1 Probability Concepts

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(B|A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

$$= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

2 Random Variables

$F_X(x) = \mathbb{P}(X \leq x)$ and $f_X(x) = \mathbb{P}(X = x_i)$

2.1 More than one variable

$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$

DRV: $P_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$

CRV: $P((X, Y) \in A) = \int\int_A f(x, y)dx\, dy$

Marginal Distribution

DRV: $$p_x(x_i) = \sum_j p(x_i, y_j)$$

CRV: $$f_x(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

Conditional Distribution

DRV: $$P(X = x_i|Y = y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} = \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)}$$

CRV: $$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Law of total probability

DRV: $$P_X(x_i) = \sum_{y_j} P_{X|Y}(x_i|y_j)P_Y(y_j)$$

CRV: $$f_X(x_i) = \int_{-\infty}^{\infty} f_{X|Y}(x_i|y_j)f_Y(y_j)dy$$

2.2 Expectations

DRV:
$$\mathbb{E}[X] = \sum_i x_i p(x_i)$$
$$\mathbb{E}[g(X)] = \sum_i g(x_i)p(x_i)$$

CRV:
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)dx$$
$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

$$\mathbb{E}[X] = \int_0^{\infty} 1 - F(X)dx \quad \text{for any nonnegative RV}$$

Conditional Expectation

DRV:
$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y)$$
$$\mathbb{E}[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y)$$

CRV:
$$\mathbb{E}[X|Y = y] = \int x f_{X|Y}(x|y)dx$$
$$\mathbb{E}[g(X)|Y = y] = \int g(x)f_{X|Y}(x|y)dx$$

Law of Iterated Expectations

$$\mathbb{E}_Y[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}(Y|X = x)]$$

2.3 Transformation of RVs

Let $Y = g(X)$

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |\frac{d}{dy}g^{-1}(y)|$$

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |J|$$

2.4 Moments

$$M(t) = \mathbb{E}(e^{tX}) \quad \text{mgf}$$

$$M^{(j)}(t) = \int x^j e^{tx} f_X(x)dx$$

$$M^{(j)}(0) = \mathbb{E}(X^j)$$

2.5 (Co)variance and Correlation

DRV: $$Var(X) = \sum_i (x_i - \mu)^2 p(x_i)$$

CRV: $$Var(x) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$Cov(X, X) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathbb{E}[XY] = \int\int xyf(x, y)\, dx\, dy$$

$$Cov(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) \cdot f(x, y)\, dx\, dy$$

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

3 Selected Probability Distributions

3.1 Binomial

$$p(k) = \binom{n}{k}p^k(1 - p)^{n-k}$$

$$E[X] = np$$

$$Var(X) = np(1 - p)$$

$$M(t) = (1 - p + pe^t)^n$$

Note: if $n = 1$, it's a Bernoulli distribution

3.2 Poisson

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E[X] = \lambda$$

$$Var(X) = \lambda$$

$$M(t) = e^{\lambda(e^t - 1)}$$

3.3 Uniform

$$f(x) = \frac{1}{b - a}$$

$$E[X] = \frac{1}{2}(a + b)$$

$$Var(X) = \frac{1}{12}(b - a)^2$$

$$M(t) = \frac{e^{bt} - e^{at}}{(b - a)t}$$

3.4 Univariate Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$

$$M(t) = e^{\mu t}e^{\frac{\sigma^2 t^2}{2}}$$

χ^2 Distribution

$$U := \sum_{i=1}^{n} Z_i^2$$

$$U \sim \chi_n^2$$

$$E[U] = n, \forall n \geq 1$$

$$Var(U) = 2n, \forall n \geq 1$$

t Distribution

$$T := \frac{Z}{\sqrt{U/n}}$$

$$T \sim t_n$$

$$E[T] = 0, \forall n \geq 2$$

$$Var(T) = \frac{n}{n - 2}, \forall n \geq 3$$

F Distribution

$$W := \frac{U/m}{V/n}$$

$$W \sim F_{m,n}$$

$$E[W] = \frac{n}{n - 2} \qquad \forall n \geq 3$$

$$Var(W) = \frac{2n^2(m + n - 2)}{m(n - 2)^2(n - 4)} \qquad \forall n \geq 5$$

3.5 Multivariate Normal

Notes on matrix algebra: Let Σ be a positive definite matrix. Then it can be factored as $\Sigma = AA'$. $A = \Sigma^{1/2}$, and then $\Sigma^{-1} = (A')^{-1}A^{-1}$. $|A|^{-1} = |A^{-1}|$ and $|A| = |A'|$.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$

