = \mathbf{0}$.
   Let $\mathbf{x}_1 = \mathbf{Ax}$ and $\mathbf{x}_2 = \mathbf{Bx}$, then $\mathbf{x}'\mathbf{Ax} = \mathbf{x}_1'\mathbf{x}_1$ and $\mathbf{x}'\mathbf{Bx} = \mathbf{x}_2'\mathbf{x}_2$. Since $Cov(\mathbf{x}_1, \mathbf{x}_2) = E[(\mathbf{Ax})(\mathbf{Bx})'] - \mathbf{0} = \mathbf{A}Var(\mathbf{x})\mathbf{B}' = \mathbf{AB}$, $\mathbf{AB} = \mathbf{0}$ would imply the independence of the two quadratic forms.
   (g) To prove that $\Sigma(x_i - \bar{x})^2$ and $n\bar{x}^2$ are independent, it suffices to prove $\mathbf{M}^0(\mathbf{I} - \mathbf{M}^0) = \mathbf{0}$, which is apparently true.

(3) F distribution
If $\mathbf{x} \sim N(\mathbf{0}, \mathrm{I})$, $\mathbf{A}$ and $\mathbf{B}$ are idempotent with rank r_A and r_B, then $\mathbf{AB} = \mathbf{0}$ implies that

$$\frac{\mathbf{x}'\mathbf{Ax}/r_A}{\mathbf{x}'\mathbf{Bx}/r_B} \sim F(r_A, r_B).$$

Extension to $\mathbf{x} \sim N(0, \sigma^2\mathrm{I})$:

$$\frac{\mathbf{x}'\mathbf{Ax}/(\sigma^2 r_A)}{\mathbf{x}'\mathbf{Bx}/(\sigma^2 r_B)} \sim F(r_A, r_B).$$

(4) Full rank quadratic form
If $\mathbf{x} \sim N(\mu, \Sigma)$, then $\Sigma^{-\frac{1}{2}}(\mathbf{x} - \mu) \sim N(\mathbf{0}, \mathbf{I})$ and $(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu) \sim \chi^2(N)$.

(5) If $\mathbf{x} \sim N(\mathbf{0}, \mathrm{I})$ and $\mathbf{A}$ is idempotent, then $\mathbf{Lx}$ and $\mathbf{x}'\mathbf{Ax}$ are independent if $\mathbf{LA} = \mathbf{0}$.

(6)

$$t(J) = \frac{N(0,1)}{[\chi^2(J)/J]^{\frac{1}{2}}}.$$

(7) $\sqrt{n}\bar{x} \sim N(0,1)$ and $\Sigma(x_i - \bar{x})^2 \sim \chi^2(n-1)$ implies that

$$\frac{\sqrt{n}\bar{x}}{[\Sigma(x_i - \bar{x})^2/(n-1)]^{\frac{1}{2}}} \sim t(n-1),$$

i.e., $\sqrt{n}\bar{x}/S \sim t(n-1)$.

CHAPTER 2

Distribution of Functions of Random Variables

1. Basic Definitions

That X_1, ..., X_n is a random sample of size n is equivalent to the statement that X_1, ..., X_n are independently identically distributed (i.i.d.) random variables.

A statistic is a function of one or more random variables that doesn't depend upon any unknown parameters. Note that a statistic is a random variable and that the p.d.f. of a statistic may involve unknown parameters.

If the addition of two random variables with the same distribution has the same distribution as the original ones, then we say this distribution has reproductive property. For example, Gamma, Poisson and Chi-square distributions are all reproductive.

2. General Approaches

Given that X_1, ..., X_n is a random sample from p.d.f. $f(x)$. Let $Y = u(X_1, ..., X_n)$. How can we find out the p.d.f. $g(y)$? There are two basic techniques.

(1) Direct approach
(Find a good enough one-to-one transformation, calculate the Jacobian of the inverse transformation, find the joint density and finally get the marginal density. Note that in the last step we have to pay extra care to the range of variables.)

(2) m.g.f. approach
(Try to use the m.g.f. corresponding to $f(x)$ to figure out the m.g.f. for Y. Identify the distribution of Y by identify its m.g.f. This approach is typically simpler than the direct approach.)

3. Direct Approach

Given that X_1, ..., X_n is a random sample from p.d.f. $f(x_1, ..., x_n; \theta)$. Let $Y_1 = u_i(X_1, ..., X_n)$. How can we find out the p.d.f. $g(y_1; \theta)$? Here are the steps for the direct approach.

(1) Construct a good enough one-to-one transformation system between Y_i and X_i, i.e., $Y_i = u_i(X_1, ..., X_n)$ and $X_i = w_i(Y_1, ..., Y_n)$, where $i = 1, ..., n$. Note that the "goodness" of the transformation depends on the construction of $Y_i = u_i(X_1, ..., X_n)$ for $i \neq 1$.

(2) Calculate the Jacobian $J = \left| \frac{\partial w_i}{\partial Y_i} \right|$.

(3) Find the joint density $g(y_1, ..., y_n) = abs(|J|) f[w_1(\cdot), ..., w_n(\cdot)]$.

(4) Find the marginal density $g(y_1)$.

Here are some examples for exercise.

Example 1. X_1, ..., X_n is a random sample from $f(x_1, ..., x_n; \theta) = \theta x^{\theta-1}$, $0 \leqslant x \leqslant 1$, $\theta > 1$. Let $Y_1 = X_1 X_2 \cdots X_n$. Find $g(y_1)$.

Example 2. $X_1 \sim f(x_1)$ and $X_2 \sim f(x_2)$ are independent, and $Y = X_1 + X_2$. Find $g(y)$.

4. M.G.F. Approach

Given that X_1, ..., X_n is a random sample from p.d.f. $f(x)$. Let $Y = u(X_1, ..., X_n)$. How can we find out the p.d.f. $g(y)$? Basically we are trying to use the m.g.f. corresponding to X to figure out the m.g.f. for Y. Finally we can identify the distribution of Y by identify its m.g.f. This approach is typically simpler than the direct approach.

Example 1. $X \sim N(0, 1)$, prove that $X^2 \sim \chi^2(1)$.

Proof: Denote $Y \equiv X^2$. Let's find out the m.g.f. for Y as follows.

$$\begin{aligned}
M_Y(t) &= E[\exp(tX^2)] \\
&= \int_{-\infty}^{+\infty} \exp(tx^2) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(1-2t)x^2\right) dx \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{1-2t}x)^2\right) d\sqrt{1-2t}x \\
&= \frac{1}{\sqrt{1-2t}} \\
&= (1-2t)^{-\frac{1}{2}}.
\end{aligned}$$

Clearly $M_Y(t)$ is the m.g.f. of Chi-square distribution with $r = 1$, i.e., $X^2 \sim \chi^2(1)$.

Example 2. X_1, ..., X_n is a random sample from $N(0, 1)$, prove that $\Sigma_i X_i^2 \sim \chi^2(T)$.

Proof: We know from Example 1 that $X_i^2 \sim \chi^2(1)$ and $M_{X^2}(s) = (1-2s)^{-\frac{1}{2}}$. Denote $Y \equiv \Sigma_i X_i^2$ and let's find the m.g.f. for Y.

$$\begin{aligned}
M_Y(s) &= E[\exp(sY)] \\
&= E[\exp(sX_1^2)\exp(sX_2^2)\cdots\exp(sX_n^2)] \\
&= E[\exp(sX_1^2)]E[\exp(sX_2^2)]\cdots E[\exp(sX_n^2)] \\
&= M_{X_1^2}(s)M_{X_2^2}(s)\cdots M_{X_n^2}(s) \\
&= (1-2s)^{-\frac{1}{2}n}.
\end{aligned}$$

Clearly $M_Y(s)$ is the m.g.f. of Chi-square distribution with $r = n$, i.e., $\Sigma_i X_i^2 \sim \chi^2(T)$.

Example 3. $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$ are independent. Prove that $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^1 + \sigma_2^2 - 2\sigma_{12})$.

Example 4. $X_1, ..., X_n$ is a random sample from $N(\mu, \sigma^2)$. Prove that $\Sigma_i X_i \sim N(n\mu, n\sigma^2)$.

Example 5. $X_1, ..., X_n$ is a random sample from $N(\mu, \sigma^2)$. Prove that $\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2)$.

Example 6. $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$ are independent, prove that $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

5. Order Statistics

X_1, ..., X_n is a random sample from $f(x)$, and Y_1, ..., Y_n are the t^{th} order statistics, i.e., $y_1 < y_2 < \cdots < y_n$. The joint density and marginal density of the order statistics are given by:

(1) $g(y_1, ..., y_n) = n! f(y_1) f(y_2) \cdots f(y_n), -\infty < y_1 < y_2 < \cdots < y_n < +\infty$;
(2) $g_i(y_i) = \frac{n!}{(n-i)!(i-1)!}[F(y_i)]^{i-1}[1 - F(y_i)]^{n-i} f(y_i)$.

That Y_i is the i^{th} order statistic requires two things, the i^{th} position is reserved for Y_i and y_i is ranked as the i^{th} item. Here is an intuitive account for the marginal density of the i^{th} order statistic in terms of permutation.

Consider a line of n slots. First, populate the first $(i-1)$ slots and the probability is $\binom{n}{i-1} = \frac{n!}{(i-1)!(n-i+1)!}$. Next, pick one lucky item out of the remaining $(n-i+1)$ items and put it on the i^{th} slot. The probability for this step is $\binom{n-i+1}{1} = n - i + 1$. The two steps above complete the process of securing the i^{th} slot in the line of n slots, delivering a probability of $\binom{n}{i-1}\binom{n-i+1}{1} = \frac{n!}{(n-i)!(i-1)!}$.

Note that merely securing the i^{th} slot is not enough to get the i^{th} order statistic, and we have to make sure that the rank is in line with the position. That is, all previous $(i-1)$ items are indeed smaller than y_i, with probability $\Pr(Y_j \leqslant y_i, 1 \leqslant j \leqslant i-1) = [F(y_i)]^{i-1}$, and the future $(n-i)$ items are indeed larger than y_i, with probability $\Pr(Y_k \geqslant y_i, i+1 \leqslant k \leqslant n) = [1 - F(y_i)]^{n-i}$.

Combining these two components, the marginal density of the i^{th} order statistics can be written as $g_i(y_i) = \frac{n!}{(n-i)!(i-1)!}[F(y_i)]^{i-1}[1 - F(y_i)]^{n-i} f(y_i)$.

6. Student-t, F and Sampling Distribution

(1) If $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2(r)$ are independent, then $\frac{X_1}{\sqrt{X_2/r}}$ follows a student-t distribution with r degrees of freedom, i.e., $\frac{X_1}{\sqrt{X_2/r}} \sim t(r)$.
(2) If $X_1 \sim \chi^2(r)$ and $X_2 \sim \chi^2(s)$ are independent, then $\frac{X_1/r}{X_2/s} \sim F(r,s)$.
(3) If X_1, ..., X_n is a random sample from $N(\mu, \sigma^2)$ and $\overline{X} = \frac{1}{n}\Sigma_i X_i$ and $S^2 = \frac{1}{n}\Sigma_i (X_i - \overline{X})^2$, then we have
   (a) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, which implies that $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$.
   (b) $\overline{X}$ and S^2 are independent.
   (c) $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$, which implies that $\frac{\sqrt{n-1}(\overline{X}-\mu)}{S} \sim t(n-1)$.

CHAPTER 3

Limiting Distributions

Note that in this chapter we use X_n as a short-hand for the random sample $X_1, ..., X_n$. We will explain the definitions of convergence in probability, convergence in distribution, convergence in mean square, rules for probability limit, and finally the central limit theorem.

1. Convergence

(1) Convergence in Distribution

A random sample X_n (with c.d.f. $F_n(x)$) converges in distribution to a random variable X (with c.d.f. $F(x)$), if $\lim_{n\to\infty} |F_n(x) - F(x)| = 0$ for every point x at which $F(x)$ is continuous. It is denoted as $X_n \xrightarrow{d} X$.

(2) Convergence in Probability

X_n converges in probability to a constant c iff X_n is getting ever closer to c as $n \to \infty$. That is, $p\lim X_n = c \Leftrightarrow \lim_{n\to\infty} \Pr(|X_n - c| \geqslant \varepsilon) = 0, \forall \varepsilon > 0$ or equivalently $X_n \xrightarrow{p} c \Leftrightarrow \lim_{n\to\infty} \Pr(|X_n - c| < \varepsilon) = 1, \forall \varepsilon > 0$.

(3) Convergence in Mean Square

A random sample X_n has mean μ_n and variance σ_n^2. If $\lim_{n\to\infty} \mu_n = c$ and $\lim_{n\to\infty} \sigma_n^2 = 0$, then X_n converges in mean square to c. It's denoted by $p\lim X_n = c$. This result can be easily proved using Chebyshev's Inequality, covered in an earlier chapter.

(4) Relationship between Three Forms of Convergences

If a random sample X_n converges in distribution to a degenerate distribution X that has probability 1 at point c, then $X_n \xrightarrow{p} c$ or $p\lim X_n = c$. If a random sample converges in mean square to c, then it also converges in probability to c, but the converse need not be true.

2. Rules for Probability Limit

(1) Slutsky Theorem: For continuous function $g(x_n)$ that is not a function of n, we have $p\lim[g(x_n)] = g[p\lim(x_n)]$.

(2) $p\lim x_n = c$ and $p\lim y_n = d$ imply the following: $p\lim(x_n + y_n) = c + d$; $p\lim(x_n y_n) = cd$; $p\lim(\frac{x_n}{y_n}) = \frac{c}{d}$, if $d \neq 0$.

(3) If $\mathbf{W}_n$ is a matrix whose elements are random variables, i.e., random matrix and $p\lim \mathbf{W}_n = \Omega$, then $p\lim \mathbf{W}_n^{-1} = \Omega^{-1}$.

(4) If $\mathbf{X}_n$ and $\mathbf{Y}_n$ are random matrices, and $p\lim \mathbf{X}_n = \mathbf{A}, p\lim \mathbf{Y}_n = \mathbf{B}$, then $p\lim \mathbf{X}_n\mathbf{Y}_n = \mathbf{AB}$.

3. Rules for Limiting Distributions

(1) If $x_n \xrightarrow{d} x$ and $p\lim y_n = c$, then $x_n y_n \xrightarrow{d} cx$, $x_n + y_n \xrightarrow{d} x + c$, and $\frac{x_n}{y_n} \xrightarrow{d} \frac{x}{c}$, $(c \neq 0)$.
(2) If $x_n \xrightarrow{d} x$ and $g(x_n)$ is a continuous function, then $g(x_n) \xrightarrow{d} g(x)$. Particularly, $F(1, n) = t^2(n) \xrightarrow{d} \chi^2(1)$.
(3) If y_n has a limiting distribution and $p\lim(x_n - y_n) = 0$, then x_n has the same limiting distribution as y_n.

4. Methods of Finding Limiting Distributions

Generally speaking, there are two ways to do it. One is to calculate the probability limits and the other is to calculate the limiting moment generating function. Here are the steps for the plim approach:

(1) find $F_n(x)$; (We may need the method of finding the distribution of functions of random variables, covered in an earlier chapter.)
(2) calculate $\lim_{n\to\infty} F_n(x)$ and define $F(x) \equiv \lim_{n\to\infty} F_n(x)$;
(3) identify what distribution X has corresponding to $F(x)$;
(4) conclude that $X_n \xrightarrow{d} X$.

In particular, if X has any one of the following properties:

(a) $f(x) = \begin{cases} 1, & \text{if } x = \mu; \\ 0, & \text{if } x \neq \mu; \end{cases}$
(b) $F(x) = \begin{cases} 1, & \text{if } x \geqslant \mu; \\ 0, & \text{if } x < \mu; \end{cases}$
(c) $M_X(s) = \exp(s\mu)$,

then conclude that $X_n \xrightarrow{p} \mu$ or $X_n \xrightarrow{d} X$, where X is a degenerate distribution that has probability of 1 at point $x = \mu$.

Here are the steps for the limiting m.g.f. approach.

(1) find $M_{X_n}(s)$;
(2) calculate $\lim_{n\to\infty} M_{X_n}(s)$ and define $M_X(s) \equiv \lim_{n\to\infty} M_{X_n}(s)$;
(3) identify what distribution X has according to $M_X(s)$;
(4) conclude that $X_n \xrightarrow{d} X$.

In particular, if $M_{X_n}(s) = \left[M_Z\left(\frac{s}{n}\right)\right]^n$ or $M_{X_n}(s) = \left[M_Z\left(\frac{s}{\sqrt{n}}\right)\right]^n$, take advantage of the following two tricks.

(a) Use Taylor expansion as follows,

$$\begin{aligned} M_Z\left(\tfrac{s}{n}\right) &= M_Z(0) + \tfrac{s}{n}M_Z'(0) + o(\tfrac{1}{n}) \\ &= 1 + \tfrac{s}{n}E(Z) + o(\tfrac{1}{n}); \\ M_Z\left(\tfrac{s}{\sqrt{n}}\right) &= M_Z(0) + \tfrac{s}{\sqrt{n}}M_Z'(0) + \tfrac{s^2}{2n}M_Z''(0) + o(\tfrac{1}{n}) \\ &= 1 + \tfrac{s}{\sqrt{n}}E(Z) + \tfrac{s^2}{2n}E(Z^2) + o(\tfrac{1}{n}); \end{aligned}$$

(b) Use $\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = e^a$, when computing $\lim_{n\to\infty} M_{X_n}(s)$.

5. Central Limit Theorems

(1) Univariate, same distribution, finite μ and σ^2. $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

(2) Univariate, different distribution, finite μ_i and σ_i^2. If $\lim_{n\to\infty} \bar{\sigma}_n^2 = \bar{\sigma}^2$, i.e., $\lim_{n\to\infty} Var(\sqrt{n}\bar{X}_n) = \bar{\sigma}^2$, then $\sqrt{n}(\bar{X}_n - \bar{\mu}_n) \xrightarrow{d} N(0, \bar{\sigma}^2)$.
(3) Multivariate, same distribution, finite vector μ and finite positive definite covariance matrix $\mathbf{Q}$. $\sqrt{n}(\bar{\mathbf{X}}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q})$.
(4) Multivariate, different distribution, finite vector μ_i and finite positive definite covariance matrix $\mathbf{Q}_i$. If $\lim_{n\to\infty} Var(\bar{\mathbf{Q}}_n) = \mathbf{Q}$, then $\sqrt{n}(\bar{\mathbf{X}}_n - \bar{\mu}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q})$.
(5) Limiting normal distribution of a function. If $\sqrt{n}(Z_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, and if $g(z_n)$ is a continuous function not involving n, then $\sqrt{n}[g(Z_n) - g(\mu)] \xrightarrow{d} N\{0, [g'(Z_n)]^2\sigma^2\}$.
(6) Limiting normal distribution of a set of functions. If $\sqrt{n}(\mathbf{Z}_n - \mu) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ and $\mathbf{c}(\mathbf{Z}_n)$ is a set of J continuous functions not involving n, then $\sqrt{n}[\mathbf{c}(\mathbf{Z}_n) - \mathbf{c}(\mu)] \xrightarrow{d} N(\mathbf{0}, \mathbf{C}\Sigma\mathbf{C}')$, where $\mathbf{C} = \frac{\partial \mathbf{c}(\mathbf{Z}_n)}{\partial \mathbf{Z}_n'}$.

6. Asymptotic Distributions

(1) If $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, then $\bar{X}_n \xrightarrow{a} N(\mu, \frac{\sigma^2}{n})$.
(2) If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$, then $\hat{\theta} \xrightarrow{a} N(\theta, \frac{\mathbf{V}}{n})$. $Asy.Var(\hat{\theta}) = \frac{\mathbf{V}}{n}$.
(3) If $\hat{\theta} \xrightarrow{a} N(\theta, \frac{\sigma^2}{n})$ and $g(\theta)$ is a continuous function not involving n, then $g(\hat{\theta}) \xrightarrow{a} N\{g(\theta), [g'(\theta)]^2\sigma^2\}$.
(4) If $\hat{\theta} \xrightarrow{a} N(\theta, \frac{\mathbf{V}}{n})$ and $\mathbf{c}(\theta)$ is a set of J continuous functions not involving n, then

$$\mathbf{c}(\hat{\theta}) \xrightarrow{a} N[\mathbf{c}(\theta), \frac{\mathbf{CVC}'}{n}],$$

where $\mathbf{C} = \frac{\partial \mathbf{c}(\theta)}{\partial \theta'}$.

7. Example of Limiting Distribution

A random sample X_n has mean μ and variance $\sigma^2 < \infty$. Prove that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S} \xrightarrow{d} N(0, 1).$$

(1) By Central Limit Theorem, we have $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$;
(2) Prove $p\lim(S^2) = \sigma^2$ as follows.

$$\begin{aligned}
S^2 &= \tfrac{1}{n}\Sigma(X_t - \bar{X})^2 = \tfrac{1}{n}\Sigma X_t^2 - \bar{X}^2 \\
p\lim(\tfrac{1}{n}\Sigma X_t^2) &= p\lim(\overline{X_t^2}) = E(X^2) \\
p\lim[(\bar{X})^2] &= [p\lim(\bar{X})]^2 = [E(X)]^2 \\
p\lim(S^2) &= p\lim(\tfrac{1}{n}\Sigma X_t^2) - p\lim(\bar{X}^2) \\
&= E(X^2) - [E(X)]^2 \\
&= \sigma^2
\end{aligned}$$

Note that we have twice deployed the weak law of large number, which says that $p\lim(\bar{Y}_n) = E(Y)$ as long as the population distribution has finite mean and variance. Refer to an ealier chapter for detailed coverage.

(3) From $p\lim(S^2) = \sigma^2$, we know $p\lim(S/\sigma) = 1$ or $S/\sigma \xrightarrow{p} 1$.
(4) Using results from steps (1) and (3), we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma \cdot \frac{S}{\sigma}} \xrightarrow{d} N(0,1).$$

CHAPTER 4

Statistical Inference

1. Desirable Statistic Properties

(1) Unbiased estimator

In the case of single parameter θ, the estimator $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$; in the case of vector θ, it requires equality of element by element.

(2) Efficient estimator

In the case of single parameter θ, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$. In the case of vector θ, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $Var(\hat{\theta}_2) - Var(\hat{\theta}_1)$ is a nonnegative definite matrix.

(3) Efficiency

If the case of single parameter, $\hat{\theta}$ is efficient if it achieves the CRLB, $[I(\theta)]^{-1}$, where

$$I(\theta) = -E(\tfrac{\partial^2 \ln L}{\partial \theta^2}) = E(\tfrac{\partial \ln L}{\partial \theta})^2.$$

In the case of vector θ,

$$\mathbf{I}(\theta) = -E(\tfrac{\partial^2 \ln L}{\partial \theta \partial \theta'}) = E[(\tfrac{\partial \ln L}{\partial \theta})(\tfrac{\partial \ln L}{\partial \theta'})].$$

(4) For normal distribution, $\left[I\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right]^{-1} = \begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{bmatrix}$.

Let $\hat{\mu}$ and $\hat{\sigma}^2$ be unbiased estimators, and $Var(\hat{\mu}\ \sigma^2) = \mathbf{V}$, then $\mathbf{V} - [\mathbf{I}(\hat{\mu}\ \hat{\sigma}^2)]^{-1}$ is a nonnegative definite matrix. Particularly, $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2$ implies that

$$Var(\hat{\mu}) = \tfrac{\sigma^2}{n},$$

i.e., $\hat{\mu}$ achieves the CRLB, and that

$$Var(\hat{\sigma}^2) = \tfrac{2\sigma^4}{n-1},$$

i.e., $\hat{\sigma}^2$ doesn't achieve the CRLB.

(5) Consistency

For the single parameter case, the estimator $\hat{\theta}$ is *consistent* if and only if $p\lim \hat{\theta} = \theta$; for the vector case, it requires the equality of element by element. If x_n has finite μ and σ^2, then $\bar{x}$ is a consistent estimator for μ. For any function $g(x)$, if $E[g(x)]$ and $Var[g(x)]$ are finite, then $p\lim \frac{1}{n}\Sigma g(x) = E[g(x)]$.

2. Sufficient Statistics

(1) Sufficient Statistics

$X_1, ..., X_n$ is a random sample from $f(x;\theta)$. $Y = u(X_1, ..., X_n)$ is a sufficient statistic for θ if given any other statistic $Z = v(X_1, ..., X_n)$,

the conditional density function doesn't depend upon θ, i.e., $g(y,z;\theta) = h(z|y)\cdot g(y;\theta)$ and $h(z|y)$ doesn't depend upon θ.

(2) Fisher-Neyman Theorem

$X_1,...,X_n$ is a random sample from $f(x;\theta)$. $Y = u(X_1,...,X_n)$ is a sufficient statistic for θ iff $f(x_1,...,x_n;\theta) = g(y;\theta)\cdot h(x_1,...,x_n)$. Note that $h(\cdot)$ doesn't depend upon θ.

(3) Factorization Criterion

$X_1,...,X_n$ is a random sample from $f(x;\theta)$. $Y = u(X_1,...,X_n)$ is a sufficient statistic for θ iff $f(x_1,...,x_n;\theta) = k_1(y;\theta)\cdot k_2(x_1,...,x_n)$, where both $k_2(\cdot)$ and the domain of $k_2(\cdot)$ don't depend upon θ.

Example 1: $X_1,...,X_n$ are i.i.d. Poisson(λ). So the joint density function can be factorized as follows,

$$f(x_1,...,x_n;\lambda) = \frac{e^{-\lambda n}\lambda^{\Sigma x_i}}{\Pi x_i!} = (e^{-\lambda n}\lambda^{\Sigma x_i})\cdot\frac{1}{\Pi x_i!}.$$

Hence, Σx_i is a sufficient statistic for λ.

Example 2: $X_1,...,X_n$ are i.i.d. Bernoulli(p). So the joint density function can be factorized as follows,

$$f(x_1,...,x_n;p) = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} = [p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}]\cdot 1.$$

Hence, Σx_i is a sufficient statistic for p.

(4) Rao-Blackwell Theorem

Let Y be sufficient for θ, and W be any unbiased estimator for θ, consider $E(W|Y=y) = \phi(y)$, then

(a) $\phi(y)$ is a statistic, i.e., there is no θ hidden in $\phi(y)$;
(b) $\phi(y)$ is unbiased for θ;
(c) $var[\phi(y)] < var(W)$;
(d) for many distribution (known as "complete" families), $\phi(y)$ is unique and the minimum variance unbiased estimator (MVUE) for θ.

(5) Transformation of Sufficient Statistic

If Y is sufficient for θ, then any function of Y is also sufficient for θ as long as the function itself doesn't dependent upon θ.

If Y is sufficient for θ, and $W = u(Y)$, which is a one-on-one correspondence, is unbiased for θ, then W is the minimum variance unbiased estimator (MVUE) for θ. Intuitively speaking, an unbiased estimator W that utilizes the minimum information Y necessary to describe θ has to be the "leanest" among all unbiased estimators for θ.

(6) Technique for Finding MVUE using Rao-Blackwell Theorem

(a) prove that Y is sufficient for θ by using the factorization criterion;
(b) find an unbiased estimator W for $g(\theta)$, i.e., $E[W] = g(\theta)$; (Sometimes we find that $E(Y) = a\cdot g(\theta) + b$, then $W = (Y-b)/a$ will be an unbiased estimator for $g(\theta)$.)
(c) calculate the conditional expectation $E(W|Y=y) = \phi(y)$;
(d) by Rao-Blackwell theorem, conclude that $\phi(y)$ is a MVUE for $g(\theta)$.