$$M(t) = e^{\mu t}e^{\frac{\sigma^2 t^2}{2}}$$

Theorems

A	Linear functions of X are normally distributed $Y = \eta + BX \sim N_k\left(\eta + B\mu, B\Sigma B'\right)$
B	X has density given by $f_X(x) = \frac{1}{(2\pi)^{p/2}\lvert\Sigma\rvert^{1/2}}\exp\left\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\right\}$
C	Independent normally distributed RVs are jointly normal. $X = (X_1', X_2')' \sim N_{p+q}(\mu, \Sigma)$ $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$
D	Conditional normal distribution $(X_1 \lvert X_2 = x_2) \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$
E	Suppose $X_2 \sim N(\mu_2, \Sigma_{22})$ and $X_1 \lvert X_2 = x_2 \sim N(A + Bx_2, \Omega)$. Then $X = (X_1', X_2')'$ has a multivariate normal distribution $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} A + B\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} B\Sigma_{22}B' + \Omega & B\Sigma_{22} \\ \Sigma_{22}B' & \Sigma_{22} \end{pmatrix}\right)$
F	Sums of independent normals $X_1 + X_2 \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$

Let $X \sim N_p(\mu, \Sigma)$. Also let $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

G	The marginal distribution of X_1 is $N_k(\mu_1, \Sigma_{11})$
H	For a normal, a zero correlation implies independence.
I	Characterizing independence of linear combinations of normal variables. If $X \sim N_p(\mu, \Sigma)$, B is a $p \times k$ matrix, and C is a $p \times m$ matrix, then $B'X$ and $C'X$ are independent iff $B'\Sigma C = 0$. Note that $B'X$ and $C'X$ are jointly normal and $B'\Sigma C$ is the covariance.

Quadratics: Assume A is symmetric, then $Y'AY$ is a quadratic form.

J	If $X \sim N_p(\mu, \Sigma)$ where Σ has rank p, then $(X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi_p^2$.
K	Let M denote an idempotent $p \times p$ matrix with rank k, then $Z'MZ \sim \chi_k^2$ $M = P\Lambda P'$, where Λ contains the eigenvalues of M on the diagonal, and the rows of P are the orthonormal eigenvectors. Then $M = [P_1\ P_2]\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} P_1' \\ P_2' \end{bmatrix} = P_1P_1'$ Thus, $P_1'Z \sim N(0, P_1'P_1)$, where $P_1'P_1 = I_k$
L	Let $X = PZ$, and $Q = Z'AZ$, where $PA = 0$, then X and Q are independent.
M	Let $Q_1 = Z'A_1Z$, and $Q_2 = Z'A_2Z$, where $A_1A_2 = 0$. Then Q_1 and Q_2 are independent.

4 Some Useful Inequalities

4.1 Jensen's Inequality

$g(x)$ concave: $E[g(x)] \leq g(E[X])$

$g(x)$ convex: $E[g(x)] \geq g(E[X])$

4.2 Chebyshev's and Markov's Inequality

If X is a random Variable with mean μ and variance σ^2, then

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \qquad \forall\varepsilon > 0$$

$$P(|X| \geq \varepsilon) \leq \frac{\mathbb{E}(|X|^p)}{\varepsilon^p} \qquad \forall\varepsilon > 0 \text{ (Markov)}$$

5 Large Sample Theory

$X_n \xrightarrow{as} X$ if $P\{\omega | \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$