Example 3: From bi-variate normal distribution, we know that $Y_i|x_i \sim N(\beta x_i, 1)$, where the variance is standardized to 1. The joint density function can be factored as follows,

$$\begin{aligned} f(y_1, ..., y_n\,|x_1, ..., x_n; \beta\,) &= (2\pi)^{-\frac{n}{2}} \exp\{-\tfrac{1}{2}\Sigma_i[y_i - \beta x_i]^2\} \\ &= \left\{\exp[\beta\Sigma_i x_i y_i - \tfrac{1}{2}\beta^2\Sigma_i x_i^2]\right\} \\ &\quad \cdot \left\{(2\pi)^{-\frac{n}{2}} \exp[-\tfrac{1}{2}\Sigma_i y_i^2]\right\}. \end{aligned}$$

Because $E(\Sigma_i x_i y_i) = \Sigma_i[x_i E(y_i)] = \beta\Sigma_i x_i^2$, we know $E[\Sigma_i x_i y_i/\Sigma_i x_i^2] = \beta$, i.e., $\Sigma_i x_i y_i/\Sigma_i x_i^2$ is an unbiased estimator for β. By the factorization criterion, we know

$$\begin{aligned} k_1(y_1, ..., y_n; x_1, ..., x_n; \beta) &= \exp\left\{\beta\Sigma_i x_i y_i - \frac{1}{2}\beta^2\Sigma_i x_i^2\right\} \\ &= \exp\left\{\Sigma_i x_i^2\left[\beta\frac{\Sigma_i x_i y_i}{\Sigma_i x_i^2} - \frac{1}{2}\beta^2\right]\right\} \end{aligned}$$

implies that $\Sigma_i x_i y_i/\Sigma_i x_i^2$ is sufficient for β. Therefore, $\Sigma_i x_i y_i/\Sigma_i x_i^2$ is a minimum variance unbiased estimator for β.

3. Cramer-Rao Lower Bound

Let X be a random variable with p.d.f. $f(x;\theta)$. Define a score vector $\mathbf{g} = \frac{\partial \ln f(x;\theta)}{\partial\theta}$ and $\mathbf{H} = \frac{\partial^2 \ln f(x;\theta)}{\partial\theta\partial\theta'}$. We can show that $E(\mathbf{g}) = 0$ and $var(\mathbf{g}) = -E(\mathbf{H})$ as follows.

Starting from $\int f(x;\theta)dx = 1$ and taking derivatives with respect to θ, we have

$$\begin{aligned} \int \frac{\partial f(x;\theta)}{\partial\theta}dx &= 0 \\ \int \frac{\partial f(x;\theta)/f(x;\theta)}{\partial\theta}f(x;\theta)dx &= 0 \\ \int \frac{\partial \ln f(x;\theta)}{\partial\theta}f(x;\theta)dx &= 0 \\ E(\mathbf{g}) &= 0. \end{aligned}$$

Starting from $\int \frac{\partial \ln f(x;\theta)}{\partial\theta}f(x;\theta)dx = 0$ and taking derivatives with respect to θ again, we have

$$\begin{aligned} \int \left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta\partial\theta'}f(x;\theta) + \frac{\partial \ln f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}\right]dx &= 0 \\ \int \left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta\partial\theta'}f(x;\theta) + \frac{\partial \ln f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)/f(x;\theta)}{\partial\theta}f(x;\theta)\right]dx &= 0 \\ \int \left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta\partial\theta'}f(x;\theta) + \left(\frac{\partial \ln f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)\right]dx &= 0 \\ E(\mathbf{H}) + E(\mathbf{g}^2) &= 0 \\ var(\mathbf{g}) &= -E(\mathbf{H}). \end{aligned}$$

Let $\hat{\theta}(x_1, ..., x_n)$ be any unbiased estimator for θ, then $E[\hat{\theta}(x_1, ..., x_n)] = \theta$ implies

$$\int \cdots \int \hat{\theta}(x_1, ..., x_n) f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n = \theta$$

$$\int \cdots \int \hat{\theta}(x_1, ..., x_n) f(x_1; \theta) \cdots f(x_n; \theta) \sum\nolimits_{i=1}^{n} \frac{\partial \ln f(x_i; \theta)}{\partial \theta} dx_1 \cdots dx_n = 1$$

$$E\left[\hat{\theta}(x_1, ..., x_n) \Sigma_i \mathbf{g}_i\right] = 1$$

$$cov\left[\hat{\theta}(x_1, ..., x_n), \Sigma_i \mathbf{g}_i\right] + E[\hat{\theta}(x_1, ..., x_n)] \cdot E(\Sigma_i \mathbf{g}_i) = 1$$

$$cov\left[\hat{\theta}(x_1, ..., x_n), \Sigma_i \mathbf{g}_i\right] = 1.$$

Using the fact that $\left|\rho_{\hat{\theta}, \Sigma_i \mathbf{g}_i}\right| \leqslant 1$ and $\rho_{\hat{\theta}, \Sigma_i \mathbf{g}_i} \sqrt{var(\hat{\theta})} \sqrt{var(\Sigma_i \mathbf{g}_i)} = 1$, we have $var(\hat{\theta}) \geqslant \frac{1}{var(\Sigma_i \mathbf{g}_i)} = \frac{1}{nvar(\mathbf{g}_i)} = \frac{1}{nE(\mathbf{g}\mathbf{g}')} = -\frac{1}{nE(\mathbf{H})} = \frac{1}{n}[\mathbf{I}(\theta)]^{-1}$, where $\mathbf{I}(\theta) \equiv -E[\mathbf{H}]$ is known as Fisher information matrix.

For univariate case, the CRLB for an unbiased estimator $\hat{\theta}$ is $\frac{1}{n}[\mathbf{I}(\theta)]^{-1}$. For multivariate case, the CRLB for an unbiased estimator $\hat{\theta}$ is $[\mathbf{I}(\theta)]^{-1}$. If an unbiased estimator achieves the CRLB, then it is efficient, i.e., MVUE. The converse need not be true.

4. Maximum Likelihood Estimators

(1) Regularity conditions:
   D1: $\ln L(\mathbf{x}; \theta), \mathbf{g} = \frac{\partial \ln L(\mathbf{x}; \theta)}{\partial \theta}, \mathbf{H} = \frac{\partial^2 \ln L(\mathbf{x}; \theta)}{\partial \theta \partial \theta'}$;
   D2: $E(\mathbf{g}) = 0$;
   D3: $Var(\mathbf{g}) = -E(\mathbf{H})$.

(2) MLE properties:
   M1: consistency: $p \lim \hat{\theta}_{ML} = \theta$;
   M2: asymptotic normality: $\hat{\theta}_{ML} \xrightarrow{a} N\{\theta, [\mathbf{I}(\theta)]^{-1}\}$,
   where $\mathbf{I}(\theta) = -E(\mathbf{H}) = E(\mathbf{g}\mathbf{g}') = Var(\mathbf{g})$;
   M3: asymptotic efficiency: $\hat{\theta}_{ML}$ is asymptotically efficient and achieves the CRLB;
   M4: invariance: the MLE of $\mathbf{c}(\theta)$ is $\mathbf{c}(\theta_{ML})$.

(3) Estimating the asymptotic variance of the MLE:
   Using one of the following three alternatives expressions evaluated at $\hat{\theta}_{ML}$ to get $Est.Asy.Var(\hat{\theta})$.

$$[\mathbf{I}(\hat{\theta})]^{-1} = \left\{ -E\left[\frac{\partial^2 \ln L(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}'}\right] \right\};$$

$$[\hat{\mathbf{I}}(\hat{\theta})]^{-1} = \left\{ -\left[\frac{\partial^2 \ln L(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}'}\right] \right\};$$

$$[\hat{\hat{\mathbf{I}}}(\hat{\theta})]^{-1} = [\hat{\mathbf{G}}'\hat{\mathbf{G}}]^{-1},$$

where

$$\hat{\mathbf{G}}' = (\hat{\mathbf{g}}_1 \ \hat{\mathbf{g}}_2 \ ... \ \hat{\mathbf{g}}_n), \text{and } \hat{\mathbf{g}}_i = \partial \ln L(\hat{\theta}) / \partial \hat{\theta}.$$

5. Concepts on Hypothesis Testing

(1) A test is a specification of a critical region whereas a critical region is the region used to reject H_0.
(2) Power function is the probability of rejecting H_0.
(3) Significance level (size of the critical region, or the α level of the test) is the maximum of the power function given H_0 is true.

	H_0 is true ($\theta = \theta_0$)	H_0 is false ($\theta = \theta_A$)
Reject H_0	Tyep I Error (α)	Correctness ($1-\beta$)
Accept H_0	Correctness ($1-\alpha$)	Type II Error (β)

(4) The power function $K(\theta_0)$ is the probability of rejecting H_0 given H_0 is true ($\theta = \theta_0$), and the power function $K(\theta_A)$ is the probability of rejecting H_0 given H_0 is false ($\theta = \theta_A$). Denote the critical region by C, then $K(\theta_0) = \Pr(x \in C; \theta = \theta_0)$ and $K(\theta_A) = \Pr(x \in C; \theta = \theta_A)$.
(5) For a fixed α, we want to maximize the power $1-\beta$, i.e., to maximize the probability of rejecting H_0 given H_0 is false. Equivalently, we want to minimize type II error, i.e., to minimize the probability of not rejecting H_0 given H_0 is false.
(6) C is a best critical region of size α if
   (a) $\Pr(x \in C; \theta = \theta_0) = \alpha$;
   (b) For any other critical region A of size α, i.e., $\Pr(x \in A; \theta = \theta_0) = \alpha$, we have $\Pr(x \in C; \theta = \theta_A) \geqslant \Pr(x \in A; \theta = \theta_A)$.
   (Intuitively, C corresponds to the highest power $1-\beta$.)
(7) Neyman-Pearson Theorem
   Given $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_A$. If for some $k > 0$
   (a) $\frac{L(\theta=\theta_0;x)}{L(\theta=\theta_A;x)} \leqslant k, \forall x \in C$;
   (b) $\frac{L(\theta=\theta_0;x)}{L(\theta=\theta_A;x)} \geqslant k, \forall x \in \bar{C}$;
   (c) $\Pr(x \in C; \theta = \theta_0) = \alpha$;

then C is the best critical region of size α for testing H_0 vs. H_A.

Intuitively, this theorem says that given the size if the sample likelihood under the null hypothesis is minimal, relative to the alternative, for realizations inside the critical region, and if the sample likelihood under the null is maximal, relative to the alternative, for realizations outside the critical region, then the critical region concerned is the best.

6. Hypothesis Testing Statistics

$H_0 :$ $\mathbf{c}(\theta) = \mathbf{q}$ (J equations)

(1) Likelihood ratio test statistic: (compute both the restricted and the unrestricted model)

$$\lambda = \frac{\bar{L}_R}{\bar{L}_U}.$$

Under the null hypothesis, $-2\ln\lambda \xrightarrow{a} \chi^2(J)$.
(2) Wald test statistic: (compute only the unrestricted model)

$$W = [\mathbf{c}(\hat{\theta}) - \mathbf{q}]'\{Var[\mathbf{c}(\hat{\theta}) - \mathbf{q}]\}^{-1}[\mathbf{c}(\hat{\theta}) - \mathbf{q}].$$

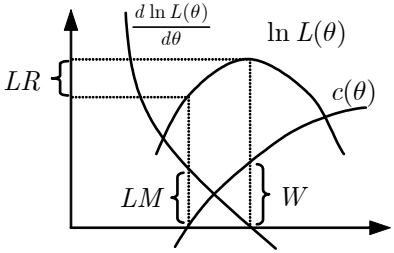
Under the null hypothesis, $W \xrightarrow{a} \chi^2(J)$.

(3) Lagrangian multiplier test statistic: (compute only the restricted model)

$$LM = \left[\frac{\partial \ln(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right]' \{\mathbf{I}(\hat{\theta}_R)\}^{-1} \left[\frac{\partial \ln(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right].$$

Under the null hypothesis, $LM \xrightarrow{a} \chi^2(J)$.

(4) Graphical representation of the three test statistics.



Part 2

Basic Econometrics

CHAPTER 5

Matrix Algebra

1. Algebraic Manipulation of Matrices

(1) For $\mathbf{C}_{(N\times K)} = \mathbf{A}_{(N\times N)}\mathbf{B}_{(N\times K)}$, we want to write them into the format of summation so as to facilitate the computation. But how should we arrange the matrices into vectors?

If we write $\mathbf{C} = (\mathbf{c}_1 \ ... \ \mathbf{c}_K)$ in stack of columns, then we have $\mathbf{c}_k = \mathbf{A}\mathbf{b}_k$. In this process, we arrange matrix $\mathbf{B}$ in the same format of $\mathbf{C}$ so that the two sides of $\mathbf{c}_k = \mathbf{A}\mathbf{b}_k$ are conformable. In writing the right-hand-side into a summation, we can't break $\mathbf{b}_k$, so the only choice left is to break up $\mathbf{A}$ while keeping the format of columns conforming with $\mathbf{c}_k$. So we have to arrange $\mathbf{A}$ in the format of $\mathbf{A} = (\mathbf{a}_1 \ ... \ \mathbf{a}_N)$ and each element of $\mathbf{b}_k$ will enter the summation individually. (Keep in mind the usual regression function in matrix notation $\mathbf{y} = \mathbf{X}\mathbf{b}$.) Therefore, we have $\mathbf{c}_k = \sum_n b_{nk}\mathbf{a}_n$, and we say each column of $\mathbf{C}$ is a linear combination of columns of $\mathbf{A}$.

If we write $\mathbf{C} = (\mathbf{c}^{1\prime}...\mathbf{c}^{N\prime})'$ in stack of rows, then we have $\mathbf{c}^n = \mathbf{a}^n\mathbf{B}$. In this process, we arrange matrix $\mathbf{A}$ in the same format of $\mathbf{C}$ so that the two sides of $\mathbf{c}^n = \mathbf{a}^n\mathbf{B}$ are conformable. In writing the right-hand-side into a summation, we can't break $\mathbf{a}^n$, so the only choice left is to break up $\mathbf{B}$ while keeping the format of rows conforming with $\mathbf{c}^n$. So we have to arrange $\mathbf{B}$ in the format of $\mathbf{B} = (\mathbf{b}^{1\prime}...\mathbf{b}^{N\prime})'$ and each element of $\mathbf{a}^n$ will enter the summation individually. Therefore, we have $\mathbf{c}^n = \sum_k \mathbf{a}_{nk}\mathbf{b}^n$, and we say each row of $\mathbf{C}$ is a linear combination of rows of $\mathbf{B}$.

(2) For $\mathbf{X}_{N\times K}$, we want to write $\mathbf{X}'\mathbf{X}$ into the format of a summation of matrices. How should we arrange $\mathbf{X}_{N\times K}$? First of all, we realize that each element of the summation is a $K \times K$ matrix. If we arrange $\mathbf{X}_{N\times K}$ as $(\mathbf{x}_1...\mathbf{x}_K)$, we are going to end up with 1×1 element for the final matrix. Clearly this is not the result we want. So if we arrange $\mathbf{X}_{N\times K}$ as $(\mathbf{x}^{1\prime}...\mathbf{x}^{N\prime})'$, we will get what we wanted: $\mathbf{X}'\mathbf{X} = \sum_n \mathbf{x}^{n\prime}\mathbf{x}^n$.

(3) A couple of rules:
$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$;
$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$,if all matrices concerned are nonsingular;
$\mathbf{a}'\mathbf{a} = tr(\mathbf{a}'\mathbf{a}) = tr(\mathbf{a}\mathbf{a}')$;
$tr(\mathbf{ABCD}) = tr(\mathbf{BCDA}) = tr(\mathbf{CDAB}) = tr(\mathbf{DABC})$.
(The trace of a square matrix is the sum of its diagonal elements.)

(4) A few rules regarding summations:
$\Sigma x_i = \mathbf{i}'\mathbf{x}$;
$\frac{1}{n}\Sigma x_i = (\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{x}$;
$\Sigma x_i^2 = \mathbf{x}'\mathbf{x}$;

$\Sigma(x_i - \bar{x})^2 = (\mathbf{x} - \mathbf{i}\bar{x})'(\mathbf{x} - \mathbf{i}\bar{x})$;
$\Sigma(\mathbf{x}_i - \bar{x})(\mathbf{y}_i - \bar{y}) = (\mathbf{x} - \mathbf{i}\bar{x})(\mathbf{y} - \mathbf{i}\bar{y})$.

(5) Define $\mathbf{M}^0 = \mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'$, which is an idempotent matrix, and we have the following results.
$\mathbf{M}^0\mathbf{i} = \mathbf{0}$;
$\mathbf{x} - \mathbf{i}\bar{x} = \mathbf{M}^0\mathbf{x}$;
$\Sigma(x_i - \bar{x}) = \mathbf{i}'\mathbf{M}^0\mathbf{x} = 0$;
$\Sigma(x_i - \bar{x})^2 = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{x}) = \mathbf{x}'\mathbf{M}^0\mathbf{x}$;
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{y}) = \mathbf{x}'\mathbf{M}^0\mathbf{y}$.
(If $\mathbf{M}$ is an idempotent matrix, then $\mathbf{MM} = \mathbf{M}$; if $\mathbf{M}$ is idempotent and symmetric, then $\mathbf{MM}' = \mathbf{M}$.)

2. Geometry of Matrices

(1) The K elements of a column vector can be viewed as the *coordinates* of a point in a K-dimensional space. In particular, the two-dimensional *plane*, $\mathbb{R}^2$, is the set of all vectors with two real-valued coordinates. This plane has two important properties. $\mathbb{R}^2$ is *closed under scalar multiplication*; every scalar multiple of a vector in the plane is also in the plane. $\mathbb{R}^2$ is also *closed under addition*; the sum of any two vectors in the plane is always a vector in the plane. Now we define a *vector space* as any set of vectors that is closed under scalar multiplication and addition.

(2) A set of vectors in a vector space is a *basis* for that vector space if any vector in the vector space can be written as a linear combination of them. The basis of a vector space is not unique, since any set of vectors that satisfies the definition will do. But for any particular basis, only one linear combination of them will produce another particular vector in the vector space. Note that exactly K vectors are required to form a basis for $\mathbb{R}^K$.

(3) Although the basis for a vector space is not unique, not every set of K vectors will suffice. That is because it may be the case that some of the vectors are linearly dependent. A set of vectors is *linearly dependent* if any one of the vectors in the set can be written as a linear combination of the others. A set of vectors is *linearly independent* if and only if the only solution to $\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + ... + \alpha_K\mathbf{a}_K = 0$ is $\alpha_1 = \alpha_2 = ... = \alpha_K = 0$. Otherwise, we can always choose one non-zero $\alpha_i \neq 0$ to scale all vectors other than $\mathbf{a}_i$ to reach the vector $\mathbf{a}_i$. Now we know that a basis for a vector space of K dimensions is any set of K linearly independent vectors in that space and that any set of more than K vectors in $\mathbb{R}^K$ must be linearly dependent.

(4) The set of all linear combinations of a set of vectors is the vector space that is *spanned* by those vectors. For example, by definition, the space spanned by a basis for $\mathbb{R}^K$ is $\mathbb{R}^K$. Consider two three-coordinate vectors whose third element is zero. These two vectors don't span the three-dimensional space $\mathbb{R}^3$ in that any linear combinations of these two vectors will have a third coordinate of zero and any vector with nonzero third coordinate is not covered. The plane spanned by these two vectors is called a *subplane*, or two-dimensional *subspace* in $\mathbb{R}^3$. Note that this subplane is not $\mathbb{R}^2$; it is the set of vectors in $\mathbb{R}^3$ whose third coordinate is zero. By the same

logic, any line in $\mathbb{R}^3$ is a one-dimensional subspace, in this case, the set of all vectors in $\mathbb{R}^3$ whose coordinates are multiples of those of the vector that define the line.

The space spanned by a set of vectors in $\mathbb{R}^K$ has at most K dimensions. If this space has fewer than K dimensions, it is a subspace, or *hyperplane*. But the important point is that every set of vectors spans some space; it may be the entire space in which the vectors reside, or it may be some subspace of it.

(5) We view a matrix as a set of column vectors. The *column space* of a matrix is the vector space that is spanned by its column vectors. If the matrix contains K columns, its column space might have K dimensions, but it certainly can have less than K dimensions if not all K columns are linearly independent. The *column rank* of a matrix is the dimension of the vector space that is spanned by its columns. It follows that the column rank of a matrix is equal to the largest number of linearly independent column vectors it contains.

(6) The column rank and row rank of a matrix are equal and the row space and column space of a matrix have the same dimension. If the column rank of a matrix happens to equal the number of columns it contains, the matrix is said to have *full column rank*. Since the row and column ranks of a matrix are always equal, we can speak unambiguously of the *rank of a matrix*. For either the row rank or the column rank, we have

$$rank(\mathbf{A}) = rank(\mathbf{A}') \leq \min(\#rows, \#cols).$$

(7) A matrix is said to have *full rank* if its rank is equal to the number of columns it contains. Of particular interest will be the distinction between *full rank* and *short rank* matrices. The distinction turns of the solutions to $\mathbf{Ax} = \mathbf{0}$. If a nonzero vector $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{0}$ exists, $\mathbf{A}$ does not have full rank.

(8) In a product matrix $\mathbf{C} = \mathbf{AB}$, every column of $\mathbf{C}$ is a linear combination of the columns of $\mathbf{A}$, so each column of $\mathbf{C}$ is in the column space of $\mathbf{A}$. It is possible that the set of columns in $\mathbf{C}$ could span this space, but it is not possible for them to span a higher-dimension space. At best, they could be a full set of linearly independent vectors in $\mathbf{A}$'s column space. We conclude that the column rank of $\mathbf{C}$ could not be greater than that of $\mathbf{A}$. Similarly, we have the conclusion that the row rank of $\mathbf{C}$ could not be greater than that of $\mathbf{B}$. Therefore, we have

$$rank(\mathbf{AB}) \leq \min(rank(\mathbf{A}), rank(\mathbf{B})).$$

In particular, for matrices $\mathbf{A}_{(M\times N)}$ and $\mathbf{B}_{(N\times N)}$, we have

$$rank(\mathbf{AB}) = rank(\mathbf{A}).$$

For any matrix $\mathbf{A}$, we also have

$$rank(\mathbf{A}) = rank(\mathbf{A}') = rank(\mathbf{A}'\mathbf{A}) = rank(\mathbf{AA}').$$

(9) The determinant of a matrix is nonzero if and only if it has full rank. For a diagonal matrix $\mathbf{D}_{(K\times K)}$, its column vectors define a "box" in $\mathbb{R}^K$ whose sides are all at right angles to one another. (Each column vector defines a segment on one of the axes.) Its "volume," or determinant, is simply the product of the lengths of the sides, i.e., the product of the

diagonal elements of the matrix $\mathbf{D}$. Two useful conclusions for general square matrices $\mathbf{D}$, $\mathbf{C}$ and scalar c are:

$$|c\mathbf{D}| = c^K |\mathbf{D}| \text{ and } |\mathbf{DC}| = |\mathbf{D}| \cdot |\mathbf{C}| .$$

(10) Given a column vector $\mathbf{y}$ and matrix $\mathbf{X}$, we are interested in expressing $\mathbf{y}$ as a linear combination of the columns of $\mathbf{X}$. There are two possibilities.

(1) If $\mathbf{y}$ lies in the column space of $\mathbf{X}$, we shall be able to find a vector $\mathbf{b}$ such that $\mathbf{y} = \mathbf{Xb}$.

(2) If $\mathbf{y}$ is not in the column space of $\mathbf{X}$, then there is no $\mathbf{b}$ such that $\mathbf{y} = \mathbf{Xb}$ holds. What we can do instead is to find a $\mathbf{b}$ that produces the smallest $\mathbf{e}$ such that $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ holds. That is, we are trying to find a $\mathbf{b}$ such that the distance between $\mathbf{y}$ and $\mathbf{Xb}$ will be shortest. The *length*, or *norm*, of a vector $\mathbf{e}$ is $\|\mathbf{e}\| = \sqrt{\mathbf{e}'\mathbf{e}}$. It turns out that $\mathbf{e}$ with the shortest length must be perpendicular, or *orthogonal*, to $\mathbf{Xb}$. We can then use the definition of *orthogonal* vectors to find out the vector $\mathbf{b}$. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are *orthogonal*, denoted as $\mathbf{a} \perp \mathbf{b}$, if and only if $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \mathbf{0}$.

Note that the linear combination $\mathbf{Xb}$ is called the *projection* of $\mathbf{y}$ into the column space of $\mathbf{X}$. Suppose that the projection of another vector $\mathbf{y}^*$ shares the same projection with $\mathbf{y}$ in the column space of $\mathbf{X}$, then how can we determine whether $\mathbf{y}$ or $\mathbf{y}^*$ is closer to its projection? We cannot use the length of the residual vector $\mathbf{e}$ or $\mathbf{e}^*$ to determine the closeness, because the length of the residual vector will be affected by the lengths of the original vectors. In this case, we would use the angle between the original vector ($\mathbf{y}$ or $\mathbf{y}^*$) and its projection to determine the closeness. The angle θ between two vectors $\mathbf{a}$ and $\mathbf{b}$ satisfies

$$\cos\theta = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|},$$

which takes care of the length of vectors.

(What will happen if $\mathbf{y}$ is orthogonal to the column space of $\mathbf{X}$? The trivial answer would be that the projection will be zero so that there is no projection, meaning that we shouldn't really regress $\mathbf{y}$ on $\mathbf{X}$.) Add one section on geometric representation of $SST = SSR + SSE$ and F-statistic.

3. Miscellaneous

(1) If $\mathbf{A}$ is positive definite, then for any nonzero vector $\mathbf{v}$, then the quadratic form $\mathbf{v}'\mathbf{A}\mathbf{v}$ is also positive definite; if $\mathbf{A}$ is positive definite, so is $\mathbf{A}^{-1}$; if $\mathbf{A}_{(N\times K)}$ has full column rank and $N > K$, then $\mathbf{A}'\mathbf{A}$ is positive definite and $\mathbf{A}\mathbf{A}'$ is nonnegative definite; if $\mathbf{A}$ is positive definite and $\mathbf{B}$ is a nonsingular matrix, then the quadratic form $\mathbf{B}'\mathbf{A}\mathbf{B}$ is positive definite.

(2) Some important differentiation rules:
$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}} = \mathbf{A}'$;
$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}'} = \mathbf{A}$;
$\frac{\partial \mathbf{x}'\mathbf{Ax}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$;
$\frac{\partial \mathbf{x}'\mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}$;
$\frac{\partial \mathbf{x}'\mathbf{A}}{\partial \mathbf{x}'} = \mathbf{A}'$;
$\frac{\partial \mathbf{x}'\mathbf{Ax}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}'$;
$\frac{\partial \mathbf{Ac(x)}}{\partial \mathbf{x}} = \mathbf{C}'\mathbf{A}'$,where $\mathbf{C} = \frac{\partial \mathbf{c(x)}}{\partial \mathbf{x}'}$.

CHAPTER 6

Classical Regression Model

1. Basic Estimation

(1) Assumptions:
  (a) linearity: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$;
  (b) identification condition: $\mathbf{X}_{(n\times K)}$ has column rank K;
  (c) conditional zero mean: $E(\varepsilon|\mathbf{X}) = \mathbf{0}$;
  (d) homoskedasticity and non-autocorrelation: $E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\mathbf{I}$;
  (e) non-stochastic regressors;
  (f) normality of disturbance: $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$;

Typical violations of assumption (a) are: wrong regressors (such as inclusion of irrelevant explanatory variables or exclusion of relevant variables), non-linearity or random coefficients. Multicollinearity refers to the violation of assumption (b). Once assumption (c) is violated, there will be an bias in the intercept. There may be many forms and shapes of violations of assumption (d), but we only consider two special cases: heteroskedasticity or autocorrelated errors. Assumption (e) implies that it is possible to repeat the sample with the same independent variables, and some problems arise from the violation of this assumption. For example, measurement error in independent variables, autoregression, or using lagged values of dependent variables as independent variables, or simultaneous equation system in which dependent variables are determined by the interaction of multiple equations. The assumption (f) is not mandatory in most cases. But without assumption (f), we often don't know the small sample properties of the estimators.

(2) $\mathbf{y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}\mathbf{b} + \mathbf{e}$
  Normal equations: $\mathbf{X}'\mathbf{e} = \mathbf{0}$ or $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$;
  Coefficients: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(3) Simple regression: $\mathbf{y} = \alpha + \beta\mathbf{x} + \varepsilon = a + b\mathbf{x} + \mathbf{e}$

$$a = \bar{y} - b\bar{x}; \quad b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}.$$

If $\varepsilon \sim N(0, \sigma^2)$, then we have

$$a \sim N\left(\alpha, \frac{\sigma^2\Sigma x_i^2}{n\Sigma(x_i - \bar{x})^2}\right), \text{and } b \sim N\left(\beta, \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}\right).$$

(4) Regression with a constant term:
  $\sum e_i = 0 (\mathbf{X}'\mathbf{e} = \mathbf{0} \Rightarrow \mathbf{x}_1{}'\mathbf{e} = 0 \Rightarrow \mathbf{i}'\mathbf{e} = 0)$
  $\bar{y} = \bar{x}'b (\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \Rightarrow \mathbf{i}'\mathbf{y} = \mathbf{i}'\mathbf{X}\mathbf{b} + \mathbf{i}'\mathbf{e} \Rightarrow \bar{y} = \bar{x}'b)$
  $\bar{\hat{y}} = \bar{y} (\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} \Rightarrow \mathbf{i}'\hat{\mathbf{y}} = \mathbf{i}'\mathbf{y} - \mathbf{i}'\mathbf{e} \Rightarrow \bar{\hat{y}} = \bar{y})$

2. Special Matrices

(1) Residual Maker $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$\mathbf{My} = \mathbf{e} \Rightarrow \mathbf{e}$ is the residual from regressing $\mathbf{y}$ on $\mathbf{X}$, where $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$;

$\mathbf{MX} = \mathbf{0} \Rightarrow \mathbf{0}$ is the residual from regressing $\mathbf{X}$ on $\mathbf{X}$;

$\mathbf{Me} = \mathbf{e} \Rightarrow \mathbf{Me} = \mathbf{M}(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{My} - \mathbf{MXb} = \mathbf{My} = \mathbf{e}$;

$\mathbf{M}\varepsilon = \mathbf{e} \Rightarrow \mathbf{e} = \mathbf{My} = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{MX}\beta + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon$;

$\mathbf{My} = \mathbf{Me} = \mathbf{M}\varepsilon = \mathbf{e}$.

In the special case of simple regression model, we have $\mathbf{M}^0 = \mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'$ and $\mathbf{M}^0\mathbf{y} = \mathbf{e}$,where $\mathbf{e} = \mathbf{y} - \mathbf{i}\bar{y}$; $\mathbf{M}^0\mathbf{i} = \mathbf{0}$; $\mathbf{M}^0\mathbf{e} = \mathbf{e}$ for regression with a constant term since we have $\mathbf{i}'\mathbf{e} = 0$ for regression with a constant term.

(2) Projection Matrix (Fitted Value Maker) $\mathbf{P} = \mathbf{I} - \mathbf{M}$

$\mathbf{Py} = \hat{\mathbf{y}} \Rightarrow \hat{\mathbf{y}}$ is the fitted value from regressing $\mathbf{y}$ on $\mathbf{X}$;

$\mathbf{PX} = \mathbf{X} \Rightarrow \mathbf{X}$ is the fitted value from regressing $\mathbf{X}$ on $\mathbf{X}$; particularly, $\mathbf{PM} = \mathbf{MP} = \mathbf{0}$.

(3) $\mathbf{X}'\mathbf{e} = \mathbf{0}, \mathbf{e}'\mathbf{X} = \mathbf{0}'$;

$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{M}'\mathbf{My} = \mathbf{y}'\mathbf{My} = \mathbf{y}'\mathbf{e} = \mathbf{e}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{Xb} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\mathbf{Xb}$.

(4) Matrix representation of simple regression with a constant term.

$SST = \sum(y_i - \bar{y})^2$ is $SST = \mathbf{y}'\mathbf{M^0}\mathbf{y}$ in matrix notation, with degrees of freedom $n - 1$;

$SSR = \sum(\hat{y}_i - \bar{y})^2$ is $SSR = \hat{\mathbf{y}}'\mathbf{M^0}\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{M^0}\mathbf{y}$ in matrix notation, with degrees of freedom $K - 1$;

$SSE = \sum e_i^2$ is $SSE = \mathbf{e}'\mathbf{e}$ in matrix notation, with degrees of freedom $n - K$.

The degrees of freedom for $\mathbf{e}'\mathbf{e}$ can be showed as $\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}\varepsilon$ and $tr(\mathbf{M}) = tr[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = tr(\mathbf{I}_n) - tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = tr(\mathbf{I}_n) - tr(\mathbf{I}_K) = n - K$.

(5) $\mathbf{y} = \mathbf{Xb} + \mathbf{e} \Rightarrow \mathbf{y}'\mathbf{M^0}\mathbf{y} = \hat{\mathbf{y}}'\mathbf{M^0}\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$

We can get this result by pre-multiplying $\mathbf{y}$ by $\mathbf{M}^0$ and then $\mathbf{y}'$ and use $\mathbf{M}^0\mathbf{e} = \mathbf{e}$ and $\mathbf{X}'\mathbf{e} = \mathbf{0}, \mathbf{e}'\mathbf{X} = \mathbf{0}'$. Note that this is valid only for regression with a constant term since $\mathbf{M^0}\mathbf{e} = \mathbf{e}$ comes from a regression with constant term.

(6)

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{Xb}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = \rho^2_{\mathbf{y},\hat{\mathbf{y}}}; \bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n-1)}.$$

To understand why the adjusted $\bar{R}^2$ is necessary, let's first consider the fact that the inclusion of irrelevant explanatory variables can never reduce R^2. This is the case because the lean model is restricted relative to the fattened model and the restrictions can only make it more difficult to minimize the mean squared error. By throwing everything into the kitchen sink, we can get a superficially large R^2. The adjusted $\bar{R}^2$ addresses this particular problem by accounting for degrees of freedom. If an additional regressor covers very little of the unexplained variation in the dependent variable, then $\bar{R}^2$ falls where as R^2 rises.

(7) $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$ and $E(\mathbf{b}) = \beta$, $Var(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

(8) For $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$, by partitioned matrix rule, we have

$\mathbf{b}_1 = (\mathbf{X}_1{}'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{M}_2\mathbf{y} = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2)$

and $Var(\mathbf{b}_1) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$
$\mathbf{b}_2 = (\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2{}'\mathbf{M}_1\mathbf{y} = (\mathbf{X}_2{}'\mathbf{X}_2)^{-1}\mathbf{X}_2{}'(\mathbf{y} - \mathbf{X}_1\mathbf{b}_1)$
and $Var(\mathbf{b}_2) = \sigma^2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}$

3. Gauss-Markov Theorem

(1) The least square estimator is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$. Let's construct another linear estimator $\mathbf{b}_0 = \mathbf{C}\mathbf{y} = \mathbf{C}\mathbf{X}\beta + \mathbf{C}\varepsilon$. For $\mathbf{b}_0$ to achieve unbiasedness, it is necessary that $\mathbf{C}\mathbf{X} = \mathbf{I}$. Let $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then $\mathbf{D}\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$ and $\mathbf{D}\mathbf{X} = \mathbf{C}\mathbf{X} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{0}$. Then $\mathbf{b}_0 = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}]\mathbf{y}$ and $\mathbf{D}\mathbf{X} = \mathbf{0}$ imply that $Var(\mathbf{b}_0) = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}']$, while we know $Var(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Therefore, $Var(\mathbf{b}_0) \geqslant Var(\mathbf{b})$.

Conclusion: The least square estimator is B.L.U.E., and the minimum variance linear unbiased estimator of $\mathbf{w}'\beta$ is $\mathbf{w}'\mathbf{b}$. Note that this theorem holds regardless of the distribution of the disturbance.

4. Test Statistics

(1) Recall that for a bivariate distribution, we have $E(\mathbf{Y}) = E_{\mathbf{X}}[E(\mathbf{Y}|\mathbf{X})]$, $Cov(\mathbf{X}, \mathbf{Y}) = [\mathbf{Y}, E(\mathbf{Y}|\mathbf{X})], Var(\mathbf{Y}) = Var_{\mathbf{X}}[E(\mathbf{Y}|\mathbf{X})] + E_{\mathbf{X}}[Var(\mathbf{Y}|\mathbf{X})]$. Interpretation again.

(2) Unbiased estimator S^2 to σ^2: $S^2 = \frac{1}{n-K}\mathbf{e}'\mathbf{e} \Rightarrow Est.Var(\mathbf{b}) = S^2(\mathbf{X}'\mathbf{X})^{-1}$.

(3) t statistic:

$\frac{b_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim N(0,1)$ and $\frac{(n-K)S^2}{\sigma^2} \sim \chi^2(n-K)$ imply that

$$\frac{(b_k - \beta_k)/\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}{\sqrt{(n-K)S^2/[\sigma^2(n-K)]}} = \frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t(n-K).$$

(1) Critical region for a two-side test for b_k, given significance level λ:

$$H_0 : b_k = \beta_k; H_1 : b_k \neq \beta_k is \left| \frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \right| > t_{\frac{\lambda}{2}}.$$

(2) Critical region for a one-side test for b_k, given significance level λ:

$$H_0 : b_k = \beta_k; H_1 : b_k > \beta_k is \frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} > t_{\frac{\lambda}{2}};$$

$$H_0 : b_k = \beta_k; H_1 : b_k < \beta_k is \frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} < t_{-\frac{\lambda}{2}};$$

(3) Confidence interval for β_k, given confidence level $(1-\lambda)$ is

$$\Pr(t_{-\frac{\lambda}{2}} \leqslant \frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \leqslant t_{\frac{\lambda}{2}}) = 1 - \lambda.$$

(4) $\frac{(n-K)S^2}{\sigma^2} \sim \chi^2(n-K)$ The confidence interval for σ^2, given confidence level α, is $\Pr(\chi^2_a \leqslant \frac{(n-K)S^2}{\sigma^2} \leqslant \chi^2_b) = \alpha$.

(5) $\frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$ for $H_0 : \beta_2 = \mathbf{0}$ (coefficients other than the constant term) and $\varepsilon \sim N(0, \sigma^2)$. The decision rule is that if the F-stat falls in the critical region, or if the probability of "the F-stat falls in the critical region" is less than the desired significance level, reject the null hypothesis.

5. Asymptotics

(1) Definition of *consistency*: $\hat{\theta}$ is consistent for θ if $p\lim\hat{\theta}=\theta$. According to convergence in mean square, if $\lim_{n\to\infty} E(\hat{\theta})=\theta$ and $\lim_{n\to\infty} Var(\hat{\theta})=\mathbf{0}$, then we have $p\lim\hat{\theta}=\theta$, i.e., $\hat{\theta}$ is consistent for θ. Particularly, $E(\bar{x})=\mu$ and $Var(\bar{x})=\frac{\sigma^2}{n}$ implies that $p\lim\bar{x}=\mu$.

(2) Consistency of $\hat{\beta}_{LS}$ regardless of distribution of disturbances: ($p\lim\hat{\beta}_{LS}=\beta$)

Assume that $\lim_{n\to\infty}\frac{\mathbf{X}'\mathbf{X}}{n}=\mathbf{Q}$, and $\mathbf{Q}$ is a positive definite matrix, then we have $\hat{\beta}_{LS}=\beta+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon=\beta+(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}\frac{\mathbf{X}'\varepsilon}{n}$. Taking the probability limits, we have $p\lim\hat{\beta}_{LS}=\beta+\mathbf{Q}^{-1}p\lim(\frac{\mathbf{X}'\varepsilon}{n})=\beta+\mathbf{Q}^{-1}p\lim\bar{\mathbf{w}}$,where $\bar{\mathbf{w}}=\frac{\mathbf{X}'\varepsilon}{n}$. $E(\bar{\mathbf{w}})=E(\frac{\mathbf{X}'\varepsilon}{n})=\mathbf{0}$, $Var(\bar{\mathbf{w}})=Var(\frac{\mathbf{X}'\varepsilon}{n})=(\frac{1}{n})^2\mathbf{X}'\mathbf{X}\sigma^2=\frac{\sigma^2}{n}\frac{\mathbf{X}'\mathbf{X}}{n}$, and $\lim_{n\to\infty}Var(\bar{\mathbf{w}})=\mathbf{0}\mathbf{Q}=\mathbf{0}$.

By the definition of convergence in mean square, we have $p\lim\bar{\mathbf{w}}=E(\bar{\mathbf{w}})=\mathbf{0}$.

Finally, $p\lim\hat{\beta}_{LS}=\beta+\mathbf{Q}^{-1}p\lim\bar{\mathbf{w}}=\beta$.

Note that we don't need assume normality here. What we need is that the regressors are well behaved such that $\lim_{n\to\infty}\frac{\mathbf{X}'\mathbf{X}}{n}=\mathbf{Q}$ is a positive definite matrix.

(3) Asymptotic normality of $\hat{\beta}_{LS}$ regardless of distribution of disturbances:

$\hat{\beta}_{LS}=\beta+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon=\beta+(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}\frac{\mathbf{X}'\varepsilon}{n}\Rightarrow\sqrt{n}(\hat{\beta}_{LS}-\hat{\beta})=(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}})$.

We know that $\lim_{n\to\infty}(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}=\mathbf{Q}^{-1}$ and need to find the limiting distribution of $\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}$.

$\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}=\sqrt{n}\frac{\mathbf{X}'\varepsilon}{n}=\sqrt{n}\bar{\mathbf{w}}=\sqrt{n}(\bar{\mathbf{w}}-E(\bar{\mathbf{w}}))$, since $E(\bar{\mathbf{w}})=\mathbf{0}$.

$Var(\sqrt{n}\bar{\mathbf{w}})=nVar(\bar{\mathbf{w}})=\sigma^2(\frac{\mathbf{X}'\mathbf{X}}{n})$, $\lim_{n\to\infty}Var(\sqrt{n}\bar{\mathbf{w}})=\sigma^2\mathbf{Q}$.

By the central limit theorem, we have $\sqrt{n}(\bar{\mathbf{w}}-E(\bar{\mathbf{w}}))\xrightarrow{d}N(\mathbf{0},\sigma^2\mathbf{Q})$, i.e., $\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}\xrightarrow{d}N(\mathbf{0},\sigma^2\mathbf{Q})$ and thus $\mathbf{Q}^{-1}\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}\xrightarrow{d}N[\mathbf{0},\mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\mathbf{Q}^{-1\prime}]$, i.e.,

$\sqrt{n}(\hat{\beta}_{LS}-\beta)\xrightarrow{d}N(\mathbf{0},\sigma^2\mathbf{Q}^{-1})\Rightarrow\hat{\beta}_{LS}\xrightarrow{a}N[\beta,\sigma^2\frac{\mathbf{Q}^{-1}}{n}]$.

In practice, it is necessary to estimate $\frac{\mathbf{Q}^{-1}}{n}$ with $\frac{(\mathbf{X}'\mathbf{X})^{-1}}{n^2}$ and σ^2 with $\frac{\mathbf{e}'\mathbf{e}}{n-K}$. Note that we need not to assume normality here. What we need is that the regressors are well behaved such that $\lim_{n\to\infty}(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}=\mathbf{Q}^{-1}$ is a positive definite matrix.

(4) Consistency of S^2:

Since

$$\begin{aligned}S^2&=\tfrac{1}{n-K}\varepsilon'\mathbf{M}\varepsilon\\&=\tfrac{1}{n-K}[\varepsilon'\varepsilon-\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon]\\&=\tfrac{n}{n-K}[\tfrac{\varepsilon'\varepsilon}{n}-\tfrac{\varepsilon'\mathbf{X}}{n}(\tfrac{\mathbf{X}'\mathbf{X}}{n})^{-1}\tfrac{\mathbf{X}'\varepsilon}{n}],\end{aligned}$$

we have $p\lim S^2=1\cdot[p\lim(\frac{\varepsilon'\varepsilon}{n})-\mathbf{0}'\mathbf{Q}^{-1}\mathbf{0}]=p\lim(\frac{\varepsilon'\varepsilon}{n})$. Assume ε_i behaves well in the sense that the mean and variance of ε^2 are finite, we

have $p\lim(\frac{\varepsilon'\varepsilon}{n}) = \sigma^2$. Then $p\lim S^2 = \sigma^2$ and $p\lim S^2(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1} = \sigma^2\mathbf{Q}^{-1}$. Note again that we don't have to assume normality here. If we do, then we can get the same result much easier: $(n-K)\frac{S^2}{\sigma^2} = \chi^2(n-K)$ and thus $E(S^2) = \sigma^2$ and $Var(S^2) = \frac{2\sigma^4}{n-K}$. Since $\lim_{n\to\infty} Var(S^2) = 0$, we get $p\lim S^2 = \sigma^2$, by the definition of convergence in mean square.

6. Delta Method and Inference

(1) $Est.Asy.Var(\hat{\beta}_{LS}) = S^2(\mathbf{X}'\mathbf{X})^{-1}$.

(2) Let $f(\hat{\beta}_{LS})$ be a set of J continuous, linear or nonlinear functions of the least square estimators, and let $\mathbf{C} = \frac{\partial f(\hat{\beta}_{LS})}{\partial \hat{\beta}_{LS}'}$. By the Slutsky theorem, $p\lim f(\hat{\beta}_{LS}) = f(\beta)$and $p\lim \mathbf{C} = \partial f(\beta)/\partial\beta' \equiv \Gamma$, then

$$f(\hat{\beta}_{LS}) \xrightarrow{a} N[f(\beta), \Gamma(\frac{\sigma^2}{n}\mathbf{Q}^{-1})\Gamma'].$$

In practice, $Est.Asy.Var[f(\hat{\beta}_{LS})] = \mathbf{C}[S^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{C}'$. If any of the functions is nonlinear, the property of unbiasedness that holds for $\hat{\beta}_{LS}$ may not carry over to $f(\hat{\beta}_{LS})$, but $f(\hat{\beta}_{LS})$ is consistent for $f(\beta)$.

(3) Since $\hat{\beta}_{LS} \xrightarrow{a} N[\beta, \sigma^2\frac{\mathbf{Q}^{-1}}{n}]$, we can construct the following statistic: $t(n-K) = \frac{\hat{\beta}_k - \beta_k}{[S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}]^{\frac{1}{2}}}$.

We also know that $p\lim S^2(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1} = \sigma^2\mathbf{Q}^{-1}$, then

$$q_k = \frac{\hat{\beta}_k - \beta_k}{[\frac{\sigma^2}{n}\mathbf{Q}^{-1}_{kk}]^{\frac{1}{2}}} \xrightarrow{a} N(0,1).$$

(4)

$$\begin{aligned}\hat{F} &= \frac{(\mathbf{R}\hat{\beta}_{LS} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{LS} - \mathbf{q})/J}{[(n-K)S^2/\sigma^2]/(n-K)} \\ &= \frac{(\mathbf{R}\hat{\beta}_{LS} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{LS} - \mathbf{q})/J}{S^2/\sigma^2}.\end{aligned}$$

Since $p\lim(\frac{S^2}{\sigma^2}) = 1$, we only consider the numerator. Hence we have

$$\begin{aligned}J\hat{F} &= (\tfrac{\varepsilon}{\sigma})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\tfrac{\varepsilon}{\sigma}) \\ &= (\tfrac{\varepsilon}{\sigma})'L(\tfrac{\varepsilon}{\sigma}),\end{aligned}$$

where $rank(L) = J$. Therefore, we have $J\hat{F} \xrightarrow{a} \chi^2(J)$. Particularly, when $J = 1$, $F(1, n-K) \xrightarrow{a} \chi^2(1)$.

(5) Another version of the same property is stated as the limiting distribution of the Wald test.

If $\sqrt{n}(\hat{\beta}_{LS} - \beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$ and $H_0 : \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$ is true, then

$$\begin{aligned}W &= (\mathbf{R}\hat{\beta}_{LS} - \mathbf{q})'[S^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{LS} - \mathbf{q}) \\ &= J\hat{F} \xrightarrow{a} \chi^2(J).\end{aligned}$$

(6) Differentiate $\int_x L(x;\theta)dx = 1$ with respect to θ and use $\frac{\partial L}{\partial\theta} = L\cdot\frac{\partial \ln L}{\partial\theta}$, we get $E(\frac{\partial \ln L}{\partial\theta}) = \int_x \frac{\partial \ln L}{\partial\theta}Ldx = \mathbf{0}$. Differentiate this identity further with respect to θ, we get $Var(\frac{\partial \ln L}{\partial\theta}) = E(\frac{\partial \ln L}{\partial\theta})^2 = -E(\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}) \equiv \mathbf{I}(\theta)$.

(7) Use the Neyman-Pearson theorem to find the critical region for the test $H_0 : \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$.

$$\frac{(\mathbf{R}\hat{\beta}_{ML} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta}_{ML} - \mathbf{q})/J}{(\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})/(n-K)} \geqslant C^*.$$

Note that when we assume normality, $\hat{\beta}_{LS} = \hat{\beta}_{ML}$ implies that it is the conventional F-test.

7. OLS vs. MLE

(1) For $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, we have the following set of results.
$\hat{\beta}_{ML} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, E(\hat{\beta}_{ML}) = \beta, Var(\hat{\beta}_{ML}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;
$\hat{\beta}_{LS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, E(\hat{\beta}_{LS}) = \beta, Var(\hat{\beta}_{LS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$;
$\hat{\sigma}^2_{ML} = \frac{\mathbf{e}'\mathbf{e}}{n}, E(\hat{\sigma}^2_{ML}) = \frac{n-K}{n}\sigma^2; Var(\hat{\sigma}^2_{ML}) = \frac{2(n-K)\sigma^4}{n^2}$;
$\hat{\sigma}^2_{LS} = \frac{\mathbf{e}'\mathbf{e}}{n-K}, E(\hat{\sigma}^2_{LS}) = \sigma^2; Var(\hat{\sigma}^2_{LS}) = \frac{2\sigma^4}{n-K}$;
$[\mathbf{I}(\theta)]^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$.

Conclusion: both $\hat{\beta}_{ML}$ and $\hat{\beta}_{LS}$ are unbiased and efficient. $\hat{\sigma}^2_{ML}$ is biased and $\hat{\sigma}^2_{LS}$ is unbiased, but $\hat{\sigma}^2_{ML}$ is more efficient than $\hat{\sigma}^2_{LS}$ although none of them achieves the CRLB.

(2) Statistical Properties of Estimators
   (a) $\hat{\beta}_{LS}$ and $\hat{\sigma}^2_{LS}$ have all the nice properties of maximum likelihood estimators (MLE) under the assumption of normality.
   (b) The desirable properties of MLE:
   M1: consistency: $p\lim \hat{\theta}_{ML} = \theta$;
   M2: asymptotic normality: $\hat{\theta}_{ML} \xrightarrow{a} N(\theta, [\mathbf{I}(\theta)]^{-1})$, where $\mathbf{I}(\theta) = -E(\mathbf{H}) = E(\mathbf{g}\mathbf{g}') = Var(\mathbf{g})$;
   M3: asymptotic efficiency: $\hat{\theta}_{ML}$ is asymptotically efficient and achieves the CRLB;
   M4: invariance: the MLE of $c(\theta)$ is $c(\hat{\theta}_{ML})$.
   (c) Since $\hat{\beta}_{LS} = \hat{\beta}_{ML}$, we know that $\hat{\beta}_{LS}$ has all the MLE properties.
   (d) Does $\hat{\sigma}^2_{LS} = S^2$ have the same MLE properties?
   $E(S^2) = \sigma^2, E(\hat{\sigma}^2_{ML}) = \frac{n-K}{n}\sigma^2 < \sigma^2$.By the second property of MLE, we have $\sqrt{n}(\hat{\sigma}^2_{ML} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$.
   Then we have

$$\begin{aligned} z_n &= (1 - \tfrac{K}{n})\sqrt{n}(\hat{\sigma}^2_{ML} - \sigma^2) + \tfrac{K}{\sqrt{n}}\sigma^2 \\ &= (1 - \tfrac{K}{n})N(0, 2\sigma^4) + \tfrac{K}{\sqrt{n}}\sigma^2. \end{aligned}$$

As $n \to \infty, \frac{K}{N} \to 0, \frac{K}{\sqrt{n}} \to 0$, we have $z_n \xrightarrow{d} N(0, 2\sigma^4)$. In the meanwhile, we know $\hat{\sigma}^2_{ML} = \frac{n-K}{n}S^2$, then

$$\begin{aligned} z_n &= \tfrac{n-K}{n} \cdot \sqrt{n} \cdot (\tfrac{n-K}{n}^2 - \sigma^2) + \tfrac{K}{\sqrt{n}}\sigma^2 \\ &= \sqrt{n}(S^2 - \sigma^2). \end{aligned}$$

Hence $\sqrt{n}(S^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$, i.e., $\hat{\sigma}^2_{LS} = S^2$ has the same properties as $\hat{\sigma}^2_{ML}$.

(3) Let $\hat{\theta}_{ML}$ be the MLE for θ, and let $\mathbf{c}(\theta) == [c_1(\theta) \quad c_2(\theta) \quad ... \quad c_K(\theta)]'$. By the invariance property of MLE, we know $\mathbf{c}(\hat{\theta})$ is the MLE for $\mathbf{c}(\theta)$. How can we find the $Asy.Var[\mathbf{c}(\hat{\theta})]$ then? Refer back to G.4.11.6., we have $Asy.[nVar(\mathbf{z}_n)] = \Sigma$ and $Asy.\{nVar[\mathbf{c}(\mathbf{z}_n)]\} = \mathbf{C}\Sigma\mathbf{C}'$, where $\mathbf{C} = \frac{\partial \mathbf{c}(\mathbf{z}_n)}{\partial \mathbf{z}_n'}$, so $Asy.\{nVar[\mathbf{c}(\mathbf{z}_n)]\} = \mathbf{C}\{Asy.[nVar(\mathbf{z}_n)]\}\mathbf{C}'$. Here we have $Asy.\{nVar[\mathbf{c}(\hat{\theta})]\} = \mathbf{C}\{Asy.[nVar(\hat{\theta})]\}\mathbf{C}'$, where $\mathbf{C} = \frac{\partial \mathbf{c}(\theta)}{\partial \theta'}$. Therefore, we have

$$\begin{aligned} Asy.\{Var[\mathbf{c}(\hat{\theta})]\} &= \tfrac{1}{n} \cdot \mathbf{C}\left(\lim_{n\to\infty}\{n[\mathbf{I}(\theta)]^{-1}\}\right)\mathbf{C}' \\ &= \tfrac{1}{n} \cdot \mathbf{C}\left(\lim_{n\to\infty}[\mathbf{I}(\theta)/n]^{-1}\right)\mathbf{C}'. \end{aligned}$$

(4) Wald test: $H_0 : f(\beta) = 0; H_1 : f(\beta) \neq 0$

$$\begin{aligned} W &= f(\hat{\beta})'\{G(\hat{\beta})[S^2(\mathbf{X}'\mathbf{X})^{-1}]G(\hat{\beta})'\}^{-1}f(\hat{\beta}) \\ &\sim \chi^2(J), \end{aligned}$$

where J is the number of restrictions and $G(\hat{\beta}) = \frac{\partial f(\hat{\beta})}{\partial \hat{\beta}'}$.

8. Partitioned Regression

If we partition the explanatory variables into two subsets, $\mathbf{X}_1$ and $\mathbf{X}_2$, we know that three types of variations in explanatory variables are competing against each other in explaining the dependent variables, namely, the variations of $\mathbf{X}_1$ alone, the variation of $\mathbf{X}_2$ alone, and the covariance between $\mathbf{X}_1$ and $\mathbf{X}_2$ if they are not orthogonal. To get the coefficient vectors for $\mathbf{X}_2$ in the full model, we can take the following steps that are equivalent to running the full model.

First, remove the variation caused by $\mathbf{X}_1$ alone by regressing the dependent variable on $\mathbf{X}_1$. Second, remove the covariance between $\mathbf{X}_1$ and $\mathbf{X}_2$ by regressing $\mathbf{X}_2$ on $\mathbf{X}_1$. Third, tease out the contribution by $\mathbf{X}_2$ alone by regressing the residuals from the first step on the residuals from the second step.