$X_n \xrightarrow{p} X$ if $\forall\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$

$X_n \xrightarrow{ms} X$ if $\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$

$X_n \xrightarrow{d} X$ if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$

Relationship between convergences

- If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{p} X$
- If $X_n \xrightarrow{ms} X$ then $X_n \xrightarrow{p} X$
- If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$

Slutsky's Theorem If

$X_n \xrightarrow{D} X \in \mathbb{R}^k$, where X can be random

$Y_n \xrightarrow{P} A \in \mathbb{R}^p$, where A is fixed

$Z_n \xrightarrow{P} B \in \mathbb{R}^{p\times k}$, where B is fixed

Then, $Y_n + Z_nX_n \xrightarrow{D} A + BX$

5.1 Law of Large Numbers

The sample mean By LLN: $\bar{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

weak LLN Let $X_1, X_2, \ldots$ be a sequence of random variables with $\mathbb{E}(X_i) = \mu$, and $\text{Var}(X_i) = \sigma^2$, and $\text{Cov}(X_i, X_j) = 0\ \forall i \neq j$. Then one can use Chebyshev to show that:

$$\bar{X} \xrightarrow{p} \mu$$

strong LLN Let $X_1, X_2, \ldots$ be $i.i.d.$ with $\mathbb{E}(X_i) = \mu < \infty$, then without saying anything about 2^{nd} moments:

$$\bar{X} \xrightarrow{a.s.} \mu$$

5.2 Central Limit Theorem

Let $Y_1, Y_2, \ldots$ be a sequence of k-dimensional i.i.d. random vectors with $E[Y_i] = \mu_Y$ and $Var(Y_i) = \Sigma$.

$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, I_k)$$

$$\lim_{n\to\infty} P(\sqrt{n}\Sigma^{-\frac{1}{2}}(\bar{Y}_n - \mu) \leq y) = \Phi_k(y), \qquad \forall y \in \mathbb{R}^k$$

Delta Method Let U_n denote a sequence of scalar random variables, and let $V_n = \sqrt{n}(U_n) - a$, where a is a constant. Let $g(\cdot)$ be a continuously differentiable function. Suppose $V_n \xrightarrow{p} V \sim N(\mu, \sigma^2)$. Then

$$\sqrt{n}(g(U_n) - g(a)) \Rightarrow \frac{dg(a)}{da}V \sim N\left(0, \left[\frac{dg(a)}{da}\right]^2\sigma^2\right)$$

$$\sqrt{n}(g(U_n) - g(a))$$

$$\Rightarrow \frac{dg(a)}{da}V \sim N\left(0, \left[\frac{dg(a)}{da}\right]\Sigma\left[\frac{dg(a)}{da}\right],\right)$$

6 Estimators

- An estimator is unbiased, if $\mathbb{E}(\hat{\theta}) = \theta$. Where Bias is defined as $\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$
- Loss Function, say $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ = quadratic loss. This is not the same as expected quadratic loss, which is MSE:

$$E(L(\hat{\theta}, \theta)) = E\left((\hat{\theta} - \theta)^2\right) = \text{mse}(\hat{\theta})$$

$$= \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

- Conclusion: for unbiased estimator $\text{mse}(\hat{\theta}) = \text{Var}(\hat{\theta})$
- An estimator is consistent, if $\hat{\theta} \xrightarrow{p} \theta$

6.1 The Likelihood Function

$$L(\theta, Y) = f(Y|\theta)$$

$$S(\theta, y) = \frac{\partial \ln f(y|\theta)}{\partial\theta} = \frac{1}{f(y|\theta)}\frac{\partial f(y|\theta)}{\partial\theta}$$

$$E[S(\theta, Y)] = \int \frac{\partial f(y|\theta)}{\partial\theta}dy = \int S(\theta, y)f(y|\theta)dy = 0$$

$$I(\theta) = -E\left[\frac{\partial S(\theta, Y)}{\partial\theta}\right] = E\left[S(\theta, Y)^2\right]$$

$$= \text{Var}[S(\theta, Y)]$$

The Cramer-Rao inequality and unbiased estimators

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\text{Var}(S(\theta, Y))} = I(\theta)^{-1}$$

Maximum Likelihood Estimators

$$\max_\theta L_n(\theta) = \max_\theta \prod_{i=1}^{n} f(Y_i|\theta)$$

$$\hat{\theta}_{mle} \xrightarrow{p} \theta_0$$

$$I(\theta_o)^{1/2}\sqrt{n}\left(\hat{\theta}_{mle} - \theta_0\right) \xrightarrow{d} N(0, I)$$

$$\hat{\theta}_{mle} \overset{a}{\sim} N\left(\theta_0, n^{-1}I(\theta_0)^{-1}\right)$$

6.2 Method of Moment Estimators

Assume $\mu = h(\theta_0)$, where μ is $l \times 1$, θ_0 is $k \times 1$ with $k \leq l$. Then $\hat{\theta}_{mm}$ solves

$$\min_\theta J_n(\theta) = \min_\theta(\bar{Y} - h(\theta))'(\bar{Y} - h(\theta))$$

$$\hat{\theta}_{mm} \xrightarrow{p} \theta_0$$

$$\hat{\theta}_{mm} \overset{a}{\sim} N(\theta_0, V)$$

$$V = n^{-1}H^{-1}\left[\frac{\partial h(\theta_o)}{\partial\theta'}\right]'\Sigma\left[\frac{\partial h(\theta)}{\partial\theta'}\right]H^{-1}$$

$$H = \left[\frac{\partial h(\theta_o)}{\partial\theta'}\right]'\left[\frac{\partial h(\theta_o)}{\partial\theta'}\right]$$

7 Sufficient Statistics

Pdf of Y as $f_Y(y|\theta)$, the pdf of S as $f_S(s|\theta)$ and the conditional pdf of Y given $S = s$ as $f_{Y|S}(y|s, \theta) = f_{Y|S}(y|s)$, that is the conditional density of Y given S does not depend on θ.

Factorization Theorem

$$\hat{\theta}(Y) = \arg\max_\theta f(Y|\theta) = \arg\max_\theta g(S|\theta) = \hat{\theta}(S)$$

Rao-Blackwell Theorem Y is RV with mean μ and variance σ_Y^2. X is another RV. Let $\mu(x) = \mathbb{E}[Y|X = x]$. Then $\text{Var}(\mu(x)) \leq \sigma_Y^2$:

$$E(\mu(X)) = \mu$$

$$Y = \mu(X) + (Y - \mu(X))$$

$$\sigma_Y^2 = \text{Var}(\mu(X)) + \text{Var}(Y - \mu(X))$$

$$\text{Var}(\mu(X)) \leq \sigma_Y^2$$

Use this result in estimations: Suppose $\hat{\theta}(Y)$ is an unbiased estimator of θ, so that $\theta = E[\hat{\theta}(Y)]$, and let S be a sufficient statistic for θ. Then $E[\hat{\theta}(S)]$ is an unbiased estimator of θ but the variance is lower by Rao-Blackwell.

$$\theta = E[\hat{\theta}(Y)] = E[E[\hat{\theta}(Y)|S]] = E[\hat{\theta}(S)]$$

8 Hypothesis Tests

8.1 Wald Tests

$$H_0 : \theta = \theta_0 \quad H_a : \theta \neq \theta_0$$

$$\xi = \left(\hat{\theta} - \theta_0\right)'\Omega^{-1}\left(\hat{\theta} - \theta_0\right)$$

$$\xi \sim \chi_k^2 \quad \text{under the null}$$

Therefore, we accept H_0, if ξ is $\leq$ cv, and reject H_0 if $\xi >$ cv. cv solves $\mathbb{P}(\xi > \text{cv}\,|\theta = \theta_0) = \alpha$.