(1) For a classical regression $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, we have $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and the residual is $\mathbf{e} = \mathbf{M}\mathbf{y}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

If we let $\mathbf{X} = \mathbf{x}_1$ and $\mathbf{b} = b_1$, we have $\mathbf{y} = \mathbf{x}_1 b_1 + \mathbf{e}$, $b_1 = (\mathbf{x}_1{}'\mathbf{x}_1)^{-1}\mathbf{x}_1{}'\mathbf{y}$ and the residual is $\mathbf{e} = \mathbf{M}_1\mathbf{y}$, where $\mathbf{M}_1 = \mathbf{I} - \mathbf{x}_1(\mathbf{x}_1{}'\mathbf{x}_1)^{-1}\mathbf{x}_1{}'$.

If we let $X = (\mathbf{x}_1 \ \mathbf{x}_2)$ and $\mathbf{b} = (b_1^* \ b_2^*)$, we have $\mathbf{y} = b_1^*\mathbf{x}_1 + b_2^*\mathbf{x}_2 + \mathbf{e}$, $b_1^* = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'(\mathbf{y} - \mathbf{x}_2 b_2^*)$, $b_2^* = (\mathbf{x}_2'\mathbf{M}_1\mathbf{x}_2)^{-1}(\mathbf{x}_2{}'\mathbf{M}_1\mathbf{y})$, where $\mathbf{e} = \mathbf{M}^*\mathbf{y}$, and $\mathbf{M}^* = \mathbf{I} - (\mathbf{x}_2{}'\mathbf{M}_1)'(\mathbf{x}_2{}'\mathbf{M}_1\mathbf{x}_2)^{-1}(\mathbf{x}_2{}'\mathbf{M}_1)$.

Under the special case where $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal (so that $\mathbf{M}_1\mathbf{x}_2 = \mathbf{x}_2$), then the coefficients associated with $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same as the coefficients obtained from regressing $\mathbf{y}$ on $\mathbf{x}_1$ alone and $\mathbf{y}$ on $\mathbf{x}_2$ alone, respectively.

(2) For a classical regression $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, we have $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and the residual is $\mathbf{e} = \mathbf{M}\mathbf{y}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Suppose that $\mathbf{X} = \mathbf{X}_1$ and $\mathbf{b} = \mathbf{b}_1$, we have $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{e}$, $\mathbf{b}_1 = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{y}$ where $\mathbf{e} = \mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1$.

Suppose $X = (\mathbf{X}_1 \ \mathbf{X}_2)$ and $\mathbf{b} = (\mathbf{b}_1^* \ \mathbf{b}_2^*)$, we have $\mathbf{y} = \mathbf{b}_1^*\mathbf{X}_1 + \mathbf{b}_2^*\mathbf{X}_2 + \mathbf{e}$, $\mathbf{b}_1^* = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2^*)$, $\mathbf{b}_2^* = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2{}'\mathbf{M}_1\mathbf{y})$, where $\mathbf{e} = \mathbf{M}^*\mathbf{y}$ and $\mathbf{M}^* = \mathbf{I} - (\mathbf{X}_2{}'\mathbf{M}_1)'(\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2{}'\mathbf{M}_1)$.

Under the special case where $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal (so that $\mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$), then the coefficients associated with $\mathbf{X}_1$ and $\mathbf{X}_2$ are the same as the coefficients obtained from regressing $\mathbf{y}$ on $\mathbf{X}_1$ alone and $\mathbf{y}$ on $\mathbf{X}_2$ alone, respectively.

(3) If we regress $\mathbf{y}$ on $\mathbf{X}_1$, $\mathbf{y} = \mathbf{X}_1\mathbf{c}_1 + \mathbf{u}$, then we have $\mathbf{c}_1 = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{y}$, and the residual is $\mathbf{u} = \mathbf{M}_1\mathbf{y}$, where $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1$. Note that in $\mathbf{u}$ we don't have the variation attributed by $\mathbf{X}_1$ alone.

If we next regress each column of $\mathbf{X}_2$ on $\mathbf{X}_1$, $\mathbf{x}_{2,k} = \mathbf{X}_1\mathbf{c}_{2,k} + \mathbf{v}_k$, then we have $\mathbf{C}_2 = (\mathbf{c}_1 ... \mathbf{c}_K)$, where $c_k = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{x}_{2,k}$ and the residual matrix is $\mathbf{V} = \mathbf{M}_1\mathbf{X}_2$, where $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1$. Note that in $\mathbf{V}$ we have removed the common variation between $\mathbf{X}_1$ and $\mathbf{X}_2$.

We then regress $\mathbf{u}$ on $\mathbf{v}$, $\mathbf{u} = \mathbf{V}\mathbf{c}_3 + \mathbf{w}$, then we have $\mathbf{c}_3 = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{u} = (\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2{}'\mathbf{M}_1\mathbf{y})$ and the residual is $\mathbf{w} = \mathbf{M}_\mathbf{V}\mathbf{u}$, where $\mathbf{M}_\mathbf{V} = \mathbf{I} - \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}' = \mathbf{M}^*$. It's clear now $\mathbf{c}_3$ captures the contribution from the variation of $\mathbf{X}_2$ alone.

If we regress $\mathbf{y}$ on both $\mathbf{X}_1$ and $\mathbf{X}_2$, we should have $\mathbf{y} = \mathbf{X}_1(\mathbf{c}_1 - \mathbf{C}_2\mathbf{c}_3) + \mathbf{X}_2 \cdot \mathbf{c}_3 + \mathbf{w}$. It is easy to realize that $\mathbf{c}_3 = \mathbf{b}_2^*$, and we have the *Frisch-Waugh Theorem* as follows: The subvector $\mathbf{b}_2^*$ in regression $\mathbf{y} = \mathbf{b}_1^*\mathbf{X}_1 + \mathbf{b}_2^*\mathbf{X}_2 + \mathbf{e}$ is the set of coefficients obtained when the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone are regressed on the set of residuals obtained when each column of $\mathbf{X}_2$ is regressed on $\mathbf{X}_1$.

(4) If we set $\mathbf{X}_2 = \mathbf{z}$ in part (3), we have the following result: The coefficient c on $\mathbf{z}$ in a multiple regression of $\mathbf{y}$ on $\mathbf{W} = (\mathbf{X}\ \mathbf{z})$ is computed as $c = (\mathbf{z}'\mathbf{M}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M}\mathbf{y}) = (\mathbf{z}_*{}'\mathbf{z}_*)^{-1}(\mathbf{z}_*{}'\mathbf{y}_*)$, where z_* and $\mathbf{y}_*$ are the residual vectors from least squares regressions of $\mathbf{z}$ and $\mathbf{y}$ on $\mathbf{X}$; $\mathbf{z}_* = \mathbf{M}_X\mathbf{z}$ and $\mathbf{y}_* = \mathbf{M}_X\mathbf{y}$.

(5) If we set $\mathbf{X}_1 = \mathbf{i}$ in part (3), we have $\mathbf{M}_1 = \mathbf{M}^0 = \mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'$ and thus $\mathbf{u} = \mathbf{M}_0\mathbf{y} = \mathbf{y} - \mathbf{i}\bar{y}$ and $\mathbf{v} = \mathbf{M}_0\mathbf{X}_2 = \mathbf{X}_2 - \mathbf{i}\bar{\mathbf{X}}'_2$. Therefore, $\mathbf{c}_3$ is equivalent to the coefficients obtained from the regression $\mathbf{y} - \mathbf{i}\bar{y} = (\mathbf{X}_2 - \mathbf{i}\bar{X}_2{}')\mathbf{c}_3 + \xi$. This result can be stated as follows: The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means, then regressing the variable $\mathbf{y}$ in deviation form on the explanatory variables, also in deviation form.

CHAPTER 7

Inference and Prediction

1. Single Restriction

Omitting a variable or equivalently adding an additional variable.

(1) Comparison on Variance:
(R): $\mathbf{y} = \mathbf{X}\mathbf{b}_* + \mathbf{e}_*$
$\mathbf{b}_* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $Var(\mathbf{b}_*) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
(U): $\mathbf{y} = \mathbf{X}\mathbf{d} + \mathbf{z}c + \mathbf{e}$
$(\mathbf{d}\ c)' = [(\mathbf{X}\ \mathbf{z})'\ (\mathbf{X}\ \mathbf{z})]^{-1}(\mathbf{X}\ \mathbf{z})'\mathbf{y}$, $Var(\mathbf{d}\ c)' = \sigma^2[(\mathbf{X}\ \mathbf{z})'\ (\mathbf{X}\ \mathbf{z})]^{-1}$, and $Var(\mathbf{d}) = Var(\mathbf{b}_*) + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\mathbf{z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}/(\mathbf{z}'\mathbf{M}\mathbf{z})$. Since $Var(\mathbf{b}_*) \leqslant Var(\mathbf{d})$, our conclusion is : restrictions reduce variance.

(2) Comparison on R^2:
(R): $\mathbf{y} = \mathbf{X}\mathbf{b}_* + \mathbf{e}_*$
(U): $\mathbf{y} = \mathbf{X}\mathbf{d} + \mathbf{z}c + \mathbf{e}$
We have $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{z}c$ and $\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{z}c) = \mathbf{b}_* - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c$. Hence $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}_* - [\mathbf{z}c - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\mathbf{c}] = \mathbf{e}_* - \mathbf{M}\mathbf{z}c$ and thus $\mathbf{e}'\mathbf{e} = \mathbf{e}_*{}'\mathbf{e}_* - c^2\mathbf{z}'\mathbf{M}\mathbf{z}$. Since $\mathbf{e}_*{}'\mathbf{e}_* \geqslant \mathbf{e}'\mathbf{e} \Rightarrow R_*^2 \leqslant R^2$, our conclusion is: restrictions reduce R^2.

(3) Test on the null hypothesis that the restrictions hold.

$$t_z^2 = \frac{(R^2 - R_*^2)/1}{(1 - R^2)/(n - K)}, i.e., F(1, n - K) = \frac{R^2 - R_*^2}{(1 - R^2)/(n - K)}.$$

2. F-test on a Set of Restrictions

People often use t-test, F-test, and Chi-square test for making statistical inferences. Note, however, these tests are valid for small samples only if the disturbance terms are normally distributed. In the case of small samples with non-normal errors, we have to rely on bootstrap or Monte Carlo techniques to obtain relevant p-values.

Here is the intuition behind the F-test. Upon imposing a set of restrictions, the minimization process becomes harder to implement and results in a larger sum squared errors. The numerator of the F-statistic concerns about the "per-restriction" increase of sum squared errors, and the denominator implies the "per-error" contribution to sum squared errors. If the set of restrictions is not far away from the truth, then the "standardized friction" shouldn't be large.

Why do we care about degrees of freedom? If we were to explore a possible linear relationship between shoe size and grade average point using only two observations, we would obtain a bogus 100% fit as two points are needed to determine a line. Adding another observation would reduce the fit but it remains large. To correct for this type of bogus fit, we use only the number of "free" observations to compute statistics.

Null Hypothesis: $\mathbf{R}\beta = \mathbf{q}$

Let $\mathbf{m} = \mathbf{Rb} - \mathbf{q}$, then $E(\mathbf{m}) = \mathbf{R}E(\mathbf{b}) - \mathbf{q} = \mathbf{R}\beta - \mathbf{q} = \mathbf{0}$ and $Var(\mathbf{m}) = Var(\mathbf{Rb}) = \mathbf{R}Var(\mathbf{b})\mathbf{R}' = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$. We have $\mathbf{m} \sim N[\mathbf{0}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$ and $\mathbf{m}'[Var(\mathbf{m})]^{-1}\mathbf{m} \sim \chi^2(J)$, i.e., $(\mathbf{Rb}-\mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}) \sim \chi^2(J)$.

Since $\frac{(n-K)S^2}{\sigma^2} \sim \chi^2(n-K)$, we have the following F-stat to test H_0: $\mathbf{R}\beta = \mathbf{q}$.

$$\begin{aligned} F(J, n-K) &= \frac{(\mathbf{Rb}-\mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{[(n-K)S^2/\sigma^2]/(n-K)} \\ &= \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-K)} \\ &= \frac{(\mathbf{Rb}-\mathbf{q})'[S^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{J}. \end{aligned}$$

F-test is valid for any sample size (finite or large sample) so long as disturbances are normally distributed.

3. A Set of Restrictions

Null Hypothesis: $\mathbf{R}\beta = \mathbf{q}$

(1) Comparison on variance:
(U): $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ (without restrictions)
$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $Var(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
(R): $\mathbf{y} = \mathbf{Xb}_* + \mathbf{e}_*$ (with restrictions: $\mathbf{R}\beta = \mathbf{q}$)
$\mathbf{b}_* = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})$ and
$Var(\mathbf{b}_*) = Var(\mathbf{b}) - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$. Since $Var(\mathbf{b}_*) \leqslant Var(\mathbf{b})$, our conclusion is that restrictions reduce variance.

(2) Comparison on R^2:
(U): $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$
(R): $\mathbf{y} = \mathbf{Xb}_* + \mathbf{e}_*$
We have $\mathbf{e}_* = \mathbf{y} - \mathbf{Xb}_* = (\mathbf{y} - \mathbf{Xb}) - (\mathbf{Xb}_* - \mathbf{Xb}) = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})$, and thus $\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b})$. Since $\mathbf{e}_*'\mathbf{e}_* \geqslant \mathbf{e}'\mathbf{e} \Rightarrow R_*^2 \leqslant R^2$, our conclusion is that restrictions reduce R^2.

(3) Test on the null hypothesis: $\mathbf{R}\beta = \mathbf{q}$

$$\begin{aligned} \mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} &= (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \\ &= (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \\ &= \sigma^2\mathbf{m}'[Var(\mathbf{m})]^{-1}\mathbf{m}. \end{aligned}$$

Since $(\mathbf{Rb}-\mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}) \sim \chi^2(J)$ and $(n-K)S^2/\sigma^2 \sim \chi^2(J)$, we have

$$F(J, n-K) = \frac{(e_*'e_* - e'e)/J}{e'e/(n-K)} = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)}.$$

Consider the independence of the numerator and the denominator:

$$\mathbf{Rb} - \mathbf{q} = \mathbf{R}[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon] - \mathbf{q} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \equiv \mathbf{T}\varepsilon \text{ and } \mathbf{e} = \mathbf{M}\varepsilon.$$

It suffices to prove $\mathbf{TM} = \mathbf{0}$, which follows from $\mathbf{TM} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathbf{0}$.

(4) Particularly, we have $\frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F(K-1, n-K)$ for $H_0 : \beta_2 = 0$ (coefficients other than the constant term). It is obvious that $J = K-1$. Since the restricted model will be $\mathbf{y} = \mathbf{i}\bar{y} + \mathbf{e}$, which has no explanatory power, $R_*^2 = 0$, we get the conventional F test.

4. Test a Subset of Coefficients

For $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$, test $H_0 : \beta_2 = \mathbf{0}$.

(1) (R): $\mathbf{y} = \mathbf{X}_1\mathbf{b}_{1*} + \mathbf{e}_*$ vs. (U): $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$
$\mathbf{b}_2 = (\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2{}'\mathbf{M}_1\mathbf{y}$ and $Var(\mathbf{b}_2) = \sigma^2(\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1}$.

(2) To test the null hypothesis, set up the hypothesis design matrix $\mathbf{R}$ as: $\mathbf{R} = (\mathbf{0}\ \mathbf{I})$ and $\mathbf{q} = \mathbf{0}$. Hence $\mathbf{m} = \mathbf{Rb} - \mathbf{q} = (\mathbf{0}\ \mathbf{I})(\mathbf{b}_1{}'\ \mathbf{b}_2{}')' - \mathbf{q} = \mathbf{b}_2$, which implies $Var(\mathbf{m}) = Var(\mathbf{b}_2)$, i.e.,

$$\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' = \sigma^2(\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2)^{-1} \Leftrightarrow [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} = \mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2.$$

(3)

$$\begin{aligned}\mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e} &= \sigma^2\mathbf{m}'[Var(\mathbf{m})]^{-1}\mathbf{m} \\ &= \sigma^2\mathbf{b}_2{}'[\sigma^2]^{-1}\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 \\ &= \mathbf{b}_2{}'\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2.\end{aligned}$$

Particularly, if we let $\mathbf{b}_2 = c$ and $\mathbf{X}_2 = \mathbf{z}$, we have $\mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = c^2\mathbf{z}'\mathbf{M}\mathbf{z}$. Without surprise, this is the same result we get, after painful work, for 1.(2).

(4)

$$F(J, n-K) = \frac{(\mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)} = \frac{(\mathbf{b}_2{}'\mathbf{X}_2{}'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2)/J}{\mathbf{e}'\mathbf{e}/(n-K)}.$$

The independence of the numerator and the denominator follows the generalized case $\mathbf{TM} = \mathbf{0}$, which we proved in 3.(3).

5. A List of Important Facts

Let subscript $_*$ denote results from the restricted model. Let $\mathbf{T} \equiv \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{m} = \mathbf{Rb} - \mathbf{q}$ and $\mathbf{N} \equiv \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$.

(1)

$$\mathbf{b}_* = \mathbf{b} - \mathbf{N}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}.$$

(2)

$$\begin{aligned}&Var(\mathbf{b}_*) = Var(\mathbf{b}) - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \text{ or} \\ &Var(\mathbf{b}_*) = Var(\mathbf{b}) - \sigma^2\mathbf{N}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{N}.\end{aligned}$$

(3)

$$\begin{aligned}\mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e} &= (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \\ &= (\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q}) \text{ or} \\ \mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e} &= \sigma^2\mathbf{m}'[Var(\mathbf{m})]^{-1}\mathbf{m}.\end{aligned}$$

(4)

$$\begin{aligned}
F(J, n-K) &= \frac{(\mathbf{Rb}-\mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{[(n-K)S^2/\sigma^2]/(n-K)} \\
&= \frac{\mathbf{m}'[Var(\mathbf{m})]^{-1}\mathbf{m}/J}{\mathbf{e}'\mathbf{e}/[\sigma^2(n-K)]} \\
&= \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-K)} \\
&= \frac{(\mathbf{e}_*{}'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)} \\
&= \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)} \\
&= \frac{(\mathbf{Rb}-\mathbf{q})'[S^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{J} \\
&= \frac{\mathbf{m}'[S^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}}{J}.
\end{aligned}$$

Part 3

Advanced Econometrics

CHAPTER 8

Functional Form, Nonlinearity, and Specification

1. Omission of Relevant Variables

(1) (PRF): $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$ v.s. (SRF): $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1^* + \mathbf{e}^*$

We have

$$\begin{aligned}\mathbf{b}_1^* &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon) \\ &= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)\mathbf{X}_1'\varepsilon\end{aligned}$$

(2) $E(\mathbf{b}_1^*) = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$

Define $\mathbf{P}_{1,2} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ from $\mathbf{X}_2 = \mathbf{X}_1\mathbf{P}_{1,2} + \mathbf{w}$. Unless $\mathbf{P}_{1.2} = \mathbf{0}$, i.e., $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal, $\mathbf{b}_1^*$ is a biased estimator for β_1.

(3) $Var(\mathbf{b}_1^*) = Var[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$.

If we use the correct SRF $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$, we would have $Var(\mathbf{b}_1) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$. It is easy to see $Var(\mathbf{b}_1^*) \leqslant Var(\mathbf{b}_1)$, and our conclusion is that although $\mathbf{b}_1^*$ is biased, it is more precise than $\mathbf{b}_1$ which results from using the correct SRF.

(4)

$$\begin{aligned}\mathbf{e}_1^{*\prime}\mathbf{e}_1^* &= \mathbf{y}'\mathbf{M}_1\mathbf{y} \\ &= (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon)'\mathbf{M}_1(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon) \\ &= (\mathbf{X}_2\beta_2 + \varepsilon)'\mathbf{M}_1(\mathbf{X}_2\beta_2 + \varepsilon) \\ &= \beta_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\beta_2 + \varepsilon'\mathbf{M}_1\varepsilon + 2\beta_2'\mathbf{X}_2'\mathbf{M}_1\varepsilon\end{aligned}$$

Hence $E(\mathbf{e}_1^{*\prime}\mathbf{e}_1^*) = \beta_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\beta_2 + (n-K_1)\sigma^2$. Hence our conclusion is : $(n-K_1)\sigma^2$ is a biased estimator for $\mathbf{e}_1^{*\prime}\mathbf{e}_1^*$, and we cannot find a proper S^2 to estimate σ^2.

(5) As we know, restrictions reduce R^2 because the minimization process becomes harder, i.e., $R_*^2 \leqslant R^2$.

Conclusions: If we omit relevant variables from the regression, our estimators of β_1 and σ^2 are biased. It is possible for $\mathbf{b}_1^*$ to be more precise than $\mathbf{b}_1$ which results from using the correct SRF, but this should be of limited comfort since we cannot estimate σ^2 consistently, and we cannot test hypothesis about β_1. Moreover, the goodness of fit is reduced.

2. Inclusion of Irrelevant Variables

(1) (PRF): $\mathbf{y} = \mathbf{X}_1\beta_1 + \varepsilon$ v.s. (SRF): $\mathbf{y} = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$

We have

$$\begin{aligned}\mathbf{b}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}\\ &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2(\mathbf{X}_1\beta_1+\varepsilon)\\ &= \beta_1 + (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\varepsilon\end{aligned}$$

(2) $E(\mathbf{b}_1)=\beta_1$ and our conclusion is: $\mathbf{b}_1$ is an unbiased estimator for β_1.
(3) $Var(\mathbf{b}_1)=Var[(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\varepsilon]=\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$.
If we use the correct SRF $\mathbf{y}=\mathbf{X}_1\mathbf{b}_1^*+\mathbf{e}^*$, we would have $Var(\mathbf{b}_1^*)=\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$. It is easy to see $Var(\mathbf{b}_1^*)\leqslant Var(\mathbf{b}_1)$, and our conclusion is : Although $\mathbf{b}_1$ is unbiased, it is less precise than $\mathbf{b}_1^*$ which results from using the correct SRF.
(4) Since in the true PRF, $\beta_2=\mathbf{0}$, it is obvious that $(n-K_1)\sigma^2$ is a unbiased estimator for $\mathbf{e}_1^{*\prime}\mathbf{e}_1^*$.
(5) As we know, restrictions reduce R^2, i.e., $R_*^2\leqslant R^2$. We know the goodness of fit is increased if we include irrelevant variables.
Conclusions: If we include irrelevant variables in the regression, our estimates of both β_1 and σ^2 are unbiased. But we get less precise estimators. We also find the goodness of fit is increased.

3. Dummy Variables

(1) Allow intercept difference only: $y_i=\beta_1+\beta_2x_{i2}+\beta_3d_i+\varepsilon_i$, where d_i is a dummy variable. Then $y_i=\beta_1+\beta_2x_{i2}$, when $d_i=0$; $y_i=(\beta_1+\beta_3)+\beta_2x_{i2}$, when $d_i=1$.
(2) Allow slope difference only: $y_i=\beta_1+\beta_2x_{i2}+\beta_3x_{i2}d_i+\varepsilon_i$, where d_i is a dummy variable. Then $y_i=\beta_1+\beta_2x_{i2}$, when $d=0$; $y_i=\beta_1+(\beta_2+\beta_3)x_{i2}$, when $d_i=1$.
(3) Allow both intercept and slope difference : $y_i=\beta_1+\beta_2x_{i2}+\beta_3d_i+\beta_4x_{i2}d_i+\varepsilon_i$, where d_i is a dummy, and $x_{i2}d_i$ is the interaction term. Then $y_i=\beta_1+\beta_2x_{i2}$, when $d_i=0$; $y_i=(\beta_1+\beta_3)+(\beta_2+\beta_4)x_{i2}$, when $d_i=1$.
To test whether there exists intercept difference or slope difference, construct the following null hypothesis: $H_0:\beta_3=0$ or $H_0:\beta_4=0$ or $H_0:\beta_3=\beta_4=0$, and use F-test of the form: $\frac{(SSE_R-SSE_U)/J}{SSE_U/(n-K)}$.
(4) Allow kinks to occur, i.e., a spline regression.
If we want to use dummy variables to express the following model:

$$income=\begin{cases}\alpha^0+\beta^0 age & \text{if } age<t_1^*\\ \alpha^1+\beta^1 age & \text{if } t_1^*\leqslant age<t_2^*\\ \alpha^2+\beta^2 age & \text{if } age\geqslant t_2^*\end{cases}$$

Let $d_1=1$ if $age\geqslant t_1^*$; $=0$, otherwise. Let $d_2=1$ if $age\geqslant t_2^*$; $=0$, otherwise.
Set up the following model:
$income=\beta_1+\beta_2 age+\delta_1d_1(age-t_1^*)+\delta_2d_2(age-t_2^*)$
(using a special interaction term for each knot.)
(5) Avoid dummy variable trap in two categories:
To analyze the models with different mean $y_i=\mu+\varepsilon_i$ and $y_i=\mu+\delta+\varepsilon_i$, we can set up dummies in two ways:

(a) with an overall intercept and a dummy $y_i = \mu + \delta d_i + \varepsilon_i$,then $X = \begin{pmatrix} \mathbf{i}_1 & \mathbf{0} \\ \mathbf{i}_2 & \mathbf{i}_2 \end{pmatrix}$;
(b) with no constant term and two dummy $y_i = \mu h_i + \delta d_i + \varepsilon_i$,then $X = \begin{pmatrix} \mathbf{i}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{i}_2 \end{pmatrix}$;
(c) if we set up the model with an overall intercept and two dummies, then we fall into the trap. $y_i = \mu + \delta_1 h_i + \delta_2 d_i + \varepsilon_i$,where $h_i = 1 - d_i$, then $X = \begin{pmatrix} \mathbf{i}_1 & \mathbf{i}_1 & \mathbf{0} \\ \mathbf{i}_2 & \mathbf{0} & \mathbf{i}_2 \end{pmatrix}$ which doesn't have full column rank.

(6) Avoid dummy variables trap in Multi-categories.
Consider the model $C_t = \beta_1 + \beta_2 x_t + \varepsilon_t$.
There are two ways of setting up the seasonal model to avoid dummy variable trap.
(a) include the overall constant and drop the dummy variable for the fourth quarter $C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t$
(b) drop the overall constant and include the fourth dummy $C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t$
Another example:
Consider the model: $income = \beta_1 + \beta_2 age + \varepsilon$
Suppose there are four possible categories of education level that might affect income, namely high school, BA, MA, Ph.D. We could set up the model in the following way: $income = \beta_1 + \beta_2 age + \delta_1 BA + \delta_2 MA + \delta_3 Ph.D. + \varepsilon$

(7) Test on pooling sample:

$y_i = \alpha_1 + \alpha_2 X_{i2} + ... + \alpha_K X_{iK} + \varepsilon_i,\ i = 1, ..., N$ with SSE_{U1}
$y_i = \delta_1 + \delta_2 X_{i2} + ... + \delta_K X_{iK} + \varepsilon_i,\ i = N+1, ..., N+M$ with SSE_{U2}.

(a) construct the restricted model as the following:
$y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_K X_{iK} + u_i$ with $H_0 : \alpha_1 = \delta_1, \alpha_2 = \delta_2, ..., \alpha_K = \delta_K$
F-test:

$$\frac{(SSE_R - SSE_U)/K}{SSE_U/(N+M-2K)} \sim F(K, N+M-2K),$$

where $SSE_U = SSE_{U1} + SSE_{U2}$.
(b) construct the restricted model as the following:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 X_{i2} + ... + \beta_K X_{iK} \\ &\quad + h_1 D_i + h_2 D_i X_{i2} + ... + h_K D_i X_{iK} + \nu_i \\ H_0 &: h_1 = h_2 = ... = h_K = 0 \end{aligned}$$

Wald test: $R = [\ O_k \quad I_k\ ]$, $q = 0$.

$$\frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})/K}{\mathbf{e}'\mathbf{e}/(N+M-2K)} \sim F(K, N+M-2K).$$

4. Test on Pooling Sample

(1) $y_i = a_1 + b_1 X_{1i} + e_{1i}, i = 1, ..., n_1$ vs. $y_i = a_2 + b_2 X_{2i} + e_{2i}, i = 1, ..., n_2$
$H_0 : b_1 = b_2$

(2) Method I: suppose two sample disturbances have the same variance.
The test statistic is

$$\frac{b_1 - b_2}{\sqrt{\frac{S^2}{\sum (X_{1i} - \bar{X}_1)^2} + \frac{S^2}{\sum (X_{2i} - \bar{X}_2)^2}}} \sim t(n_1 + n_2 - 4),$$

where $S^2 = \frac{SSE_1 + SSE_2}{n_1 + n_2 - 4}$.