Power $= \mathbb{P}(\xi > \text{cv}\,|H_a \text{ is true}) = \mathbb{P}(\xi > \text{cv}\,|\theta \neq \theta_0)$. Because H_a has many values for θ, the power differs for each value. But assuming that the distribution of $\hat{\theta}$ was based on CLT ($\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$). Then $\Omega = n^{-1}V$, and our test statistic becomes

$$\xi = n\left(\hat{\theta} - \theta_0\right)'V^{-1}\left(\hat{\theta} - \theta_0\right)$$

If the mean of $\hat{\theta}$ is equal to a fixed constant (that is not θ_0), then $\xi \to \infty$, and $\mathbb{P}(\xi > \text{cv}) \to 1$. The test has therefore power $= 1$ for any fixed value of θ under the alternative. When power $\to 1$, a test is said to be consistent.

Hypotheses involving linear functions of θ

$H_0 : R\theta = r_0$ where R is a $j \times k$ matrix with rank j

$H_a : R\theta \neq r_0$ where R is a $j \times k$ matrix with rank j

$$R\hat{\theta} \sim N\left(R\theta, R\Omega R'\right)$$

$$\xi = \left(R\hat{\theta} - r_0\right)'\left(R\Omega R'\right)^{-1}\left(R\hat{\theta} - r_0\right)$$

$$\xi \sim \chi_j^2 \quad \text{under } H_0$$

Hypotheses involving nonlinear functions of θ

$$H_0 : R(\theta) = r_0$$

$$H_a : R(\theta) \neq r_0$$

$$\sqrt{n}(R(\hat{\theta}) - R(\theta)) \Rightarrow N\left(0, HVH'\right) \quad \text{by delta method}$$

$$H = \frac{\partial R(\theta)}{\partial\theta'}$$

$$R(\hat{\theta}) \overset{a}{\sim} N(R(\theta), \tilde{\Omega}) \text{ where } \tilde{\Omega} = n^{-1}HVH'$$

$$\xi = \left(R(\hat{\theta}) - r_0\right)'\tilde{\Omega}^{-1}\left(R(\hat{\theta}) - r_0\right)$$

8.2 Neyman-Pearson Tests

We choose the probability of a type 1 error (accept H_a when H_0) beforehand (size), and then minimize the probability of making a type 2 error (maximize power).

Neyman-Pearson Lemma Choose critical region based on the LR to maximize power.

$$LR(Y) = \frac{L_a(Y)}{L_o(Y)}$$

$$W = \{y|LR(y) > cv\}$$

$$P[LR(Y) > cv|Y \sim F_o] = \alpha \text{ to determine cv}$$

We reject $H_0 : \mu = \mu_0$, if LR is large when $H_a : \mu > \mu_0$. We reject $H_0 : \mu = \mu_0$, if LR is small when $H_a : \mu < \mu_0$. Since the LR critical regions are the same for all of the simple hypotheses making up H_a and each is most powerful, then the LR procedure is said to be Uniformly Most Powerful (UMP) for H_0 vs. H_a.

8.3 Maximizing Weighted Average Power

$H_o : \theta = \theta_0$ simple null hypothesis

$H_a : \theta \in \Theta_a$ composite alternative hypothesis

$w(\theta)$ weight function for values of $\theta \in \Theta_a$

$f(y|\theta)$ density of y, conditional on a value of θ

$\int_W f(y|\theta)dy$ power of the test for a particular θ

$$WAP = \int_{\Theta_a}\left[\int_W f(y|\theta)dy\right]w(\theta)d\theta$$

$$= \int_W\left[\int_{\Theta_a} f(y|\theta)w(\theta)d\theta\right]dy$$

$$= \int_W g(y)dy$$

$g(Y)$ is the density of Y under the assumption that θ is random with density $w(\theta)$. Thus, the problem is equivalent to testing with a simple alternative $\tilde{H}_a : y \sim g(y)$. Then use NP test:

$$LR(Y) = \frac{g(Y)}{f(Y|\theta_o)} = \frac{\int_{\Theta_a} f(Y|\theta)w(\theta)d\theta}{f(Y|\theta_o)}$$