(3) Method II: suppose two sample disturbances have different variance.
The test statistic is

$$\frac{b_1 - b_2}{\sqrt{Est.Var(b_1) + Est.Var(b_2)}} \sim N(0, 1).$$

CHAPTER 9

Data Problem

1. Missing Observations on Simple Regressions

(1) Don't attempt to fill the missing dependent variables.

(2) Simple regressions with constant $\mathbf{y} = \alpha\mathbf{i}+\beta\mathbf{x}+\varepsilon$, where $\mathbf{y} = \begin{pmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{pmatrix}, \mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix}$, and $\mathbf{x}_B$ is missing. Consider the following two approaches.

(a) Replace $\mathbf{x}_B$ with $\bar{\mathbf{x}}_A$. This is equivalent to dropping the missing observations in that $\hat{\bar{\mathbf{x}}}_B - \bar{\mathbf{x}}_A = 0$ implies no change in the sample moments. The only thing gets worse is the R^2 because of the more number of observations.

(b) Fill $\mathbf{x}_B$ with zeros and add a dummy variable that takes value one for missing observations and zero for complete ones. This is algebraically identical to simply filling the gaps with $\bar{\mathbf{x}}_A$.

2. Missing Observations on Multiple Regressions

Multiple regressions with constant $\mathbf{y} = \alpha\mathbf{i}+\beta\mathbf{x}+\gamma\mathbf{z}+\varepsilon$, where $\mathbf{y} = \begin{pmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{pmatrix}, \mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix}, \mathbf{z} = \begin{pmatrix} \mathbf{z}_A \\ \mathbf{z}_B \end{pmatrix}$, and $\mathbf{x}_B$ is missing. Here are three approaches handling this particular problem.

(1) Suppose that it were valid to impose a linear relationship between $\mathbf{x}$ and $\mathbf{z}$. Then if $\mathbf{x} = \delta\mathbf{z} + \mathbf{u}$, the model may be rewritten in three equations:

$$\begin{aligned} \mathbf{y}_A &= \alpha\mathbf{i}_A + \beta\mathbf{x}_A + \gamma\mathbf{z}_A + \varepsilon_A\mathbf{x}_A \\ &= \delta\mathbf{z}_A + \mathbf{u}_A\mathbf{y}_B \\ &= \alpha\mathbf{i}_B + (\gamma + \beta\delta)\mathbf{z}_B + \varepsilon_B + \beta\mathbf{u}_B. \end{aligned}$$

Each of the first two equations can be estimated by OLS. Let $\hat{\bar{\mathbf{x}}}_B$ be the predicted mean value of the missing $\mathbf{x}_B$ obtained by using $\hat{\delta}$ and $\mathbf{z}_B$. Consider combining the two data sets in one regression model as the following,

$$\begin{pmatrix} \mathbf{y}_A - \hat{\beta}\mathbf{x}_A \\ \mathbf{y}_B - \hat{\beta}\mathbf{x}_B \end{pmatrix} = \alpha \begin{pmatrix} \mathbf{i}_A \\ \mathbf{i}_B \end{pmatrix} + \gamma \begin{pmatrix} \mathbf{z}_A \\ \mathbf{z}_B \end{pmatrix} + \nu.$$

Assuming that ν and $\mathbf{z}$ are uncorrelated (at least asymptotically), γ can be estimated by least squares. Note that although we have done nothing to the original estimate of β, some new information is being used to estimate γ in the second regression, which can be expected to provide added efficiency.

(2) Fill $\mathbf{x}_B$ with zeros and add a dummy that takes value one for missing observations and zero for complete ones. The regression model is now

$$\begin{pmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{pmatrix} = \alpha \begin{pmatrix} \mathbf{i}_A \\ \mathbf{i}_B \end{pmatrix} + \beta \begin{pmatrix} \mathbf{x}_A \\ \mathbf{0} \end{pmatrix} + \gamma \begin{pmatrix} \mathbf{z}_A \\ \mathbf{z}_B \end{pmatrix} + \eta \begin{pmatrix} \mathbf{D}_A \\ \mathbf{D}_B \end{pmatrix} + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix},$$

where $\mathbf{D}_A = \mathbf{0}$ and $\mathbf{D}_B = \mathbf{1}$. Then we have the following results

$$\mathbf{y}_A = \alpha \mathbf{i}_A + \beta \mathbf{x}_A + \gamma \mathbf{z}_A + \varepsilon_A \Rightarrow \bar{\mathbf{y}}_A - \hat{\alpha} - \hat{\gamma}\bar{\mathbf{z}}_A = \hat{\beta}\bar{\mathbf{x}}_A$$

$$\mathbf{y}_B = \alpha \mathbf{i}_B + \gamma \mathbf{z}_B + \eta \mathbf{i}_B + \varepsilon_B \Rightarrow \hat{\eta} = \bar{\mathbf{y}}_B - \hat{\alpha} - \hat{\gamma}\bar{\mathbf{z}}_B = \hat{\beta}\bar{\mathbf{x}}_B$$

$$\therefore \hat{\eta} = \hat{\beta}\bar{\mathbf{x}}_B.$$

(3) Fill $\mathbf{x}_B$ with $\bar{\mathbf{x}}_A$ and add a dummy as defined in Approach Two. The regression model is now

$$\begin{pmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{pmatrix} = \alpha \begin{pmatrix} \mathbf{i}_A \\ \mathbf{i}_B \end{pmatrix} + \beta \begin{pmatrix} \mathbf{x}_A \\ \bar{\mathbf{x}}_A \end{pmatrix} + \gamma \begin{pmatrix} \mathbf{z}_A \\ \mathbf{z}_B \end{pmatrix} + \eta \begin{pmatrix} \mathbf{D}_A \\ \mathbf{D}_B \end{pmatrix} + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix},$$

and we have the following results:

$$\mathbf{y}_A = \alpha \mathbf{i}_A + \beta \mathbf{x}_A + \gamma \mathbf{z}_A + \varepsilon_A \Rightarrow$$

$$\bar{\mathbf{y}}_A - \hat{\alpha} - \hat{\gamma}\bar{\mathbf{z}}_A = \hat{\beta}\bar{\mathbf{x}}_A$$

$$\mathbf{y}_B = \alpha \mathbf{i}_B + \beta \bar{\mathbf{x}}_A + \gamma \mathbf{z}_B + \eta \mathbf{i}_B + \varepsilon_B \Rightarrow$$

$$\hat{\eta} = \bar{\mathbf{y}}_B - \hat{\alpha} - \hat{\gamma}\bar{\mathbf{z}}_B - \hat{\beta}\bar{\mathbf{x}}_A = \hat{\beta}(\bar{\mathbf{x}}_B - \bar{\mathbf{x}}_A).$$

We actually have a natural test statistic $E(\hat{\eta}) = 0$ for whether the data $\mathbf{x}_B$ are missing at random.

CHAPTER 10

Generalized Least Square Model

1. Classical Model

This model is also known as a model with spherical disturbances, i.e., $y = X\beta + \varepsilon$, $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon') = \sigma^2 I$. OLS: $b = \beta + (X'X)^{-1}X'\varepsilon$. The OLS estimators have the following properties: **b** is BLUE and CAN; **b** is also asymptotically efficient assuming normally distributed disturbances; $S^2 = e'e/(n-K)$ is an unbiased estimator for σ^2.

2. Generalized Model

Here is a generalized model with non-spherical disturbances, $y = X\beta+\varepsilon$, $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon') = \sigma^2\Omega$.Recall that the OLS procedure minimizes (equally weighted) sum of squared errors. The GLS procedure is utilized in presence of heteroskedasticity or auto-correlation, by using a different weighting scheme for the weighted sum of squared errors. In particular, the errors that are known to have large variance themselves are assigned with a smaller weight, so are the errors that are known to have large covariance with other errors.

If we still use OLS estimator, then $b = \beta + (X'X)^{-1}X'\varepsilon$ implies the following properties: **b** is still unbiased and CAN; **b** is no longer efficient even if the disturbances are normally distributed; $S^2 = e'e/(n-K)$ may be biased for σ^2.

The conventional estimated variances of OLS estimates are no longer unbiased due to the nature of heteroskedasticity or auto-correlation, and the extent of biasness (upward or downward) varies in different applications. Therefore, we can't draw proper inference based on OLS estimators.

If the disturbances are normally distributed, the OLS estimates are not the same as MLE, but the GLS estimates coincides with MLE. Despite the superiority of GLS over OLS in this particular context, the implementation of GLS requires advance knowledge about the variance structure of the disturbances, a rare case in reality. So people developed FGLS that uses the estimated covariance matrix of errors from the OLS, and others simply use the "White-washed", or heteroskedasticity-corrected, variance for the OLS estimates. One alternative version of heteroskedasticity-corrected variance, known as Newey-West correction, is also very popular.

Because the covariance matrix of the disturbance term is often large, it's computationally hard to find FGLS estimators. An easier way will be to perform an OLS on transformed data that make the errors spherical. In the case of heteroskedasticity, the transformation is to divide all the data (including the constant term) by the square root of the respective error variance. It's demonstrated that the OLS on the post-transformed data is equivalent to the GLS on the original data.

We typically use the Durbin-Waston statistic to detect the auto-correlation in the regression errors, whereas a D-W value of 2 indicates no auto-correlation. If the null hypothesis of zero (first-order) auto-correlation is rejected, then we can use an estimated $\hat{\rho}$ to transform the data as follows: replace all the data x_t with $x_t - \hat{\rho}x_{t-1}$ for $t \geq 2$ and replace x_1 with $\sqrt{1-\hat{\rho}^2}x_1$. The OLS estimates on the post-transformed data boils down to the FGLS estimates for the original specification. Be cautions of the source of auto-correlation though, as a mis-specfied model (say, omitting a relevant explanatory variable) can lead to auto-correlation.

(1) Finite Sample properties of b in GLR model
   (a) $E(b) = E_X[E(b|X)] = \beta$
   If regressors and disturbances are uncorrelated, LS estimator is still unbiased in GLR model.
   (b)

$$\begin{aligned} Var(b|X) &= E[(b-\beta)(b-\beta)'|X] \\ &= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\left(\frac{X'X}{n}\right)^{-1} \end{aligned}$$

   If the disturbances are normally distributed,

$$b \sim N[\beta, \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}].$$

(2) Asymptotic properties of b in GLR model
   (a) Consistency
   We have $E(b) = \beta$, and $Var(b) = \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}$. If $p\lim Var(b) = 0$, according to the definition of convergence in mean square, then we have $p\lim b = \beta$. Therefore, if $p\lim\left(\frac{X'X}{n}\right)$ and $p\lim\left(\frac{X'\Omega X}{n}\right)$ are both finite positive definite matrices, then b is consistent for β, i.e., $p\lim b = \beta$.
   (b) Asymptotic normal distribution

$$\begin{aligned} \hat{\beta}_{LS} &= \beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X/n)^{-1}(X'\varepsilon/n) \Rightarrow \\ \sqrt{n}(\hat{\beta}_{LS} - \beta) &= (X'X/n)^{-1}(X'\varepsilon/\sqrt{n}) \end{aligned}$$

   Let $p\lim\left(\frac{X'X}{n}\right) = Q$. If $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$, then recall 6.25 and we get $X'\varepsilon/\sqrt{n} \xrightarrow{d} N(0, \sigma^2 Q) \Rightarrow \sqrt{n}(\hat{\beta}_{LS} - \beta) = Q^{-1}(X'\varepsilon/\sqrt{n}) \xrightarrow{d} N[0, Q^{-1}(\sigma^2 Q)Q^{-1}] = N(0, \sigma^2 Q^{-1})$ i.e., $\hat{\beta}_{LS} \xrightarrow{a} N[\beta, \sigma^2 Q^{-1}/n]$.
   Note that we estimate Q^{-1}/n with $(X'X)^{-1}$ and $\sigma^2$with $e'e/(n-K)$.
   If $\varepsilon \sim \mathrm{N}(0, \sigma^2\Omega)$, similarly we get $X'\varepsilon/\sqrt{n} \xrightarrow{d} N\left[0, \sigma^2\left[p\lim\left(\frac{X'\Omega X}{n}\right)\right]\right]$
   $\Rightarrow \sqrt{n}(\hat{\beta}_{LS} - \beta) = Q^{-1}(X'\varepsilon/\sqrt{n}) \xrightarrow{d} N[0, \sigma^2 Q^{-1} p\lim(X'\Omega X/n))Q^{-1}]$
   i.e., $\hat{\beta}_{LS} \xrightarrow{a} N[\beta, (\sigma^2/n)Q^{-1}p\lim(X'\Omega X/n)Q^{-1}]$.
   (c) Conclusions:
      (i) In the heteroskedastic case, if the variances of disturbances are finite and not dominated by any single term, then the least

squares estimator is asymptotically normally distributed with covariance matrix:

$$Asy.Var(b) = (\sigma^2/n)Q^{-1}p\lim(X'\Omega X/n)Q^{-1};$$

(ii) In case that Ω is known,

$$Est.Var(b) = \frac{1}{n}\cdot\left(\frac{X'X}{n}\right)^{-1}\cdot\left(\frac{X'\sigma^2\Omega X}{n}\right)\cdot\left(\frac{X'X}{n}\right)^{-1};$$

(iii) In case that Ω is unknown, using White heteroskedasticity consistent estimator:

$$\begin{aligned} Est.Asy.Var(b) &= \frac{1}{n}\cdot\left(\frac{X'X}{n}\right)^{-1}\cdot\left(\frac{\sum e_i^2 x_i x_i'}{n}\right)\cdot\left(\frac{X'X}{n}\right)^{-1} \\ &= n(X'X)^{-1}S_0(X'X)^{-1}. \end{aligned}$$

(iv) Without specifying the distribution of disturbances, we cannot use F statistic, and the likelihood ratio and Lagrangian Multiplier tests are also not available. However, we can use the Wald statistic as well as asymptotic "t-ratio".

(3) Efficient estimation when Ω is a known, symmetric, positive definite matrix: $y = X\beta + \varepsilon$,where $\varepsilon \sim (0, \sigma^2\Omega)$.

Define $PP' \equiv \Omega^{-1}$, then we have $P'\Omega P = I$.

Pre-multiply $y = X\beta + \varepsilon$ by P', we have $P'y = P'X\beta + P'\varepsilon$, i.e., $y^* = X^*\beta + \varepsilon^*$.

From $E(\varepsilon^*\varepsilon^{*\prime}) = E(P'\varepsilon\varepsilon' P) = \sigma^2 P'\Omega P = \sigma^2 I$, we know $y^* = X^*\beta + \varepsilon^*$is the classical model, where $\varepsilon^* \sim (0, \sigma^2 I)$.

According to the OLS rules, we have

$$\begin{aligned} b^* &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X'PP'X)^{-1}X'PP'y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, \\ E(b^*) &= \beta, \\ Var(b^*) &= \sigma^2(X'\Omega^{-1}X)^{-1}. \end{aligned}$$

Hence we have

$$\begin{aligned} \hat{\beta}_{GLS} &= b^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y, \\ E(\hat{\beta}_{GLS}) &= E(b^*) = \beta, \\ Var(\hat{\beta}_{GLS}) &= Var(b^*) = \sigma^2(X'\Omega^{-1}X)^{-1}, \\ \hat{\sigma}^2_{GLS} &= (y^* - X^*b^*)'(y^* - X^*b^*)/(n-K) \\ &= (y - X\hat{\beta}_{GLS})'\Omega^{-1}(y - X\hat{\beta}_{GLS})/(n-K), \\ E(\hat{\sigma}^2_{GLS}) &= \sigma^2. \end{aligned}$$

For the sake of comparison, we list the following OLS results:

$$\begin{aligned}\hat{\beta}_{OLS} &= (X'X)^{-1}X'y,\\ E(\hat{\beta}_{OLS}) &= \beta,\\ Var(\hat{\beta}_{OLS}) &= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1},\\ \hat{\sigma}^2_{OLS} &= (y - X\hat{\beta}_{OLS})'(y - X\hat{\beta}_{OLS})/(n-K).\end{aligned}$$

(4) Properties of $\hat{\beta}_{GLS}$ inherited from b^*:
   (a) b^* is unbiased, thus $\hat{\beta}_{GLS}$ is also unbiased;
   (b) b^* is consistent as long as $p\lim X^{*\prime}X^*/n = Q^*$ is a finite and positive definite matrix; $\hat{\beta}_{GLS}$ is consistent as long as $p\lim X^{*\prime}X^*/n = p\lim X'\Omega^{-1}X/n = Q^*$ is finite and positive definite.
   (c) b^* is asymptotically normally distributed , with mean β and variance $\sigma^2(X^{*\prime}X^*)^{-1}$; $\hat{\beta}_{GLS}$ is asymptotically normally distributed , with mean β and variance $\sigma^2(X'\Omega^{-1}X)^{-1}$.
   (d) b^* is MVLUE by Gauss-Markov theorem; $\hat{\beta}_{GLS}$ is MVLUE by extended Gauss-Markov theorem.
   (e) Testing $H_0 : R\beta = q$, we use the following F test.

$$\begin{aligned}F(J, n-K) &= (Rb^* - q)'[S^{*2}R(X^{*\prime}X^*)^{-1}R']^{-1}(Rb^* - q)/J\\ &= (R\hat{\beta}_{GLS} - q)'[S^{*2}R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\hat{\beta}_{GLS} - q)/J\\ &= \frac{(e_c^{*\prime}e_c^* - e^{*\prime}e^*)/J}{e^{*\prime}e^*/(n-K)}\\ &= \frac{(R\hat{\beta}_{GLS} - q)'[\sigma^2R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\hat{\beta}_{GLS} - q)/J}{S^{*2}/\sigma^2}\end{aligned}$$

$$\begin{aligned}S^{*2} &= e^{*\prime}e^*/(n-K)\\ &= (y^* - X^*b^*)'(y^* - X^*b^*)/(n-K)\\ &= (y - X\hat{\beta}_{GLS})'\Omega^{-1}(y - \hat{\beta}_{GLS})/(n-K)\end{aligned}$$

$$\begin{aligned}b_c^* &= b^* - (X^{*\prime}X^*)^{-1}R'[R(X^{*\prime}X^*)^{-1}R']^{-1}(Rb^* - q)\\ &= b^* - (X'\Omega^{-1}X)^{-1}R'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\hat{\beta}_{GLS} - q)\end{aligned}$$

(5) Assuming normality, i.e., $\varepsilon \sim \mathrm{N}(0, \sigma^2\Omega)$, then we have the following MLE:

$$\begin{aligned}\hat{\beta}_{ML} &= \hat{\beta}_{GLS} = b^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y,\\ E(\hat{\beta}_{ML}) &= E(\hat{\beta}_{GLS}) = E(b^*) = \beta,\\ Var(\hat{\beta}_{ML}) &= Var(\hat{\beta}_{GLS}) = Var(b^*) = \sigma^2(X'\Omega^{-1}X)^{-1},\\ \hat{\sigma}^2_{ML} &= (y - X\hat{\beta}_{ML})'\Omega^{-1}(y - X\hat{\beta}_{ML})/n.\end{aligned}$$

Again, we find that both $\hat{\beta}_{ML}$ and $\hat{\beta}_{GLS}$ are unbiased and efficient. $\hat{\sigma}^2_{ML}$is biased and $\hat{\sigma}^2_{GLS}$ is unbiased.

(6) Estimation when Ω is unknown.
   (a) Overall idea:

(i) when Ω is unknown, $\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ is not feasible. Hence we need a good enough estimator $\hat{\Omega}$ for Ω. If we plug $\hat{\Omega}$ into $\hat{\beta}_{GLS}$, we get $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$.

(ii) Suppose we know the structure of Ω is $\Omega = \Omega(\theta)$. If $\hat{\theta}$ is a consistent estimator for θ, then by Slutsky theorem, we know $\hat{\Omega} = \Omega(\hat{\theta})$ is consistent for $\Omega(\theta)$, and we say $\hat{\Omega}$ is good enough.

(iii) Moreover, for $\hat{\beta}_{FGLS}$ to be asymptotically equivalent to $\hat{\beta}_{GLS}$ so that $\hat{\beta}_{FGLS}$ have all desired properties that $\hat{\beta}_{GLS}$ inherited from b^*_{LS}, there are certain conditions to be satisfied: $p\lim(\frac{X'\hat{\Omega}^{-1}X}{n} - \frac{X'\Omega^{-1}X}{n}) = 0$ and $p\lim(\frac{X'\hat{\Omega}^{-1}\varepsilon}{\sqrt{n}} - \frac{X'\Omega^{-1}\varepsilon}{\sqrt{n}}) = 0$. Of course, we don't need to worry about those conditions so far.

(iv) An important theorem: for $\hat{\beta}_{FGLS}$ to be asymptotically efficient, we don't need to have an efficient estimator of θ, and only a consistent one is required to achieve full efficiency for $\hat{\beta}_{FGLS}$.

(b) As long as the information matrix is block diagonal, the GLS, FGLS and ML estimators of β have the same asymptotically distribution. Particularly, $Asy.Var(\hat{\beta}_{ML}) = \sigma^2(X'\Omega^{-1}X)^{-1}$.

(c) For the case of group-wise heteroskedasticity, to get the $\hat{\beta}_{MLE}$, follow the following steps:
Step 1: get OLS estimator b for the pooling data;
Step 2: get ML estimator $\hat{\sigma}^2_g$ for each group using $\hat{\sigma}^2_g = e'_g e_g/n$, and $e_g = y_g - X_g b$;
Step 3: get ML estimator $\hat{\beta}_{ML}$ using

$$\hat{\beta}_{ML} = \left[\sum_{g=1}^{G} \frac{1}{\hat{\sigma}^2_g} X_g X'_g\right]^{-1} \left[\sum_{g=1}^{G} \frac{1}{\hat{\sigma}^2_g} X_g y_g\right];$$

Step 4: if $\hat{\beta}_{ML}$ has not yet converged, go to step 2 while substituting b with $\hat{\beta}_{ML}$; otherwise, exit.
Notes for step 3: Refer back to 9.3 about grouped data:

$$b^* = (\bar{x}^{*\prime}\bar{x}^*)^{-1}(\bar{x}^{*\prime}\bar{y}^*) = \left[\sum_{g=1}^{G} n_g \bar{x}_g \bar{x}'_g\right]^{-1} \left[\sum_{g=1}^{G} n_g \bar{x}_g \bar{y}_g\right],$$

where $\bar{x}_g, \bar{y}_g$ are the grouped data mean. Since $\sigma^2_g = \sigma^2/n_g$ (suppose σ^2 is the variance of disturbances for the ungrouped data), we use $\sqrt{n_g}$ as weight to get $[\sqrt{n_g}\sigma_g]^2 = \sigma^2$.
Here we have $n_g = \sigma^2/\sigma^2_g$, $\bar{x}_g = X_g$, $\bar{y}_g = y_g$, since σ^2 will be cancelled out, we can equally well normalize it to 1.

3. Heteroskedasticity

(1) For a heteroskedasticity model with $\sigma^2_i = \sigma^2\omega_i$, we let P_{ii} be $1/\sqrt{\omega_i}$ so that $[\sigma_i/\sqrt{\omega_i}]^2 = \sigma^2$.

(2) OLS estimator for heteroskedasticity model

(a) $Asy.Var(b) = \frac{\sigma^2}{n}\left[p\lim\left(\frac{X'X}{n}\right)\right]^{-1} p\lim\left(\frac{X'\Omega X}{n}\right)\left[p\lim\left(\frac{X'X}{n}\right)\right]^{-1}$

Assuming $p\lim(X'X/n) = Q$, $p\lim(X'\Omega X/n) = p\lim(\frac{1}{n}\sum_{i=1}^{n}\omega_i x_i x_i')$ $= Q^*$, we have $b \xrightarrow{a} N[\beta,\ \frac{\sigma^2}{n}Q^{-1}Q^*Q^{-1}]$ and

$$Est.Asy.Var(b) = (X'X)^{-1}(\sigma^2\sum_{i=1}^{n}\omega_i x_i x_i')(X'X)^{-1}.$$

(b)

$$Var(b) = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1};$$
$$Var(\hat{\beta}_{GLS}) = \sigma^2(X'\Omega^{-1}X)^{-1}.$$

b is less efficient than $\hat{\beta}_{GLS}$.

(c) If the heteroskedasticity is not correlated with the variables in the model, then at least in large samples, it is tolerable, although not optimal, to use $Est.Var(b) = S^2(X'X)^{-1}$ to estimate $Var(b) = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$.

(d) One of the appropriate way of estimating Var(b) for OLS is White estimator:

$$Var(b) = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}.\frac{1}{n}\sigma^2 X'\Omega X$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 x_i x_i'.$$

Define $S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$, where e_i is the ith least squares residual .
$Est.Var(b) = n(X'X)^{-1}S_0(X'X)^{-1}$.

(3) Testing for group-wise heteroskedasticity
$H_0 : \sigma_1^2 = ... = \sigma_G^2 (G-1$ restrictions)
Statistic: $n\ln S^2 - \sum_{g=1}^{G} n_g \ln S_g^2 \sim \chi^2(G-1)$, where $S^2 = e'e/n$,
$S_g^2 = e_g'e_g/n_g$ are from grouped data.

(4) GLS when Ω is known

(a) For the case of $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2\omega_i$, let P_{ii} be $1/\sqrt{\omega_i}$ so that $Var(\varepsilon_i/\sqrt{\omega_i}) = \sigma^2$. Or equivalently, $diag(\Omega) = (\ \omega_1 \quad \omega_2 \quad ... \quad \omega_n\ )$
$\Rightarrow diag(\Omega^{-1}) = (1/\ \omega_1 \quad 1/\omega_2 \quad ... \quad 1/\omega_n\ )$
$\Rightarrow diag(P) = (\ 1/\sqrt{\omega_1} \quad 1/\sqrt{\omega_2} \quad ... \quad 1/\sqrt{\omega_n}\ )$,
$P'y_i = y_i/\sqrt{\omega_i}$, $P'x_i = x_i/\sqrt{\omega_i}$, $P'\varepsilon_i = \varepsilon_i/\sqrt{\omega_i}$.
According to the formula of $\hat{\beta}_{GLS}$ for group-wise model, we have

$$\hat{\beta}_{GLS} = [(P'X)'(P'X)]^{-1}[(P'X)'(P'y)]$$
$$= \left(\sum_{i=1}^{n} x_i x_i'/\omega_i\right)\left(\sum_{i=1}^{n} x_i y_i/\omega_i\right)$$
$$= \left(\sum_{i=1}^{n} w_i x_i x_i'\right)\left(\sum_{i=1}^{n} w_i x_i y_i\right)$$

where $w_i = 1/\omega_i = P_{ii}^2$.
We also name it as weighted least squares estimator.

Note: Observations with smaller variance receive a larger weight in the computation of the sums and therefore have greater influence in the estimates obtained.

(b) In the case of $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 x_k^2$, let P_{ii} be $1/x_k$ so that $Var(\varepsilon_i/x_k) = \sigma^2$. We have weights $w_i = 1/x_k^2$ for WLS.

(c) In the case of $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 x_k$, let P_{ii}be $1/\sqrt{x_k}$ so that $Var(\varepsilon_i/\sqrt{x_k}) = \sigma^2$. We have weights $w_i = 1/x_k$for WLS.