9 Confidence Sets

$$C(Y) = \left\{\theta||(\hat{\theta} - \theta)'\left[\frac{1}{n}\hat{V}\right]^{-1}(\hat{\theta} - \theta) \leq \chi^2_{\kappa,1-\alpha}\right\}$$

$$C(Y) = \left\{\hat{\theta} \pm Z_{1-\alpha/2} \times \sqrt{\frac{1}{n}\hat{V}}\right\}$$

10 The Bayes Approach to Estimation and Inference

10.1 Basic concepts and some jargon

$$f_{\theta|Y}(\tilde{\theta}|y) = \frac{f_{Y,\theta}(y, \tilde{\theta})}{f_Y(y)} = \frac{f_{Y|\theta}(y|\tilde{\theta})f_\theta(\tilde{\theta})}{\int f_{Y,\theta}(y, \tilde{\theta})d\tilde{\theta}}$$

$$= \frac{f_{Y|\theta}(y|\tilde{\theta})f_\theta(\tilde{\theta})}{\int f_{Y|\theta}(y|\tilde{\theta})f_\theta(\tilde{\theta})d\tilde{\theta}}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal Likelihood}}$$

10.2 Bayes Estimators

$$\min_{\hat{\theta}} \mathbb{E}_{\theta|Y=y}[L(\hat{\theta}, \theta)] \quad \text{posterior risk}$$

If loss is quadratic, then we know that $\hat{\theta} = E_{\theta|Y=y}(\theta)$ minimizes the MSE, which is just the posterior mean.

10.3 Bayes Credible Sets

$$P_{\theta|Y=y}(\theta \in C(y)) = 1 - \alpha$$

where the notation emphasizes that the probability is computed using the posterior for $\theta|Y = y$. Because this holds for all y, we also have

$$P_{\theta,Y}(\theta \in C(Y)) = 1 - \alpha$$

where now the probability is computed over θ and Y.

11 Hayashi 1

11.1 Assumptions

11.1.1 Linearity

$y = X\beta + \varepsilon$, X is $(n \times k)\beta$ is $(k \times 1)$

11.1.2 Strict Exogeneity

$$E(\varepsilon_i \mid X) = 0$$
$$E(\varepsilon_i) = E[E(\varepsilon_i \mid X)] = 0$$

11.1.3 Full Rank

No multicollinearity. (XX') has to be invertible.

11.1.4 Homoskedasticity

$$E\left(\varepsilon_i^2 \mid X\right) = \sigma^2 > 0$$
$$E\left(\varepsilon_i \varepsilon_j \mid X\right) = 0 \text{ for i} \neq \text{j}$$

11.1.5 Normality

$$\varepsilon \mid X \sim N\left(0, \sigma^2 I\right)$$
$$\frac{\varepsilon}{\sigma} \mid X \sim N(0, I)$$

for inference

11.2 OLS estimation

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

for one regressor: $\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(x,y)}{\text{Var}(x)}$, and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\hat{y} = Xb$$
$$P \equiv X\left(X'X\right)^{-1} X'$$
$$M \equiv I - P$$
$$\hat{y} = Py = X\left(X'X\right)^{-1} X'y = Xb$$
$$e = My = y - Py = y - \hat{y}$$
$$PX = X$$
$$MX = 0$$
$$e = My = M(X\beta + \varepsilon) = M\varepsilon$$

11.3 Finite sample properties

OLS estimator is unbiased with following variance:

$$\hat{\beta}_{\text{OLS}} = \left(X'X\right)^{-1} X'(X\beta + \varepsilon) = \beta + \left(X'X\right)^{-1} X'\varepsilon$$
$$E[\hat{\beta}] = \beta + \left(X'X\right)^{-1} X'E[\varepsilon \mid x] = \beta$$
$$V[\hat{\beta} \mid X] = E\left[(\hat{\beta} - E[\hat{\beta}])(\hat{\beta} - E[\hat{\beta}])' \mid X\right] = \left(X'X\right)^{-1} X'E\left[\varepsilon\varepsilon' \mid X\right] X \left(X'X\right)^{-1} = \sigma^2 \left(X'X\right)^{-1}$$

CR lower bound (achieved by MLE): $\frac{2\sigma^4}{n}$.