(d) The weighted least square estimator

$$\hat{\beta}_{WLS} = \left(\sum_{i=1}^{n} w_i x_i x_i'\right)\left(\sum_{i=1}^{n} w_i x_i y_i\right)$$

is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances, but improperly weighted least squares estimator is inefficient.

4. Autocorrelated Disturbances

(1) White noise ε_t satisfies: zero mean, constant variance and zero covariance between any two disturbances in different periods. This is also the definition of covariance stationary or weakly stationary.

(2) Suppose disturbances are homoskedastic, but correlated across observations, then $E(\varepsilon\varepsilon') = \sigma^2\Omega$, where $\sigma^2\Omega$ is a full rank, positive definite matrix with a constant σ^2 on the diagonal.

Impose stationarity further, i.e., Ω_{ts} is a function of $|s-t|$, but not of t or s alone.

Auto-covariances: $\gamma_s = Cov(\varepsilon_t, \varepsilon_{t-s}) = Cov(\varepsilon_{t+s}, \varepsilon_t)$, particularly, $\gamma_0 = \sigma^2$.

Auto-correlation: $\rho_s = \gamma_s/\gamma_0 = Cov(\varepsilon_t, \varepsilon_{t-s})/\sqrt{Var(\varepsilon_t)Var(\varepsilon_{t-s})}$, particularly, $\rho_0 = 1$.

(3) Stationary AR(1) Process: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, where $|\rho| < 1$, and u_t is classical.

Then $E(\varepsilon_t) = 0$, $\sigma_\varepsilon^2 = \sigma_u^2/(1-\rho^2)$, $\gamma_s = \rho^s\sigma_u^2/(1-\rho^2)$, (particularly, $\gamma_0 = \sigma_\varepsilon^2$), $\rho_s = \rho^s$.

Thus we have $\sigma_\varepsilon^2\Omega = \frac{\sigma_u^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & 1 \end{pmatrix}$.

(4) OLS estimators

If the regression doesn't contain lagged dependent variable, then OLS estimators are unbiased, consistent, asymptotically normally distributed, and inefficient;

If the regression contains lagged dependent variable, then OLS estimators are no longer unbiased or consistent.

(5) GLS estimators ($\Phi = \sigma^2\Omega$is known.)

$$\hat{\beta}_{GLS} = [X'\Phi^{-1}X]^{-1}[X'\Phi^{-1}y],$$
$$Var(\hat{\beta}_{GLS}) = [X'\Phi^{-1}X]^{-1}.$$

For AR(1), $P' = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \cdots & 0 \\ -\rho & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$, then $Var(P'\varepsilon) = \sigma_u^2$, and $PP' = \Phi^{-1}$, $P'\Phi P = I$.

Note that this transformation matrix is called Prais-Winston one. If we drop the first row, then the transformation is called Cochrane-Orcutt one.

(6) FGLS estimators ($\Phi = \sigma^2\Omega$is unknown.)

$$\hat{\beta}_{FGLS} = [X'\hat{\Phi}^{-1}X]^{-1}[X'\hat{\Phi}^{-1}y],$$
$$Var(\hat{\beta}_{FGLS}) = [X'\hat{\Phi}^{-1}X]^{-1}.$$

For AR(1), follow the steps:

Step 1: regress $y = X\beta + \varepsilon$ and get e;

Step 2 regress $e_t = \rho e_{t-1} + u_t$ and get $\hat{\rho}$;

Step 3: $\hat{\Phi} = \Phi(\rho = \hat{\rho})$;

Step 4: plug $\hat{\Phi}$into the formula.

(7) Durbin-Watson Test

For regression satisfying : (1) with a constant term; (2) without lagged dependent variables.

Statistic:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1-r_1) - \frac{(e_1^2 + e_T^2)}{\sum_{t=1}^{T} e_t^2}$$
$$\approx 2(1-r_1),$$

where e_t are the residuals from the original regression, and

$$r_1 = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_t^2},$$

i.e., $r_1 = \hat{\rho}$ for $e_t = \rho e_{t-1} + u_t$.

Decision rule:

For testing positive autocorrelation, reject $H_0 : \rho = 0$ if $d < d_L$; accept $H_0 : \rho = 0$if $d > d_U$; inconclusive if $d_L < d < d_U$.

For testing negative autocorrelation, reject $H_0 : \rho = 0$ if $d > 4 - d_L$; accept $H_0 : \rho = 0$ if $d < 4 - d_U$; inconclusive if $4 - d_U < d < 4 - d_L$.

(8) Durbin-H test:

For regression $Y_t = \beta_1 Y_{t-1} + \beta_2 X_t + \varepsilon_t$, statistic:

$$h = r_1\sqrt{T\Big/(1 - T\cdot\hat{V}(\hat{\beta}_1))} \xrightarrow{a} N(0,1),$$

where T is the number of observations, and $\hat{V}(\hat{\beta}_1)$ is the estimated variance of the coefficient on Y_{t-1}.

CHAPTER 11

Models for Panel Data

1. Panel Data Models

The major advantage of using panel data rather than cross sectional data is that panel data provides us with great flexibility discussing different behavior across individual. The conventional panel data model is: $y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + \varepsilon_{it}$. Note that constant is not part of the K regressors in $\mathbf{x}_{it}$. When α_i is considered stationary across time and fixed within group i, the model is called *fixed effect panel data model*; when α_i is considered a group specific disturbance, the model is called *random effect panel data model*.

2. Fixed Effects

(1) The regression model takes the form

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & & \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \alpha + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or $\mathbf{y} = \mathbf{D}\alpha + \mathbf{X}\beta + \varepsilon$. Suppose that we have T_i observations in the i^{th} group. We can run a regular least square regression on this equation to get the estimators. However, under normal circumstances, we would have way too many dummy variables to handle, and we take the partial regression approach.

(2) Note that $\mathbf{b} = (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}(\mathbf{X}'\mathbf{M}_D\mathbf{y})$, where $\mathbf{M}_D = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$. This is equivalent to run a regression from $\mathbf{M}_D\mathbf{y}$ onto $\mathbf{M}_D\mathbf{X}$. Note further that $\mathbf{M}_D$ has the following nice feature,

$$\mathbf{M}_D = \begin{bmatrix} \mathbf{M}^0_{T_1} & 0 & \cdots & 0 \\ 0 & \mathbf{M}^0_{T_2} & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{M}^0_{T_n} \end{bmatrix},$$

where $\mathbf{M}^0_{T_i} = \mathbf{I}_{T_i} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}' = \mathbf{I}_{T_i} - \frac{1}{T_i}\mathbf{i}\mathbf{i}'$. It is obvious that $\mathbf{M}^0_{T_i}\mathbf{X}_i = \mathbf{X}_i - \bar{\mathbf{X}}_i\mathbf{i}$ and $\mathbf{M}^0_{T_i}\mathbf{y}_i = \mathbf{y}_i - \bar{y}_i\mathbf{i}$. Hence we could get $\mathbf{b}$ from a regression using pooling sample deviations from their respective group means, i.e.,

$$\begin{bmatrix} \mathbf{y}_1 - \bar{y}_1\mathbf{i} \\ \mathbf{y}_2 - \bar{y}_2\mathbf{i} \\ \vdots \\ \mathbf{y}_n - \bar{y}_n\mathbf{i} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 - \bar{\mathbf{X}}_1\mathbf{i} \\ \mathbf{X}_2 - \bar{\mathbf{X}}_2\mathbf{i} \\ \vdots \\ \mathbf{X}_n - \bar{\mathbf{X}}_n\mathbf{i} \end{bmatrix} \mathbf{b} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_n \end{bmatrix}.$$

Similarly, we have $\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b})$.

(3) $Est.Var(\mathbf{b}) = S^2(\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}$ and $Est.Var(a_i) = \frac{1}{T_i}S^2 + \bar{\mathbf{X}}_i Var(\mathbf{b})\bar{\mathbf{X}}_i'$. $S^2 = \frac{\sum_{i=1}^n \mathbf{e}_i'\mathbf{e}_i}{\sum_{i=1}^n T_i - K - n}$.(Note the correction of degrees of freedom for $n \quad a_i$'s. $R^2 = 1 - \frac{\sum_{i=1}^n \mathbf{e}_i'\mathbf{e}_i}{\sum_{i=1}^n \mathbf{y}_i'\mathbf{M}^0_{T_i}\mathbf{y}_i}$.

(4) $H_0 : \alpha_i = 0 \ \forall i$ We have $\frac{\alpha_i}{\sqrt{Est.Var(\alpha_i)}} \sim t(\sum_{i=1}^n T_i - K - n)$, but this is not a useful hypothesis to test. $H_0 : a_1 = a_2 = ... = a_n (n-1$ restrictions) We have $\frac{(R^2 - R_*^2)/(n-1)}{(1-R^2)/(\sum_{i=1}^n T_i - n - K)} \sim F(n-1, \sum_{i=1}^n T_i - n - K)$. Note that under the null, we use the pooling sample data to get the efficient estimators and R_*^2. Note further that in the unrestricted model we could have estimated the model with an overall constant and $n-1$ dummy variables. But this alternative method will produce the same results except that the interpretation for coefficients associated with the dummies would be different.

(5) Comparison on three alternative models

(a) Overall Model (OA): $y_{it} = \alpha_i + \beta'\mathbf{x}_{it} + \varepsilon_{it}$
Within-group Model (WG): $y_{it} - \bar{y}_{i\cdot} = \beta'(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot}) + \varepsilon_{it} - \bar{\varepsilon}_{i\cdot}$.
Between-group Model (BG): $\bar{y}_{i\cdot} = \alpha_i + \beta'\bar{\mathbf{x}}_{i\cdot} + \bar{\varepsilon}_{i\cdot}$.

(b) Sample moments
Corresponding to three models, we have the sample moments as the following:

$$\mathbf{S}_{XX}^{OA} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})';$$

$$\mathbf{S}_{Xy}^{OA} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\bar{\mathbf{x}}})(y_{it} - \bar{\bar{y}});$$

$$\mathbf{S}_{XX}^{WG} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot})' = \sum_i^n \mathbf{X}_i'\mathbf{M}^0_{T_i}\mathbf{X}_i;$$

$$\mathbf{S}_{Xy}^{WG} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i\cdot})(y_{it} - \bar{y}_{i\cdot}) = \sum_i^n \mathbf{X}_i'\mathbf{M}^0_{T_i}\mathbf{y}_i;$$

$$\mathbf{S}_{XX}^{BG} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\bar{\mathbf{x}}_{i\cdot} - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_{i\cdot} - \bar{\bar{\mathbf{x}}})';$$

$$\mathbf{S}_{Xy}^{BG} = \sum_{i=1}^n \sum_{t=1}^{T_i} (\bar{\mathbf{x}}_{i\cdot} - \bar{\bar{\mathbf{x}}})(\bar{y}_{i\cdot} - \bar{\bar{y}}).$$

Interrelationships:
$\mathbf{S}_{XX}^{OA} = \mathbf{S}_{XX}^{WG} + \mathbf{S}_{XX}^{BG}$and $\mathbf{S}_{Xy}^{OA} = \mathbf{S}_{Xy}^{WG} + \mathbf{S}_{Xy}^{BG}$.

(c) Estimators: In terms of estimators for β, we have

$$\mathbf{b}^{OA} = (\mathbf{S}_{XX}^{OA})^{-1}\mathbf{S}_{Xy}^{OA},\ \mathbf{b}^{WG} = (\mathbf{S}_{XX}^{WG})^{-1}\mathbf{S}_{Xy}^{WG},\ \text{and}\ \mathbf{b}^{BG} = (\mathbf{S}_{XX}^{BG})^{-1}\mathbf{S}_{Xy}^{BG}.$$

Interrelaionships:

$$\mathbf{b}^{OA} = \mathbf{F}^{WG}\mathbf{b}^{WG} + \mathbf{F}^{BG}\mathbf{b}^{BG},\ \text{where}\ \mathbf{F}^{WG} = (\mathbf{S}_{XX}^{WG} + \mathbf{S}_{XX}^{BG})^{-1}\mathbf{S}_{XX}^{WG} = \mathbf{I} - \mathbf{F}^{BG}.$$

3. Random Effects

The regression model takes the form $y_{it} = \alpha + \beta'\mathbf{x}_{it} + u_i + \varepsilon_{it}$, where u_i and ε_{it} are independent disturbances with zero mean and variances σ_u^2 and σ_ε^2. It is further assumed that there is no autocorrelation inside ε.

Consider the combined disturbance $\eta_{it} = u_i + \varepsilon_{it}$, we have $\Omega_i = E(\eta\eta') = \sigma_u^2\mathbf{i}\mathbf{i}' + \sigma_\varepsilon^2\mathbf{I}$ and

$$\Sigma = \begin{bmatrix} \Omega_1 & \mathbf{0} & \vdots & \mathbf{0} \\ \mathbf{0} & \Omega_2 & \vdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \vdots & \Omega_n \end{bmatrix}.$$

4. Preparation for Factor Analysis

(1) Multivariate normal density

Let $\mathbf{Y} \sim N(\mu_Y, \Sigma_{YY})$ with $\mathbf{Y}_{N\times K}$, then the individual likelihood is

$$f(\mathbf{Y}_i|\mu_{Y_i}, \Sigma_{YY}) = \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma_{YY}|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}(\mathbf{Y}_i - \mu_{Y_i})'\Sigma_{YY}^{-1}(\mathbf{Y}_i - \mu_{Y_i})],$$

and the sample log-likelihood is

$$\begin{aligned} \ln \mathcal{L} &= c - \tfrac{N}{2} \ln |\Sigma_{YY}| - \tfrac{1}{2} \sum\nolimits_{i=1}^{N} (\mathbf{Y}_i - \mu_{Y_i})'\Sigma_{YY}^{-1}(\mathbf{Y}_i - \mu_{Y_i}) \\ &= c - \tfrac{N}{2} \ln |\Sigma_{YY}| - \tfrac{N}{2} tr[\Sigma_{YY}^{-1}\hat{\mathbf{S}}_{YY}], \end{aligned}$$

which is the so-called "Wishart covariance structure."

Note that the representation of sample log-likelihood is valid only if Σ_{YY} is the same for all $\mathbf{Y}_i$'s. If we do allow difference across $\mathbf{Y}_i$'s, then we need use the MVN density function. For example, the MVN density function for the model $\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{u}_i$, or $\mathbf{Y} = (\mathbf{y}_1', ..., \mathbf{y}_N')'$, is

$$\begin{aligned} f(\mathbf{y}_i|\mathbf{0}, \Sigma_{YY}) &= \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma_{YY}|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}\mathbf{y}_i'\Sigma_{YY}^{-1}\mathbf{y}_i] \\ &= \frac{1}{(2\pi)^{\frac{K_{y_i}}{2}} |\Sigma_{uu}|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\beta)'\Sigma_{uu}^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta)] \\ &\quad \left\{ \frac{1}{(2\pi)^{\frac{K_X}{2}} |\Sigma_{XX}|^{\frac{1}{2}}} \exp[-\tfrac{1}{2}\mathbf{X}_i'\Sigma_{XX}^{-1}\mathbf{X}_i] \right\}. \end{aligned}$$

Note that the part before the curly bracket is the sufficient statistic for the density function, and in practice we maximize the sample likelihood using only the sufficient statistic part. That is,

$$\ln \mathcal{L} = c - \tfrac{N}{2} \ln \Sigma_{uu} - \tfrac{1}{2} tr[\Sigma_{uu}^{-1}\hat{\mathbf{S}}_{uu}].$$

(2) A single equation system

Assume $\mathbf{y} = \beta'\mathbf{x} + \mathbf{u}$. Let's set up the model in terms of $\mathbf{Y}$, where $\mathbf{Y} = (\mathbf{y}'\ \mathbf{x}')'$, as the following, $Y = \begin{pmatrix} \beta'\mathbf{x}+\mathbf{u} \\ \mathbf{x} \end{pmatrix}$. The parameter vector is $\theta = (\beta\ \Sigma_{xx}\ \Sigma_{uu})$. Clearly we have the following system of equations $\hat{\Sigma}(\theta) = \mathbf{S}_{YY}$, where

$$\Sigma(\theta) = \begin{bmatrix} \beta'\Sigma_{xx}\beta + \Sigma_{uu} & \beta'\Sigma_{xx} \\ \Sigma_{xx}\beta & \Sigma_{xx} \end{bmatrix}, \text{ and } \mathbf{S}_{YY} = \begin{bmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{bmatrix}.$$

We have an exactly identified system, one solution of which is:

$$\hat{\Sigma}_{xx} = \mathbf{S}_{xx}, \hat{\beta} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}, \hat{\Sigma}_{uu} = \mathbf{S}_{yy} - \mathbf{S}_{xy}'\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}.$$

Note that we stack $\mathbf{x}$ below $\mathbf{y}$ in constructing $\mathbf{Y}$ because we need the estimators for β. In the sections below where we consider only the variance

component, we won't stack $\mathbf{x}$ below $\mathbf{y}$ again since there is no need to estimate β.

(3) A simple variance component

Assume $y_{it} = \delta_i + u_{it}, t = 1, ..., T$ with $\delta_i \sim N(0, \sigma_\delta^2)$ and $u_{it} \sim$ i.i.d. $N(0, \sigma_u^2)$. In matrix notation, the T equation system is $\mathbf{y}_i = \delta_i \mathbf{i} + \mathbf{u}_i$. The parameter vector is $\theta = (\sigma_\delta^2 \ \sigma_u^2)$, and we hypothesize that $T = 2$ is enough to identify both parameters.

When $T = 2$, we have the following system of equations, $\hat{\Sigma}(\theta) = \mathbf{S_{yy}}$, where

$$\Sigma(\theta) = \sigma_\delta^2 \mathbf{i}_T \mathbf{i}_T{}' + \sigma_u^2 \mathbf{I}_T = \begin{bmatrix} \sigma_\delta^2 + \sigma_u^2 & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma_u^2 \end{bmatrix}.$$

In this case, we have an identified system, one solution of which is $\hat{\sigma}_\delta^2 = \mathbf{S}_{21}, \hat{\sigma}_u^2 = \mathbf{S}_{11} - \mathbf{S}_{21}$. The sample likelihood for individual i can be written as one of following two representations:

$$\begin{aligned} \mathcal{L}_i &= \frac{1}{(2\pi)^{\frac{2}{2}} |\Sigma_{\mathbf{yy}}|^{\frac{1}{2}}} \exp(-\tfrac{1}{2} \mathbf{y}' \Sigma_{\mathbf{yy}} \mathbf{y}) \\ &= \int_{\delta_i} \tfrac{1}{\sqrt{2\pi}\sigma_\delta} \exp[-\tfrac{1}{2}(\tfrac{\delta}{\sigma_\delta})^2] \prod_{t=1}^{T} \tfrac{1}{\sqrt{2\pi}\sigma_u} \exp[-\tfrac{1}{2}(\tfrac{y_{it} - \delta_i}{\sigma_u})^2] d\delta_i. \end{aligned}$$

(4) Variance component plus $AR(1)$

The model is $y_{it} = \delta_i + u_{it}, t = 1, ..., T$, $u_{it} = \gamma u_{it-1} + \eta_{it}$, with $\delta_i \sim N(0, \sigma_\delta^2)$ and $\eta_{it} \sim$ i.i.d. $N(0, \sigma_\eta^2)$. In matrix notation, the T equation system is $\mathbf{y}_i = \delta_i \mathbf{i} + \mathbf{u}_i$. The parameter vector is $\theta = (\gamma \ \sigma_\delta^2 \ \sigma_\eta^2)$ and we hypothesis that $T = 3$ is enough to identify θ.

Apparently, we have $Var(u_{it}) = \frac{\sigma_\eta^2}{1-\gamma^2}$ and the correlation matrix for $\mathbf{u}_i$ is

$$\mathbf{A} = \begin{bmatrix} 1 & \gamma & \gamma^2 \\ \gamma & 1 & \gamma \\ \gamma^2 & \gamma & 1 \end{bmatrix}.$$

We also have

$$\begin{aligned} \Sigma(\theta) &= \sigma_\delta^2 \mathbf{i}_T \mathbf{i}_T{}' + \tfrac{\sigma_\eta^2}{1-\gamma^2} \mathbf{A} \\ &= \begin{bmatrix} \sigma_\delta^2 + \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \gamma \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \gamma^2 \frac{\sigma_\eta^2}{1-\gamma^2} \\ \sigma_\delta^2 + \gamma \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \gamma \frac{\sigma_\eta^2}{1-\gamma^2} \\ \sigma_\delta^2 + \gamma^2 \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \gamma \frac{\sigma_\eta^2}{1-\gamma^2} & \sigma_\delta^2 + \frac{\sigma_\eta^2}{1-\gamma^2} \end{bmatrix}. \end{aligned}$$

$\hat{\Sigma}(\theta) = \mathbf{S}_{YY}$ implies that we have an identified system, one solution of which is:

$$\hat{\gamma} = \tfrac{\mathbf{S}_{31} - \mathbf{S}_{11}}{\mathbf{S}_{21} - \mathbf{S}_{11}} - 1, \hat{\sigma}_\eta^2 = (\mathbf{S}_{11} - \mathbf{S}_{21})(1 + \hat{\gamma}), \hat{\sigma}_\delta^2 = \mathbf{S}_{11} - \tfrac{\hat{\sigma}_\eta^2}{1-\hat{\gamma}^2}.$$

(5) Variance component plus $MA(1)$

The model is $y_{it} = \delta_i + u_{it}, t = 1, ..., T$, $u_{it} = \eta_{it} - \gamma \eta_{it-1}$, with $\delta_i \sim N(0, \sigma_\delta^2)$ and $\eta_{it} \sim$ i.i.d. $N(0, \sigma_\eta^2)$. In matrix notation, the T equation system is $\mathbf{y}_i = \delta_i \mathbf{i} + \mathbf{u}_i$. The parameter vector is $\theta = (\gamma \ \sigma_\delta^2 \ \sigma_\eta^2)$ and we

hypothesis that $T = 3$ is enough to identify θ. Apparently, we have $Var(u_{it}) = \frac{\sigma_\eta^2}{1+\gamma^2}$ and the correlation matrix for $\mathbf{u}_i$ is

$$\mathbf{A} = \begin{bmatrix} 1 & -\gamma & 0 \\ -\gamma & 1 & -\gamma \\ 0 & -\gamma & 1 \end{bmatrix}.$$

We also have

$$\begin{aligned}\Sigma(\theta) &= \sigma_\delta^2 \mathbf{i}_T \mathbf{i}_T{}' + (1+\gamma^2)\mathbf{A} \\ &= \begin{bmatrix} \sigma_\delta^2 + (1+\gamma^2)\sigma_\eta^2 & \sigma_\delta^2 - \gamma\sigma_\eta^2 & \sigma_\delta^2 \\ \sigma_\delta^2 - \gamma\sigma_\eta^2 & \sigma_\delta^2 + (1+\gamma^2)\sigma_\eta^2 & \sigma_\delta^2 - \gamma\sigma_\eta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 - \gamma\sigma_\eta^2 & \sigma_\delta^2 + (1+\gamma^2)\sigma_\eta^2 \end{bmatrix}.\end{aligned}$$

$\hat{\Sigma}(\theta) = \mathbf{S}_{YY}$ implies that we have an identified system.

5. Regression Based Factor Analysis

(1) Model with K indicators λ_k and one factor δ

The model is $y_{ik} = \lambda_k \delta_i + u_{ik}$, with $\delta_i \sim N(0, \sigma_\delta^2)$ and $\mathbf{u}_i \sim N(\mathbf{0}, \Omega)$, where Ω is a diagonal covariance matrix for $\mathbf{u}_i$ with heterogeneity. In matrix notation, the K equation system is $\mathbf{y}_i = \lambda\delta_i + \mathbf{u}_i$. Let's normalize $\lambda_1 = 1$ and consider the case where $K = 3$. The parameter vector is $\theta = (\lambda_2\ \lambda_3\ \sigma_\delta^2\ \sigma_{u_1}^2\ \sigma_{u_2}^2\ \sigma_{u_3}^2)$.

We also have

$$\Sigma(\theta) = \sigma_\delta^2 \lambda\lambda' + \Omega = \begin{bmatrix} \sigma_\delta^2 + \sigma_{u_1}^2 & 1\cdot\lambda_2\sigma_\delta^2 & 1\cdot\lambda_3\sigma_\delta^2 \\ 1\cdot\lambda_2\sigma_\delta^2 & \sigma_\delta^2 + \sigma_{u_2}^2 & \lambda_2\lambda_3\sigma_\delta^2 \\ 1\cdot\lambda_3\sigma_\delta^2 & \lambda_2\lambda_3\sigma_\delta^2 & \sigma_\delta^2 + \sigma_{u_3}^2 \end{bmatrix}.$$

$\hat{\Sigma}(\theta) = \mathbf{S}_{YY}$implies that we have an identified system, one solution of which is: $\hat{\lambda}_2 = \frac{\mathbf{S}_{32}}{\mathbf{S}_{31}}, \hat{\lambda}_3 = \frac{\mathbf{S}_{32}}{\mathbf{S}_{21}}, \hat{\sigma}_\delta^2 = \frac{\mathbf{S}_{31}\mathbf{S}_{21}}{\mathbf{S}_{32}}, \hat{\sigma}_{u_1}^2 = \mathbf{S}_{11} - \hat{\sigma}_\delta^2, \hat{\sigma}_{u_2}^2 = \mathbf{S}_{22} - \hat{\lambda}_2^2\hat{\sigma}_\delta^2, \hat{\sigma}_{u_3}^2 = \mathbf{S}_{33} - \hat{\lambda}_3^2\hat{\sigma}_\delta^2$.

(2) Multiple indicator multiple cause (MIMC) model with K indicators λ_k and one factor F with L regressors $\mathbf{X}$.

The model is $y_{ik} = \lambda_k F_i + u_{ik}$ and $F_i = \beta'\mathbf{x}_i + \delta_i$, with $\delta_i \sim N(0, \sigma_\delta^2)$ and $\mathbf{u}_i \sim N(\mathbf{0}, \Omega)$, where Ω is a diagonal covariance matrix for $\mathbf{u}_i$ with heterogeneity. Note that $\mathbf{x}_i$ is of L-dimension. The model above can be rewritten as $y_{ik} = \lambda_k\beta'\mathbf{x}_i + \lambda_k\delta_i + u_{ik}$, or in matrix notation, $\mathbf{y}_i = \lambda\beta'\mathbf{x}_i + \lambda\delta_i + \mathbf{u}_i$. Let's normalize $\lambda_1 = 1$ and consider the case when $K = 3$. The parameter vector is $\theta = (\lambda_2\ \lambda_3\ \sigma_\delta^2\ \sigma_{u_1}^2\ \sigma_{u_2}^2\ \sigma_{u_3}^2)$.

If we define $\sigma_F^2 \equiv \beta'\Sigma_{xx}\beta + \sigma_u^2$, we also have

$$\begin{aligned}\Sigma(\theta) &= \lambda\beta'\Sigma_{xx}\beta\lambda' + \lambda\sigma_u^2\lambda' + \Omega \\ &= \lambda\sigma_F^2\lambda' + \Omega \\ &= \begin{bmatrix} \sigma_F^2 + \sigma_{u_1}^2 & 1\cdot\lambda_2\sigma_\delta^2 & 1\cdot\lambda_3\sigma_\delta^2 \\ 1\cdot\lambda_2\sigma_\delta^2 & \sigma_F^2 + \sigma_{u_2}^2 & \lambda_2\lambda_3\sigma_\delta^2 \\ 1\cdot\lambda_3\sigma_\delta^2 & \lambda_2\lambda_3\sigma_\delta^2 & \sigma_F^2 + \sigma_{u_3}^2 \end{bmatrix}.\end{aligned}$$

$\hat{\Sigma}(\theta) = \mathbf{S}_{YY}$implies that we have an identified system.