11.3.1 Misspecification

Two models:

$$A: \quad y = X\beta + Z\gamma + \epsilon$$
$$B: \quad y = X\beta + \epsilon$$

OVB: A is true, B is false. Then $\hat{\beta}$ is biased.

$$\hat{\beta} = \beta + \underbrace{\left(X'X\right)^{-1} X'Z\gamma}_{\text{Bias}} + \left(X'X\right)^{-1} X'\varepsilon$$

Irrelevant variable: B is true, A is false. Then $\hat{\beta}$ is unbiased, but inefficient.

$$V(\hat{\beta} \mid X, Z) \geq \sigma^2 \left(X'X\right)^{-1} = V(\hat{\beta} \mid X)$$

12 Hayashi 2

12.1 Inference

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\text{se}\left(\hat{\beta}_j\right)}$$
$$\beta^0 \in \hat{\beta}_j \pm t^{\alpha}_{n-K} \text{ se}\left(\hat{\beta}_j\right)$$

12.1.1 Linear restrictions

$$F = \frac{\left(R'\hat{\beta} - r\right)' \left[R'\left(X'X\right)^{-1} R\right]^{-1} \left(R'\hat{\beta} - r\right)/p}{\left((n-k)s^2/\sigma^2\right)/(n-k)} = \frac{\left(e'^{*}e^{*} - e'e\right)/p}{e'e/(n-K)} = \frac{(SSR^{*} - SSR)/p}{SSR/(n-K)} = \frac{\left(R^2 - R^{*2}\right)/p}{\left(1 - R^2\right)/(n-K)} = \frac{(SST^{*} - SSR)/p}{SSR/(n-K)}$$

p: number of regressors w/o constant
$n - K$: number of individuals minus number of regressors with constant
last equation only for regression output useful

12.1.2 Goodness of Fit

$$R^2 = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2} = MSS/TSS = 1 - \frac{\sum e_i^2}{\sum\left(y_i - \bar{y}\right)^2} = 1 - (RSS/TSS)$$
$$\bar{R}^2 = 1 - \frac{\sum e_i^2/(n-K)}{\sum\left(y_i - \bar{y}\right)^2/(n-1)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k}$$

12.1.3 Wald Statistic

$V(\hat{\beta} \mid X)$ is given in outputs. Do not forget to square SE$(\hat{\beta})$.

$$W = \left(R'\hat{\beta} - r\right)' \left[R'V(\hat{\beta} \mid X)R\right]^{-1} \left(R'\hat{\beta} - r\right) = n\left(R'\hat{\beta} - r\right)' \left[R'\hat{\text{Avar}}(\hat{\beta})R\right]^{-1} \left(R'\hat{\beta} - r\right)$$
$$W = a(\hat{\beta})' \left[\nabla a(\hat{\beta})'\hat{V}(\hat{\beta} \mid X)\nabla a(\hat{\beta})\right]^{-1} a(\hat{\beta}) = na(\hat{\beta})' \left[\nabla a(\hat{\beta})'\hat{\text{Avar}}(\hat{\beta})\nabla a(\hat{\beta})\right]^{-1} a(\hat{\beta})$$

13 Hayashi 3: IV and GMM

13.1 Assumptions IV

13.1.1 Linearity

$$y_i = z_i'\delta + \varepsilon_i \quad \text{but} \quad E\left[z_i\varepsilon_i\right] \neq 0$$

13.1.2 Ergodicity and Stationarity

(y_i, z_i, x_i) ergodic and stationary for LLN.