CHAPTER 12

Simultaneous Equations Models

1. Simultaneous Equations Model with a Single Observation

(1) Original Model

$By_t + \Gamma x_t = \varepsilon_t$ where

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{G1} & \beta_{G2} & \cdots & \beta_{GG} \end{pmatrix}, \Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{G1} & \gamma_{G2} & \cdots & \gamma_{GK} \end{pmatrix},$$

$$y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \\ \cdots \\ y_{Gt} \end{pmatrix}, x_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ \cdots \\ x_{Gt} \end{pmatrix},$$

with

$$\begin{aligned} E(\varepsilon_{gt}) &= 0, \\ E(\varepsilon_{gt}\varepsilon_{gs}) &= 0 \text{ for } t \neq s, \\ E(\varepsilon_{gt}^2) &= \sigma_{gg}, \\ E(\varepsilon_{gt}\varepsilon_{ht}) &= \sigma_{gh} \text{ for } g, h = 1, ..., G. \end{aligned}$$

$$E(\varepsilon_t \varepsilon_t') = \sum = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1G} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2G} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{G1} & \sigma_{G2} & \cdots & \sigma_{GG} \end{pmatrix}.$$

(2) Reduced Form

$$By_t + \Gamma x_t = \varepsilon_t \Rightarrow y_t = -B^{-1}\Gamma x_t + B^{-1}\varepsilon_t \Rightarrow y_t = \Pi x_t + v_t, \text{ where}$$

$$\begin{aligned} \Pi &= -B^{-1}\Gamma, \\ v_t &= B^{-1}\varepsilon_t, \\ \Pi &= \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1G} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2G} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{G1} & \pi_{G2} & \cdots & \pi_{GG} \end{pmatrix}, \\ E(v_t v_t') &= B^{-1} \sum (B^{-1})' = \Omega. \end{aligned}$$

2. Simultaneous Equations Model of Full Observations

(1) Original Model: $\mathbf{BY}+\Gamma\mathbf{X}=\varepsilon$
(2) Reduced Form: $\mathbf{BY}+\Gamma\mathbf{X}=\varepsilon \Rightarrow \mathbf{Y}=-\mathbf{B}^{-1}\Gamma\mathbf{X}+\mathbf{B}^{-1}\varepsilon \Rightarrow \mathbf{Y}=\Pi\mathbf{X}+\mathbf{V}$ where $\Pi=-\mathbf{B}^{-1}\Gamma$ and Π are the same as that in the SEM of a single observation, $\mathbf{V}=\mathbf{B}^{-1}\varepsilon$. $p\lim\left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)=Q, p\lim\left(\frac{\mathbf{X}'\varepsilon}{T}\right)=0, p\lim\left(\frac{\varepsilon'\varepsilon}{T}\right)=\sum$ and $p\lim\left(\frac{\mathbf{V}'\mathbf{V}}{T}\right)=\mathbf{B}^{-1\prime}\Sigma\mathbf{B}^{-1}=\Omega, p\lim\left(\frac{\mathbf{X}'\mathbf{V}}{T}\right)=\mathbf{0}$.

3. Identification Conditions

(1) Order Condition (Necessary Condition)
Let R be the number of restrictions within the considered equation, the order condition is $R \geqslant G-1$, since we normalize one of the coefficients associated with endogenous variables to be -1.
(2) Rank Condition (Necessary and Sufficient Condition)
Given a considered equation, find the restricted parameters within that equation, select all columns containing those restricted parameters and form a matrix, the rank of this matrix should be equal to $G-1$.

4. Example #1

	y_{1t}	y_{2t}	y_{3t}	z_{1t}	z_{2t}	z_{3t}
Eq. 1	-1	β_{12}	β_{13}	γ_{11}	γ_{12}	γ_{13}
Eq. 2	β_{21}	-1	0	γ_{21}	0	γ_{23}
Eq. 3	β_{31}	0	-1	0	γ_{32}	0

For Eq. 1: $R=0, G-1=2$ It fails the order condition, thus is not identified.

For Eq. 2: $R=2, G-1=2$ It passes the order condition. The selected matrix is

$$\begin{pmatrix} \beta_{13} & \gamma_{12} \\ 0 & 0 \\ -1 & \gamma_{32} \end{pmatrix},$$

which has rank 2, then it also passes the rank condition. Thus Eq. 2 is just identified.

For Eq. 3: $R=3, G-1=2$ It passes the order condition. The selected matrix is

$$\begin{pmatrix} \beta_{12} & \gamma_{11} & \gamma_{13} \\ -1 & \gamma_{21} & \gamma_{23} \\ 0 & 0 & 0 \end{pmatrix},$$

which has rank 2, then it also passes the rank condition. Thus Eq. 3 is potentially over identified by 1.

5. Johnston's Approach

$\Pi=-B^{-1}\Gamma$implies $B\Pi+\Gamma=0$, i.e., $\begin{pmatrix} B & \Gamma \end{pmatrix}\begin{pmatrix} \Pi \\ I \end{pmatrix}=0$.

Let $A\equiv\begin{pmatrix} B & \Gamma \end{pmatrix}, W\equiv\begin{pmatrix} \Pi \\ I \end{pmatrix}$, then $AW=0$.

Design Φ_i as $(G+K)\times R$ to represent the restrictions imposed on the i^{th} equation. (Not like the design matrix R in Chapter 6, where we have one row for each restriction, here we set one column for each restriction.)

Let α_i be the ith row of A, then the ith equation implies $\alpha_i \begin{pmatrix} \Phi_i & W \end{pmatrix} = \begin{pmatrix} O_R & O_K \end{pmatrix}$, where $\alpha_i\Phi_i = O_R$ are the set of R restrictions, and$\alpha_i W = O_K$ are K equations coming from $AW = 0$. In α_i, the parameter associated with the normalized endogenous variable is set to be –1. We have $R + K$ equations and $G + K - 1$ unknowns.

Order condition is $G + K - 1 \leqslant R + K$, i.e., $R \geqslant G - 1$.

Rank condition is $Rank \begin{pmatrix} \Phi_i & W \end{pmatrix} = G + K - 1$ or $Rank\,(A\Phi_i) = G - 1$.

[Essentially, $A\Phi_i$ is the selected matrix in 3.(2)]

6. Kmenta's Approach

$B\Pi + \Gamma = 0 \Rightarrow \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \end{pmatrix} \Pi = -\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \end{pmatrix}$,i.e., $\beta_1\Pi = -\gamma_1$ for the first equation.

Let $^{\Delta}$ stand for endogenous variable, and * stand for exogenous variable.

$G^{\Delta}$is the number of included endogenous variables;

$G^{\Delta\Delta}$is the number of excluded endogenous variables;

$K^*$is the number of included exogenous variables;

$K^{**}$is the number of excluded exogenous variables.

Partition matrices properly as following:

$$\beta_1 = \begin{pmatrix} \beta^{\Delta} & O^{\Delta\Delta} \end{pmatrix}, \quad \gamma_1 = \begin{pmatrix} \gamma^* & O^{**} \end{pmatrix}, \quad \Pi = \begin{pmatrix} \Pi^{\Delta *} & \Pi^{\Delta **} \\ \Pi^{\Delta\Delta *} & \Pi^{\Delta\Delta **} \end{pmatrix}.$$

Thus $\beta_1\Pi = -\gamma_1$ implies $\begin{pmatrix} \beta^{\Delta} & O^{\Delta\Delta} \end{pmatrix} \begin{pmatrix} \Pi^{\Delta *} & \Pi^{\Delta **} \\ \Pi^{\Delta\Delta *} & \Pi^{\Delta\Delta **} \end{pmatrix} = -\begin{pmatrix} \gamma^* & O^{**} \end{pmatrix}$, i.e., $\beta^{\Delta}\Pi^{\Delta *} = -\gamma^*$ and $\beta^{\Delta}\Pi^{\Delta **} = O^{**}$.

From $\beta^{\Delta}\Pi^{\Delta **} = O^{**}$, we know that we have K^{**} equations and $G^{\Delta} - 1$ unknowns.

Order Condition: $K^{**} \geqslant G^{\Delta} - 1$, i.e., $G^{\Delta} - 1$ out of K^{**} equations must be independent.

Rank Condition: $Rank(\Pi^{\Delta **}) = G^{\Delta} - 1$.

After finding β from this system of equations, we can use $\beta^{\Delta}\Pi^{\Delta *} = \gamma^*$ to find γ correspondingly.

7. Indirect Least Square

Step 1: Use either Johnston's approach or Kmenta's approach to solve for β and γ in terms of Π_{ij};

Step 2: Run OLS regressions on the system of reduced form equations and get $\hat{\Pi}_{ij}$;

Step 3: Replace Π_{ij} with $\hat{\Pi}_{ij}$ to get $\hat{\beta}$ and $\hat{\gamma}$. By Slutsky's Theorem, we know both $\hat{\beta}$ and $\hat{\gamma}$ are consistent estimators.

8. Two Stage Least Square (TSLS)

$YB' + X\Gamma' = \varepsilon \Rightarrow Y = X\Pi' + V$. In particular, the first equation is: $y_1 = Y_1\beta_1' + X_1\gamma_1' + \varepsilon_1$, where y_1 is the normalized endogenous variable in the first equation, Y_1 is the included endogenous variables except the normalized one in the first reduced form equation, X_1 is the included exogenous variables in the first equation.

From $Y = X\Pi' + V$, we also have

$$\begin{pmatrix} y_1 & Y_1 & Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \Pi_{11}' & \Pi_{12}' & \Pi_{13}' \\ \Pi_{21}' & \Pi_{22}' & \Pi_{23}' \end{pmatrix} + \begin{pmatrix} \nu_1 & V_1 & V_2 \end{pmatrix},$$

where Y_2 is the excluded endogenous variables in the first equation; X_2 is the excluded endogenous variables in the first equation; Π' is the blockwise coefficients matrix, in which the second row of blocks is for excluded exogenous variables and the first row is for included exogenous variables. Also note that the first column of blocks consists of parameters in the reduced form for y_1. Similarly, the second and third columns are for $Y_1$and Y_2, respectively.

Essentially, we have

$$\begin{pmatrix} y_1 & Y_1 & Y_2 & X_1 & X_2 \end{pmatrix} \begin{pmatrix} 1 \\ -\beta_1' \\ 0 \\ -\gamma_1' \\ 0 \end{pmatrix} = \varepsilon_1.$$

(1) Let $Z_1 = \begin{pmatrix} Y_1 & X_1 \end{pmatrix}$ and $\alpha_1' = \begin{pmatrix} \beta_1' \\ \gamma_1' \end{pmatrix}$, then $y_1 = Y_1\beta_1' + X_1\gamma_1' + \varepsilon_1$ implies $y_1 = Z_1\alpha_1' + \varepsilon_1$.

If we use OLS, $\hat{\alpha}_{1,OLS}' = (Z_1'Z_1)^{-1}Z_1'y_1 = \alpha_1' + (Z_1'Z_1)^{-1}Z_1'\varepsilon_1$.

After painful work, we find $p\lim\left(\frac{Z_1'\varepsilon_1}{T}\right) \neq 0$, thus$\hat{\alpha}_1'$ is inconsistent for α_1'.

Actually, we would expect this result from the fact that part of the regressors Z_1, namely Y_1, is stochastic, which violates our classical assumption of "non-stochastic regressors".

(2) From

$$\begin{pmatrix} y_1 & Y_1 & Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \Pi_{11}' & \Pi_{12}' & \Pi_{13}' \\ \Pi_{21}' & \Pi_{22}' & \Pi_{23}' \end{pmatrix} + \begin{pmatrix} \nu_1 & V_1 & V_2 \end{pmatrix},$$

we know $Y_1 = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \Pi_{12}' \\ \Pi_{22}' \end{pmatrix} + V_1 = X\Pi_2' + V_1$. Thus by OLS, we have $\hat{\Pi}_2' = (X'X)^{-1}X'Y_1$, $\hat{V}_1 = Y_1 - X\hat{\Pi}_2'$, and $X'\hat{V}_1 = 0$. Substitute $Y_1 = X\Pi_2' + V_1$ back into $y_1 = Y_1\beta_1' + X_1\gamma_1' + \varepsilon_1$, we get $y_1 = X\Pi_2'\beta_1' + X_1\gamma_1' + \varepsilon_1 + V_1\beta_1'$. Using the estimators $\hat{\Pi}_2$ and $\hat{V}_1$, and the fact that $X\hat{\Pi}_2' = \hat{Y}_1$, we get $y_1 = \hat{Y}_1\beta_1' + X_1\gamma_1' + (\varepsilon_1 + \hat{V}_1\beta_1')$. Let $\hat{Z}_1 = \begin{pmatrix} \hat{Y}_1 & X_1 \end{pmatrix}$, then we can use $y_1 = \hat{Z}_1\alpha_1' + \varepsilon_1 + \hat{V}_1\beta_1'$ to get $\hat{\hat{\alpha}}_{1,TSLS}' = (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'y_1 = \alpha_1' + (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'\varepsilon_1 + (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'\hat{V}_1\beta_1'$. Since $X'\hat{V}_1 = 0$, $\hat{Z}_1'\hat{V}_1 = \begin{pmatrix} \hat{Y}_1' \\ X_1' \end{pmatrix}\hat{V}_1 = \begin{pmatrix} \hat{\Pi}_2X' \\ X_1' \end{pmatrix}\hat{V}_1 = 0$. Thus $\hat{\hat{\alpha}}_{1,TSLS}' = (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'y_1 = \alpha_1' + (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'\varepsilon_1$. Since $p\lim\left(\frac{\hat{Z}_1'\varepsilon_1}{T}\right) = p\lim\begin{pmatrix} \frac{\hat{\Pi}_2X'\varepsilon_1}{T} \\ \frac{X_1'\varepsilon_1}{T} \end{pmatrix} = 0$, we know that $\hat{\hat{\alpha}}_1'$ is consistent for $\hat{\alpha}_1'$. $Var\left(\hat{\hat{\alpha}}_{1,TSLS}'\right) = \sigma_1^2(\hat{Z}_1'\hat{Z}_1)^{-1} = \sigma_1^2\begin{pmatrix} \hat{Y}_1'\hat{Y}_1 & \hat{Y}_1'X_1 \\ X_1'\hat{Y}_1 & X_1'X_1 \end{pmatrix}$.

(3) Procedures of TSLS:

Stage 1: a. run OLS regression on the reduced form equations $Y_1 = X\Pi_2' + V_1$;
b. get $\hat{\Pi}_2$, $\hat{V}_1$ and $\hat{Y}_1$;
c. get $\hat{Z}_1 = \begin{pmatrix} \hat{Y}_1 & X_1 \end{pmatrix}$.

Stage 2: a. run OLS regression on the structural equation $y_1 = \hat{Z}_1\alpha_1' + \varepsilon_1$;
b. get consistent estimators $\hat{\alpha}_{1,TSLS}' = (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'y_1$.

Note: From $y_1 = \hat{Y}_1\beta_1' + X_1\gamma_1' + (\varepsilon_1 + \hat{V}_1\beta_1')$, we know $y_1 - (\hat{Y}_1\hat{\beta}_1' + X_1\hat{\gamma}_1')$ is not the correct residual associated with the disturbance term for the first equation ε_1, and the correct one should be $y_1 - (\hat{Y}_1\hat{\beta}_1' + X_1\hat{\gamma}_1' + \hat{V}_1\hat{\beta}_1') = y_1 - [(\hat{Y}_1 + \hat{V}_1)\hat{\beta}_1' + X_1\hat{\gamma}_1'] = y_1 - (Y_1\hat{\beta}_1' + X_1\hat{\gamma}_1')$.

9. Example #2

	y_{1t}	y_{2t}	1	x_{2t}	x_{3t}	x_{4t}
Eq. 1	-1	β_{12}	γ_{11}	γ_{12}	0	γ_{14}
Eq. 2	β_{21}	-1	γ_{21}	0	γ_{23}	0

Reduced forms:

$$y_{1t} = \Pi_{11} + \Pi_{12}x_{2t} + \Pi_{13}x_{3t} + \Pi_{14}x_{4t} + \nu_{1t}$$
$$y_{2t} = \Pi_{21} + \Pi_{22}x_{2t} + \Pi_{23}x_{3t} + \Pi_{24}x_{4t} + \nu_{2t}$$

Stage 1: run OLS on the reduced form equation for Y_1, here $Y_1 = y_{2t}$, and save the fitted values $\hat{Y}_1 = \hat{y}_{2t}$;

Stage 2: run OLS on the structural equation for y_1, and replace y_{2t} with $\hat{y}_{2t}$, i.e., $y_{1t} = \beta_{12}\hat{y}_{2t} + \gamma_{11} + \gamma_{12}x_{2t} + \gamma_{14}x_{4t} + \varepsilon_{1t}$.

Note: When we are doing TSLS like this, we will get a wrong residual vector, which is much less than the correct one. (How do we know it is much less? It should depend upon the sign of β_{12}.) Since when we run regression $y_{1t} = \beta_{12}\hat{y}_{2t} + \gamma_{11} + \gamma_{12}x_{2t} + \gamma_{14}x_{4t} + \varepsilon_{1t}$, we didn't do any operation to the residual term, which ends up with $y_{1t} - (\hat{\beta}_{12}\hat{y}_{2t} + \hat{\gamma}_{11} + \hat{\gamma}_{12}x_{2t} + \hat{\gamma}_{14}x_{4t})$, whereas the correct residual term is $y_{1t} - (\hat{\beta}_{12}y_{2t} + \hat{\gamma}_{11} + \hat{\gamma}_{12}x_{2t} + \hat{\gamma}_{14}x_{4t})$.

10. Instrumental Variable (IV) Approach

$$y_1 = Y_1\beta_1' + X_1\gamma_1' + \varepsilon_1 = Z_1\alpha_1' + \varepsilon_1$$

Our problem is that $p\lim\left(\frac{Z_1'\varepsilon_1}{T}\right) \neq 0$ so that the OLS estimators are not consistent. As the instrumental variable approach indicates, we need to find a proper proxy W for Z_1, which satisfies $p\lim\left(\frac{W'W}{T}\right) = \sum_{WW}$, $p\lim\left(\frac{W'Z_1}{T}\right) = \sum_{WZ_1}$, $p\lim\left(\frac{W'\varepsilon_1}{T}\right) = 0$.

We find that $W = \hat{Z}_1 = (\hat{Y}_1 \;\; X_1)$ satisfies such a need. Thus we replace $\hat{\alpha}_{1,OLS}' = (Z_1'Z_1)^{-1}Z_1'y_1$ with $\hat{\alpha}_{1,IV}' = (\hat{Z}_1'Z_1)^{-1}\hat{Z}_1'y_1$. Note that the TSLS estimators are $\hat{\alpha}_{1,TSLS}' = (\hat{Z}_1'\hat{Z}_1)^{-1}\hat{Z}_1'y_1$, and we can prove $\hat{Z}_1'\hat{Z}_1 = \hat{Z}_1'Z_1$ so that $\hat{\alpha}_{1,IV}' = \hat{\alpha}_{1,TSLS}'$. The proof is the following:

Since $\hat{Z}_1'Z_1 = \begin{pmatrix} \hat{Y}_1'Y_1 & \hat{Y}_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{pmatrix}$ and $\hat{Z}_1'\hat{Z}_1 = \begin{pmatrix} \hat{Y}_1'\hat{Y}_1 & \hat{Y}_1'X_1 \\ X_1'\hat{Y}_1 & X_1'X_1 \end{pmatrix}$, it suffices to prove $\hat{Y}_1'\hat{Y}_1 = \hat{Y}_1'Y_1$ and $\hat{Y}_1'X_1 = Y_1'X_1$, which follow the fact that :

$$\hat{Y}_1'\hat{Y}_1 = \hat{Y}_1'(Y_1 - \hat{V}_1) = \hat{Y}_1'Y_1 - \hat{\Pi}_2 X'\hat{V}_1 = \hat{Y}_1'Y_1, (X'\hat{V}_1 = 0).$$
$$\hat{Y}_1'X_1 = (Y_1 - \hat{V}_1)'X_1 = Y_1'X_1 - \hat{V}_1'X_1 = Y_1'X_1.$$
$$Var\left(\hat{\alpha}_{1,IV}'\right) = \sigma_1^2(\hat{Z}_1'Z_1)^{-1} = \sigma_1^2 \begin{pmatrix} \hat{Y}_1'Y_1 & \hat{Y}_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{pmatrix}.$$

11. Aitken's Approach (Given by Dhrymes)

$$y_1 = Y_1\beta_1' + X_1\gamma_1' + \varepsilon_1 \Rightarrow X'y_1 = X'Y_1\beta_1' + X'X_1\gamma_1' + X'\varepsilon_1$$

Assume $p\lim\left(\frac{X'y_1}{T}\right)$, $p\lim\left(\frac{X'Y_1}{T}\right)$, and $p\lim\left(\frac{X'X_1}{T}\right)$ are well behaved, and $p\lim\left(\frac{X'\varepsilon_1}{T}\right) = 0$.

We get $Var(X'\varepsilon_1) = \sigma_1^2(X'X)$.

Invent P such that $P'X'XP = I$ and $PP' = (X'X)^{-1}$, then we have $P'X'y_1 = P'X'Y_1\beta_1' + PX'X_1\gamma_1' + P'X'\varepsilon_1$, with $Var(P'X'\varepsilon_1) = \sigma_1^2 I_K$.

Define $P'X'y_1 \equiv W_1, \quad P'X'\varepsilon_1 \equiv r_1, \quad R_1 \equiv \begin{pmatrix} P'X'Y_1 & P'X'X_1 \end{pmatrix} = P'X'Z_1$.

$W_1 = R_1\alpha_1' + r_1$implies that $\hat{\alpha}_{1,A}' = (R_1'R_1)^{-1}R_1'W_1 = \alpha_1' + (R_1'R_1)^{-1}R_1'r_1$.

We also can prove that $\hat{\alpha}_{1,A}' = \hat{\hat{\alpha}}_{1,TSLS}' = \hat{\alpha}_{1,IV}'$.

We have $R_1 = P'X'Z_1, \quad W_1 = P'X'y_1, \quad P'X'XP = I, \quad PP' = (X'X)^{-1}$, thus $R_1'R_1 = (P'X'Z_1)'(P'X'Z_1) = Z_1'XPP'X'Z_1 = Z_1'X(X'X)^{-1}X'Z_1$, $R_1'W_1 = Z_1'XPP'X'y_1 = Z_1'X(X'X)^{-1}X'y_1$

Then $\hat{\alpha}_{1,A}' = \hat{\hat{\alpha}}_{1,TSLS}' = \hat{\alpha}_{1,IV}' = [Z_1'X(X'X)^{-1}X'Z_1]^{-1}Z_1'X(X'X)^{-1}X'y_1$.

From $\hat{\alpha}_{1,A}' = \alpha_1' + (R_1'R_1)^{-1}R_1'r_1$, we have $\sqrt{T}(\hat{\alpha}_{1,A}' - \alpha_1') = \left(\frac{R_1'R_1}{T}\right)^{-1}\left(\frac{R_1'r_1}{\sqrt{T}}\right)$.

$$p\lim\left\{\left[\sqrt{T}(\hat{\alpha}_{1,A}' - \alpha_1')\right]\left[\sqrt{T}(\hat{\alpha}_{1,A}' - \alpha_1')\right]^T\right\}$$
$$= p\lim\left[\left(\frac{R_1'R_1}{T}\right)^{-1}\left(\frac{R_1'r_1}{\sqrt{T}}\right)\left(\frac{r_1'R_1}{\sqrt{T}}\right)\left(\frac{R_1'R_1}{T}\right)^{-1}\right]$$
$$= \sigma_1^2 p\lim\left(\frac{R_1'R_1}{T}\right)^{-1}$$

(Since $p\lim r_1r_1' = p\lim P'X'\varepsilon_1\varepsilon_1'XP = \sigma_1^2 I$)

$$\sqrt{T}(\hat{\alpha}_{1,A}' - \alpha_1') \xrightarrow{Asy} N\left[0, \sigma_1^2 p\lim\left(\frac{R_1'R_1}{T}\right)^{-1}\right].$$

$$Est.Var.(\sqrt{T}\hat{\alpha}_{1,A}') = T\sigma_1^2 \begin{pmatrix} Y_1'X(X'X)^{-1}X'Y_1 & Y_1'X(X'X)^{-1}X'X_1 \\ X_1'X(X'X)^{-1}X'Y_1 & X_1'X(X'X)^{-1}X'X_1 \end{pmatrix}^{-1}$$

(Substitute $Z_1 = \begin{pmatrix} Y_1 & X_1 \end{pmatrix}$ into $R_1'R_1 = Z_1'X(X'X)^{-1}X'Z_1$.)

$$\hat{\sigma}_1^2 = \frac{(y_1 - Y_1\hat{\beta}_1' - X_1\hat{\gamma}_1')'(y_1 - Y_1\hat{\beta}_1' - X_1\hat{\gamma}_1')}{T} = e_1'e_1/T$$

(or replace T with $T - G^{\Delta} - K^* + 1$)

$$\hat{\sigma}_{gh} = \frac{(y_g - Y_g\hat{\beta}_g' - X_g\hat{\gamma}_g')'(y_h - Y_h\hat{\beta}_h' - X_h\hat{\gamma}_h')}{T} = e_g'e_h/T$$

12. Three Stage Least Squares

Using Aitken's approach, we get consistent estimators $\hat{\alpha}_{1,A}'$ for equation 1 by doing regression $W_1 = R_1\alpha_1' + r_1$. If we do the same procedure to every equation, we have $W = R\alpha' + r$, where

$$W = \begin{pmatrix} W_1 \\ W_2 \\ ... \\ W_G \end{pmatrix}, \alpha' = \begin{pmatrix} \alpha_1' \\ \alpha_2' \\ ... \\ \alpha_G' \end{pmatrix}, r = \begin{pmatrix} r_1 \\ r_2 \\ ... \\ r_G \end{pmatrix}, R = \begin{pmatrix} R_1 & 0 & ... & 0 \\ 0 & R_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & R_G \end{pmatrix}.$$

Then $\hat{\hat{\alpha}}_{3SLS}' = [R'V^{-1}R]^{-1}R'V^{-1}W$. (Here we are using GLS estimators formula.)

$$V = E(rr') = \begin{pmatrix} \sigma_1^2 I & \sigma_{12}I & ... & \sigma_{1G}I \\ \sigma_{21}I & \sigma_2^2 I & ... & \sigma_{2G}I \\ ... & ... & ... & ... \\ \sigma_{G1}I & \sigma_{G2}I & ... & \sigma_G^2 \end{pmatrix} = \sum \otimes I$$

The corresponding GLS estimators are:

$$\hat{\hat{\alpha}}_{3SLS}' = [R'(\overset{-1}{\sum} \otimes I)R]^{-1}R'(\overset{-1}{\sum} \otimes I)W.$$

$$\sqrt{T}(\hat{\hat{\alpha}}_{GLS}' - \alpha') \xrightarrow{Asy} N\left[0, p\lim\left(\frac{R'(\sum^{-1} \otimes I)R}{T}\right)^{-1}\right].$$

To get the FGLS estimators, following suggestion by Zellner and Theil, use TSLS residuals to compute $\hat{\sum}$: $\hat{\sigma}_g^2 = e_g'e_g/T$ and $\hat{\sigma}_{gh} = e_g'e_h/T$.

13. Comparison of Methods of Regressing SEM

First of all, we know OLS estimators are not consistent, and indirect least square estimators are consistent.

Second, if we consider one equation at a time, this method is called limited information method. Note that TSLS, IV, and Aitken estimators are exactly the same. Also note that TSLS estimators are asymptotically equivalent to the Limited Information Maximum Likelihood estimators.

Finally, if we try to consider the system of equations simultaneously in order to capture the cross-equation correlation, this method is called full information method. We can use either 3SLS or Full Information Maximum Likelihood estimators, which are asymptotically equivalent.

14. Testing

(1) Test particular parameter:

Using Aitken's approach, we have $W_1 = R_1\alpha_1' + r_1$, and $\hat{\alpha}_{1,A}' = (R_1'R_1)^{-1}R_1'W_1$.

$$\sqrt{T}(\hat{\alpha}_{1,A}' - \alpha_1') \xrightarrow{Asy} N\left[0, \sigma_1^2 p\lim\left(\frac{R_1'R_1}{T}\right)^{-1}\right]$$

We can estimate $\sigma_1^2 p\lim\left(\frac{R_1'R_1}{T}\right)^{-1}$ as $\Phi_1 = T\hat{\sigma}_1^2(R_1'R_1)^{-1}$, which is a $(G^\Delta - K^*) \times (G^\Delta - K^*)$ square and symmetric matrix.

The so-called "T-Statistic" is: $\frac{\hat{\alpha}_{1,i}' - \alpha_{1,i}'}{\sqrt{\Phi_{1,ii}/T}} \xrightarrow{Asy} N(0,1)$.

Denote by $\tilde{r}_1$ the residual for regression $W_1 = R_1\alpha_1' + r_1$. Then

$$\tilde{r}_1 = [I_K - R_1(R_1'R_1)^{-1}R_1']r_1, \text{ and}$$

$$\tilde{r}_1'\tilde{r}_1 = r_1'[I_K - R_1(R_1'R_1)^{-1}R_1']r_1 \xrightarrow{Asy} \chi^2(\nu).$$

$$\nu = trace[I_K - R_1(R_1'R_1)^{-1}R_1'] = K - K^* - (G^\Delta - 1) > 0$$

This is for over-identified equation, i.e., ν is the degree of over-identification, then we have the alternative t-statistic:

$$\frac{\hat{\alpha}_{1,i}' - \alpha_{1,i}'}{\sqrt{\tilde{r}_1'\tilde{r}_1(R_1'R_1)_{ii}^{-1}/\nu}} \xrightarrow{Asy} t(\nu).$$

(2) Test identification problem:

Recall the set-up of Kmenta's Approacch:

$$B\Pi + \Gamma = 0 \Rightarrow$$
$$\begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \end{pmatrix}\Pi = -\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \end{pmatrix}, i.e.,$$
$$\beta_1\Pi = -\gamma_1.$$

Let $^\Delta$ stand for endogenous variable, and * stand for exogenous variable.

$G^\Delta$is the number of included endogenous variables;
$G^{\Delta\Delta}$is the number of excluded endogenous variables;
$K^*$is the number of included exogenous variables;
$K^{**}$is the number of excluded exogenous variables.

Partition matrices properly as following:

$$\beta_1 = \begin{pmatrix} \beta^\Delta & O^{\Delta\Delta} \end{pmatrix}, \quad \gamma_1 = \begin{pmatrix} \gamma^* & O^{**} \end{pmatrix}, \quad \Pi = \begin{pmatrix} \Pi^{\Delta *} & \Pi^{\Delta **} \\ \Pi^{\Delta\Delta *} & \Pi^{\Delta\Delta **} \end{pmatrix}.$$

Thus $\beta_1\Pi = -\gamma_1$ implies

$$\begin{pmatrix} \beta^\Delta & O^{\Delta\Delta} \end{pmatrix}\begin{pmatrix} \Pi^{\Delta *} & \Pi^{\Delta **} \\ \Pi^{\Delta\Delta *} & \Pi^{\Delta\Delta **} \end{pmatrix} = -\begin{pmatrix} \gamma^* & O^{**} \end{pmatrix},$$

i.e., $\beta^\Delta\Pi^{\Delta *} = -\gamma^*$ and $\beta^\Delta\Pi^{\Delta **} = O^{**}$.

From $\beta^\Delta\Pi^{\Delta **} = O^{**}$, we know that we have K^{**} equations and $G^\Delta - 1$ unknowns.

Define $g \equiv G^{\Delta} - 1$. Considering the rank condition, we have the following:

If $Rank(\Pi^{\Delta **}) = g$, then the rank condition satisfies, and we can solve for β^{Δ} uniquely;
if $Rank(\Pi^{\Delta **}) > g$, then the rank condition fails due to too many exclusion, and the test for this situation is called "zero restriction test";
if $Rank(\Pi^{\Delta **}) < g$, then the rank condition fails due to too little independent relations, and the test for this situation is called "rank test".

(a) Rank Test

$H_0 : Rank(\Pi^{\Delta **}) < g$,i.e., not identified; $H_A : Rank(\Pi^{\Delta **}) = g$, i.e., identified.

Let $G^{\Delta} = 1 + G^{\varphi}$. Let y_1 be the normalized endogenous variable.

Since $\beta^{\Delta} = \begin{pmatrix} -1 & \beta_1^{\varphi} \end{pmatrix}$, where β_1^{φ} is the coefficients associated with Y_1 in the first structural equation, $\beta^{\Delta}\Pi^{\Delta **} = O^{**}$ implies $\begin{pmatrix} -1 & \beta_1^{\varphi} \end{pmatrix} \begin{pmatrix} \pi^{**} \\ \Pi^{\varphi **} \end{pmatrix} = O^{**}$, where π^{**} is the single row of coefficients corresponding to y_1 in the reduced form, and $\Pi^{\varphi **}$ is the coefficients block corresponding to Y_1 in the reduced form.

Thus $\beta_1^{\varphi}\Pi^{\varphi **} = -\pi^{**}$ implies that $Rank(\Pi^{\Delta **}) = Rank(\Pi^{\varphi **})$. Then we have the alternative set of hypothesis: $H_0 : Rank(\Pi^{\varphi **}) < g$,i.e., not identified; $H_A : Rank(\Pi^{\varphi **}) = g$, i.e., identified.

Since we have totally $g + 1$ included endogenous variables, thus we have $g+1$ possible ways of normalization. Only if all of normalization fail the rank test, can we say the considered equation fails the rank test.

The likelihood ratio statistic is $\hat{\lambda}_k = \left(1 + \hat{\phi}_g\right)^{-T/2}$, and the asymptotic distributions of the statistic are: $-T \cdot \ln(\hat{\lambda}_k) \xrightarrow{asy} \chi^2(K^{**} - g + 1)$ or $\hat{\phi}_g \cdot \frac{T-K}{K^{**}-g+1} \xrightarrow{asy} F(K^{**} - g + 1, T - K)$, where $\hat{\phi}_g$ is the smallest root of $\left|W_d^* - \phi W^{\Delta\Delta *}\right| = 0$, and $W^{\Delta\Delta *} = Y'MY$, $W_d^* = Y_1'M_dY_1$, $M = I - X(X'X)^{-1}X'$,

$$M_1 = I - X_1(X_1'X_1)^{-1}X_1', \quad M_d = M_1 - M.$$

The decision rule for this test statistic is: reject H_0 if $\hat{\phi}_g$ is significantly larger than 0.

In particular, if $G^{\Delta} = 2$, then we have exact distribution of the test statistic in the sense of the following: $G^{\Delta} = 2 \Rightarrow G^{\varphi} = 1$ *and* $g = 1 \Rightarrow \Pi^{\varphi **}$ *is* $1 \times K^{**}$, then $Rank(\Pi^{\varphi **}) < g$ is equivalent to $Rank(\Pi^{\varphi **}) = 0$. Then we can use Wald test to test the elements of $\Pi^{\varphi **}$ are all jointly zero.

(b) Zero Restriction Test

$H_0 : Rank(\Pi^{\Delta **}) > g$,i.e., not identified; $H_A : Rank(\Pi^{\Delta **}) = g$, i.e., identified.

The likelihood ratio statistic is $\hat{\lambda}_z = \left(1 + \hat{\xi}\right)^{-T/2}$, and the asymptotic distributions of the statistic are: $-T \cdot \ln(\hat{\lambda}_z) \xrightarrow{asy} \chi^2(\nu)$ or $\hat{\xi} \cdot \frac{T-K}{\nu} \xrightarrow{asy} F(\nu, T - K)$, where $\nu = K^{**} - g$ is the degree of over-identification, and $\hat{\xi}$ is the smallest root of $\left|W_d^* - \xi W^{\Delta\Delta *}\right| = 0$, and

$$W^{\Delta\Delta *} = Y'MY, \quad W_d^* = Y_1'M_dY_1, \quad M = I - X(X'X)^{-1}X',$$
$$M_1 = I - X_1(X_1'X_1)^{-1}X_1', \quad M_d = M_1 - M.$$

The decision rule for this test statistic is: reject H_0 if $\hat{\xi}$ is significantly larger than 0.

15. A Recursive Two-Equation System

A recursive Two-Equation System without Exclusion Restrictions

(1) Basics

Let's consider the following structural equations:

$$\begin{pmatrix} 1 & 0 \\ -\alpha & 1 \end{pmatrix} \begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \mathbf{X} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}.$$

The reduced-form equations are:

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \alpha\beta_1+\beta_2 \end{pmatrix} \mathbf{X} + \begin{pmatrix} \mathbf{u}_1 \\ \alpha\mathbf{u}_1+\mathbf{u}_2 \end{pmatrix}.$$

Clearly, the first equation passes the order condition but the second equation fails the order condition. Yet this recursive system is identified. Why? The rank and order condition is the necessary and sufficient condition for a system to be identified when there is no restrictions on the disturbance terms on the equations. Since we have some restrictions on the disturbance terms, the system is identified through the variance component. How?

$$\Sigma_{VV} = \begin{pmatrix} \sigma_{u_1}^2 & \alpha\sigma_{u_1}^2 + \sigma_{u_2}^2 \\ \alpha\sigma_{u_1}^2 + \sigma_{u_2}^2 & \sigma_{u_2}^2 \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{V_1V_1} & \mathbf{S}_{V_2V_1} \\ \mathbf{S}_{V_1V_2} & \mathbf{S}_{V_2V_2} \end{pmatrix}.$$

We get $\hat{\alpha} = (\mathbf{S}_{V_1V_2} - \mathbf{S}_{V_2V_2})/\mathbf{S}_{V_1V_1}$, and then $\hat{\beta}_1, \hat{\beta}_2$ can be estimated from the reduced form equations.

(2) Measurement errors in system of equations

To see why measurement errors causes inconsistent estimates, consider the following model $\mathbf{y} = \beta'\mathbf{x} + \mathbf{u}$ and $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{e}$. $\mathbf{u}$ and $\mathbf{e}$ are independent and we observe $\mathbf{x}$ with measurement error. That is, we don't observe $\mathbf{x}$ directly but $\tilde{\mathbf{x}}$. Let $\mathbf{Y} = (\mathbf{y}' \ \tilde{\mathbf{x}}')'$, then we have

$$\Sigma_{YY} = \begin{pmatrix} \beta'\Sigma_{xx}\beta + \Sigma_{uu} & \beta'\Sigma_{xx} \\ \Sigma_{xx}\beta & \Sigma_{xx} + \Sigma_{ee} \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{y\tilde{x}} \\ \mathbf{S}_{\tilde{x}y} & \mathbf{S}_{\tilde{x}\tilde{x}} \end{pmatrix}.$$

Clearly $\hat{\beta}_{OLS} = \frac{\mathbf{S}_{\tilde{x}y}}{\mathbf{S}_{\tilde{x}\tilde{x}}} = \frac{\Sigma_{xx}\beta}{\Sigma_{xx}+\Sigma_{ee}}$ is not consistent.

While we can use instrumental variables approach to tackle the measurement error problem, we can also use the approach of multiple measures. For example, suppose we have $\mathbf{x}_1 = \mathbf{x} + \mathbf{e}_1$ and $\mathbf{x}_2 = \mathbf{x} + \mathbf{e}_2$, where $\mathbf{u}$, $\mathbf{e}_1$ and $\mathbf{e}_2$ are mutually independent. Let $\mathbf{Y} = (\mathbf{y}' \ \mathbf{x}_1' \ \mathbf{x}_2')'$, then we have

$$\Sigma_{YY} = \begin{pmatrix} \beta'\Sigma_{xx}\beta + \Sigma_{uu} & \beta'\Sigma_{xx} & \beta'\Sigma_{xx} \\ \Sigma_{xx}\beta & \Sigma_{xx} + \Sigma_{e_1e_1} & \Sigma_{xx} \\ \Sigma_{xx}\beta & \Sigma_{xx} & \Sigma_{xx} + \Sigma_{e_2e_2} \end{pmatrix}.$$

Clearly β is overly identified by one and everything else is exactly identified. If we allow $\mathbf{e}_1$ and $\mathbf{e}_2$ having the same variance, then Σ_{ee} is also overly identified by one.

Suppose that the structural equations are $\mathbf{BY} + \Gamma\mathbf{X} = \varepsilon$ where $\mathbf{X}$ is measured with error, i.e., we observe $\tilde{\mathbf{X}} = \mathbf{X} + \delta$. The reduced form is $\mathbf{Y} = -\mathbf{B}^{-1}\Gamma\mathbf{X} + \mathbf{B}^{-1}\varepsilon$. Let $\mathbf{W} = (\mathbf{Y}' \ \tilde{\mathbf{X}}')'$, then we have

$$\begin{aligned}\Sigma_{WW} &= \begin{pmatrix} -(\mathbf{B}^{-1}\Gamma)\Sigma_{XX}(\mathbf{B}^{-1}\Gamma)' + \mathbf{B}^{-1}\Sigma_{\varepsilon\varepsilon}\mathbf{B}^{-1\prime} & -(\mathbf{B}^{-1}\Gamma)\Sigma_{XX} \\ -\Sigma_{XX}(\mathbf{B}^{-1}\Gamma)' & \Sigma_{XX} + \Sigma_{\delta\delta} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{S}_{YY} & \mathbf{S}_{Y\tilde{X}} \\ \mathbf{S}_{\tilde{X}Y} & \mathbf{S}_{\tilde{X}\tilde{X}} \end{pmatrix}.\end{aligned}$$

Hence we have $\mathbf{S}_{Y\tilde{X}} = -\mathbf{B}^{-1}\Gamma(\mathbf{S}_{\tilde{X}\tilde{X}} - \Sigma_{\delta\delta})$, or $\mathbf{B}\mathbf{S}_{Y\tilde{X}} = -\Gamma(\mathbf{S}_{\tilde{X}\tilde{X}} - \Sigma_{\delta\delta})$, which can also be derived by post-multiplying $\mathbf{BY} = -\Gamma(\tilde{\mathbf{X}} + \delta) + \varepsilon$ by $\tilde{\mathbf{X}}$ and then taking the expectation of both sides.

CHAPTER 13

Models with Discrete Dependent Variables

1. Truncated Model

Suppose that $y \sim N(\mu, \sigma^2)$ and all our observations are $y \geqslant y^*$. Then y^* is called the point of truncation, and the density of y is

$$f(y|y \geqslant y^*) = \tfrac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{1-\Phi[(y^*-\mu)/\sigma]},$$

where ϕ and Φ are p.d.f. and c.d.f. of standard normal, respectively.

2. Censored Model

Among the sample of size n, we have a sub-sample of size n_1, the information obtained of which is simply $y \leqslant y^*$, and we have exact values known for the rest of the sample. The joint density of the sample is

$$\tbinom{n}{n_1}\left[\Phi(\tfrac{y^*-\mu}{\sigma})\right]^{n_1} \cdot \prod_{i=1}^{n-n_1} \tfrac{1}{\sigma} \cdot \phi(\tfrac{y_i-\mu}{\sigma}).$$

3. Classification of Discrete Dependent Variables

It may be the case that some dependent variables, such as the number of patents granted to a company in a year, assume discrete values, but those discrete values are not categorical. Here we are mainly concerned with categorical values. In particular, the categorical values can be further classified as "ordered," "sequential," or "non-ordered non-sequential."

4. Probit/Logit Model for a Binary Case

Assume that $\mathbf{Y}_i^* = \mathbf{X}_i\beta + \mathbf{u}_i$ and we observe that $Y_{it} = 1$ if $Y_{it}^* > 0$ and $Y_{it} = 0$ otherwise. Clearly we have

$$\Pr(Y_{it} = 1) = \Pr(u_{it} > -X_{it}\beta) = 1 - F(-X_{it}\beta),$$

where $F(\cdot)$ is the c.d.f. for u_{it}. The sample likelihood is then

$$\mathcal{L} = \prod_{Y_{it}=0} F(-X_{it}\beta) \prod_{Y_{it}=1} [1 - F(-X_{it}\beta)].$$

If the c.d.f. of u_{it} is assumed to be logistic, we have the logit model. In this case, we have

$$\begin{aligned}
\Pr(Y_{it}=0) &= F(-X_{it}\beta) = \tfrac{\exp(-X_{it}\beta)}{1+\exp(-X_{it}\beta)} = \tfrac{1}{1+\exp(X_{it}\beta)},\\
\Pr(Y_{it}=1) &= 1-F(-X_{it}\beta) = \tfrac{1}{1+\exp(-X_{it}\beta)} = \tfrac{\exp(X_{it}\beta)}{1+\exp(X_{it}\beta)},\\
\tfrac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)} &= \tfrac{\Pr(Y_{it}=1)}{\Pr(Y_{it}=0)} = e^{X_{it}\beta}, \text{ and}\\
\ln\left[\tfrac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)}\right] &= X_{it}\beta.
\end{aligned}$$

If the c.d.f. of u_{it} is assumed to be $N(0,\sigma^2)$, we have the probit model. In this case,

$$F(-\tfrac{X_{it}\beta}{\sigma}) = \int_{-\infty}^{-X_{it}\beta/\sigma} \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{u^2}{2})du.$$

It can be easily seen from the equation above and the sample likelihood function that we can estimate only $\frac{\beta}{\sigma}$, not β and σ separately. Hence we might as well assume $\sigma=1$ to start with. In particular, we have

$$\begin{aligned}
\Pr(Y_{it}=0) &= \Phi(-X_{it}\beta) = 1-\Phi(X_{it}\beta),\\
\Pr(Y_{it}=1) &= 1-\Phi(-X_{it}\beta) = \Phi(X_{it}\beta).
\end{aligned}$$

The marginal effects for logit and probit models can be found as the following:

$$\tfrac{\partial \hat{P}_{1t}}{\partial x_{itk}} = \tfrac{\exp(X_{it}\beta)}{[1+\exp(X_{it}\beta)]^2}\beta_k, \text{ and } \tfrac{\partial \hat{P}_{1t}}{\partial x_{itk}} = \phi(X_{it}\beta)\beta_k,$$

where $\hat{P}_{1t}$ is the predicated probability for $Y_{it}=1$.

Because the c.d.f. of normal and logistic distributions are very close to each other except at tails, we are not likely to get very different results from using two models (note that only the marginal effects are directly comparable though), unless the samples are large so that we have enough observations at the tails.

5. Ordered Probit/Logit Model

Assume that $\mathbf{Y}_i^* = \mathbf{X}_i\beta + \mathbf{u}_i$ and $u_{it} \sim N(0,1)$. The $K-1$ thresholds are denoted as $\tau_1, ..., \tau_{K-1}$. Clearly we have $\tau_0 = -\infty$ and $\tau_K = +\infty$. We normalize $\tau_1 = 0$ for convenience. What we are observing is $Y_{it} = k$ with probability $F(\tau_k - X_{it}\beta) - F(\tau_{k-1} - X_{it}\beta), \forall k$. Equivalently, we have $\Pr(Y_{it} \leqslant k) = F(\tau_k - X_{it}\beta)$.

6. Sequential Probit

Assume that $Y_{ikt}^* = X_{ikt}\beta_k + u_{ikt}$ and we observe $Y_{ikt} = k$ with probability

$$\begin{aligned}
&\Pr\{Y_{ijt}^* > 0, \forall j \in [1, k-1], \text{ and } Y_{ikt}^* \leqslant 0\}, \forall k \in [2, K-1], \text{ and}\\
&\Pr(Y_{it}=1) = \Pr(Y_{i1t}^* \leqslant 0),\\
&\Pr(Y_{it}=K) = \Pr\{Y_{ikt}^* > 0, \forall k \in [1, K-1]\}.
\end{aligned}$$

Note that we are allowing different underlying schemes for different categories. For the case of no correlation across categories, we have $u_{ikt} \sim$ i.i.d. $N(0,1)$. If $K=4$, then we have

$$\begin{aligned}
\Pr(Y_{it}=1) &= \Phi(-X_{i1t}\beta_1)\\
\Pr(Y_{it}=2) &= \Phi(X_{i1t}\beta_1)\Phi(-X_{i2t}\beta_2)\\
\Pr(Y_{it}=3) &= \Phi(X_{i1t}\beta_1)\Phi(X_{i2t}\beta_2)\Phi(-X_{i3t}\beta_3)\\
\Pr(Y_{it}=4) &= \Phi(X_{i1t}\beta_1)\Phi(X_{i2t}\beta_2)\Phi(X_{i3t}\beta_3).
\end{aligned}$$

For the case allowing correlation across categories, we have $\mathbf{u}_{it} \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is both the covariance and correlation matrix of $\mathbf{u}_{it}$. If $K = 4$, then we have

$$\begin{aligned}
\Pr(Y_{it}=1) &= \Phi(-X_{i1t}\beta_1)\\
\Pr(Y_{it}=2) &= \Phi(X_{i1t}\beta_1, -X_{i2t}\beta_2| - \rho_{12})\\
\Pr(Y_{it}=3) &= \Phi(X_{i1t}\beta_1, X_{i2t}\beta_2, -X_{i3t}\beta_3|\rho_{12}, -\rho_{13}, -\rho_{23})\\
\Pr(Y_{it}=4) &= \Phi(X_{i1t}\beta_1, X_{i2t}\beta_2, X_{i3t}\beta_3|\rho_{12}, \rho_{13}, \rho_{23}).
\end{aligned}$$

Note that whenever we change $-X_{ikt}\beta_k$ into $X_{ikt}\beta_k$, we have to reverse the sign of $\rho_{km}, \forall m \in [1, K-1]$. Note further that the sample fraction of $Y_{it} = 1$ gives us one moment, and we can estimate β_1 accordingly. Although the sample fraction of $Y_{it} = 2$ gives us one additional moment, we have two more parameters, β_2 and ρ_{12}, to estimate, a mission impossible. In the case where $K = 4$, we have merely three moments yet we have six parameters to be estimated. We inevitably run into an unidentified system.

7. Unordered Non-Sequential Model

Unordered Non-Sequential Model with Mutually Exclusive and Exhaustive Categories

(1) Basics

Denote as P_k the probability associated with the k^{th} category, $k = 1, 2, ..., K$. Then the idea is to express these probabilities in binary form. Let

$$\frac{P_k}{P_k + P_K} = F(X_i\beta_k), \forall k = 1, ..., K-1,$$

then

$$\frac{P_k}{P_K} = \frac{F(X_i\beta_k)}{1 - F(X_i\beta_k)} \equiv G(X_i\beta_k).$$

Since $\sum_{k=1}^{K-1} \frac{P_k}{P_K} = \frac{1-P_K}{P_K} = \frac{1}{P_K} - 1$, we have

$$P_k = \frac{1}{1 + \sum_{k=1}^{K-1} P_k/P_K}.$$

Therefore,

$$P_k = \frac{G(X_i\beta_k)}{1 + \sum_{m=1}^{K-1} G(X_i\beta_m)}, \forall k = 1, ..., K-1.$$

(2) Multinomial logit

Assume that $U_{ik}^* = X_{ik}\beta_k + u_{ik}$, $k = 1, ..., K$ and u_{ik} is distributed as logistic. Note that we cannot allow correlation across categories when using logit model. Since the sum of the probabilities for any $K-1$ categories will uniquely determine the probability for the only category left, i.e., we have only $K-1$ equations for K β_k's, we need some sort of normalization for β.

If we choose $\beta_K = 0$ then

$$P_{ik} = \frac{\exp(X_{ik}\beta_k)}{1+\sum_{m=1}^{K-1}\exp(X_{im}\beta_m)}, k = 1, ..., K-1.$$

If we choose $\sum_{k=1}^{K}\beta_k = 0$, then

$$P_{ik} = \frac{\exp(X_{ik}\beta_k)}{\sum_{m=1}^{K}\exp(X_{im}\beta_m)}, k = 1, ..., K.$$

(Find out how to derive the probabilities above from the double-exponential distribution.)

(3) Multinomial probit

Assume that $U_{ik}^* = X_{ik}\beta_k + u_{ik}, k = 1, ..., K$ where $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{R})$, $\mathbf{R}$ is both the covariance and correlation matrix for $\mathbf{u}_i$. We have the following probabilities:

$$\begin{aligned} P_{ik} &= \Pr\{U_{im}^* \leqslant U_{ik}^*, \forall m \neq k\} \\ &= \Pr\{u_{im} - u_{ik} \leqslant -(X_{im}\beta_m - X_{ik}\beta_k), \forall m \neq k\} \\ &= \Phi_{K-1}[-(X_{i1}\beta_1 - X_{ik}\beta_k), ..., -(X_{iK}\beta_K - X_{ik}\beta_k)|\mathbf{ARA}'], \end{aligned}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_{k-1} & -\mathbf{i}_{k-1} & \mathbf{0}_{k-1} \\ \mathbf{0}_{K-k} & -\mathbf{i}_{K-k} & \mathbf{I}_{K-k} \end{bmatrix}.$$

Let's consider the case where $K = 3$ for a closely related model, $U_{ik}^* = X_i\beta_k + u_{ik}, k = 1, ..., K$. Let's use U_{i3}^* as the base case. Then we have

$$\begin{aligned} \Pr(U_{i1}^* \leqslant U_{i3}^*) &= \Pr\left[\frac{u_{i1}-u_{i3}}{\sigma_1-\sigma_3} \leqslant -\frac{X_i(\beta_1-\beta_3)}{\sigma_1-\sigma_3}\right], \\ \Pr(U_{i2}^* \leqslant U_{i3}^*) &= \Pr\left[\frac{u_{i2}-u_{i3}}{\sigma_2-\sigma_3} \leqslant -\frac{X_i(\beta_2-\beta_3)}{\sigma_2-\sigma_3}\right]. \end{aligned}$$

What we can estimated is $\frac{\beta_1-\beta_3}{\sigma_1-\sigma_3}$, $\frac{\beta_2-\beta_3}{\sigma_2-\sigma_3}$, and $\rho_{u_{i1}-u_{i3},u_{i2}-u_{i3}}$, yet we have $\beta_k, \sigma_k, (k = 1, 2, 3)$, and $\rho_{12}, \rho_{13}, \rho_{23}$ to estimate in the original model. Although we couldn't estimate the full model, we do capture the most interesting features of the model from the three parameters we can estimate.

(4) Panel/replicated data probit model

Assume $Y_{it}^* = X_{it}\beta + e_i + d_{it}, t = 1, ..., T_i$, where $e_i \sim N(0, \sigma_e^2)$ and $d_{it} \sim N(0, 1)$. We observe $Y_{it} = 0$ if $Y_{it}^* \leqslant 0$ and $Y_{it} = 1$ if $Y_{it}^* > 0$. Let's define an indicator variable for the sign, $S_{it} = 1 - 2Y_{it}$. Clearly we have $S_{it} = 1$ if $Y_{it}^* \leqslant 0$ and $S_{it} = -1$ if $Y_{it}^* > 0$. Then the sample likelihood is

$$\mathcal{L}_i = \Phi_{T_i}\left\{\frac{-S_{i1}X_{i1}\beta}{\sqrt{1+\sigma_e^2}}, \frac{-S_{i2}X_{i2}\beta}{\sqrt{1+\sigma_e^2}}, ..., \frac{-S_{iT_i}X_{iT_i}\beta}{\sqrt{1+\sigma_e^2}}|\mathbf{R}_{T_i}\right\},$$

where $\mathbf{R}_{T_i} = \mathbf{S}_i\mathbf{S}_i' \oplus (\sigma_e^2\mathbf{ii}' + \mathbf{I}_{T_i})$. $\oplus$ denotes element-by-element product.

The likelihood for the individual i is

$$\mathcal{L}_i = \int_{e_i} \tfrac{1}{\sigma_e}\phi(\tfrac{e_i}{\sigma_e}) \prod_{t=1}^{T_i} \Phi[-S_{it}(X_{it} + e_i)]de_i.$$

Part 4

Applications in Financial Markets

CHAPTER 14

GMM

CHAPTER 15

Difference in Difference