13.1.3 Exogenous Instrument

$$E\left[g_i\right] = E\left[x_i\varepsilon_i\right] = E\left[x_i\left(y_i - z_i'\delta\right)\right] = 0$$

13.1.4 Identification

$E\left[x_i z_j'\right]$ has full rank L where $\dim(z_i) = L \leq K = \dim(x_i)$

13.2 Estimator

$$\delta = E\left[x_i z_i'\right]^{-1} E\left[x_i y_i\right] = \Sigma_{xz}^{-1}\sigma_{xy}$$
$$\hat{\delta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$
$$\sqrt{n}(\hat{\delta} - \delta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{d} E\left[x_i z_i'\right]^{-1} \cdot N(0, S) \xrightarrow{d} N(0, E\left[x_i z_i'\right]^{-1} SE\left[x_i z_i'\right]^{-1}) \xrightarrow{d} N(0, E\left[x_i z_i'\right]^{-1} E\left[\varepsilon_i^2 x_i x_i'\right] E\left[x_i z_i'\right]^{-1})$$
$$S = E\left[g_i g_i'\right] = \left[(x_i\varepsilon_i)(x_i\varepsilon_i)'\right]$$

13.2.1 Case 1: exogenous error

Assume

$$E\left[\varepsilon_i \mid x_i, z_i\right] = 0$$
$$E\left[\varepsilon_i^2 \mid x_i, z_i\right] = f(x_i)$$

Then it is pretty much OLS:

$$\hat{\delta}_{\text{OLS}} \xrightarrow{p} \delta$$
$$\sqrt{n}(\hat{\delta}_{\text{OLS}} - \delta) \xrightarrow{d} N(0, V)$$
$$V = E\left[z_i z_i'\right]^{-1} E\left[\varepsilon_i^2 z_i z_i'\right] E\left[z_i z_i'\right]^{-1}$$

And for GMM estimator:

$$W = S^{-1} = E\left[\varepsilon_i^2 x_i x_i'\right]^{-1}$$
$$\sqrt{n}\left(\hat{\delta}\left(\hat{S}^{-1}\right) - \delta\right) \xrightarrow{d} N\left(0, \left(\Sigma_{xz}'S^{-1}\Sigma_{xz}\right)^{-1}\right)$$

13.2.2 Case 2: endogenous error with homoskedasticity

Assume

$$E\left[\varepsilon_i \mid z_i\right] \neq 0$$
$$E\left[\varepsilon_i \mid x_i\right] = 0$$
$$E\left[\varepsilon_i^2 \mid x_i, z_i\right] = \sigma^2$$

OLS-estimator:

$$\hat{\delta}_{\text{OLS}} \xrightarrow{p} \delta + E\left[z_i z_i'\right]^{-1} E\left[z_i\varepsilon_i\right] = \bar{\delta}$$
$$\sqrt{n}\left(\hat{\delta}_{\text{OLS}} - \bar{\delta}\right) \xrightarrow{d} N(0, V)$$
$$V = E\left[z_i z_i'\right]^{-1} E\left[u_i^2 z_i z_i'\right] E\left[z_i z_i'\right]^{-1} = E\left[z_i z_i'\right]^{-1} E\left[u_i^2 | z_i\right]$$
$$u_i = z_i'\bar{\delta} - y_i$$

Efficient GMM estimator:

$$W = S^{-1} = E\left[\varepsilon_i^2 x_i x_i'\right]^{-1}$$
$$\sqrt{n}\left(\hat{\delta}\left(\hat{S}^{-1}\right) - \delta\right) \xrightarrow{d} N\left(0, \left(\Sigma_{xz}'S^{-1}\Sigma_{xz}\right)^{-1}\right)$$

13.2.3 Case 3: endogenous error with heteroskedasticity

Assume

$$E\left[\varepsilon_i \mid z_i\right] \neq 0$$
$$E\left[\varepsilon_i \mid x_i\right] = 0$$
$$E\left[\varepsilon_i^2 \mid x_i, z_i\right] = f(x_i)$$

Then we use the GMM estimator which also works in the case of overidentification with $W = \mathbb{E}\left[x_i x_i'\right]^{-1}$:

$$\sqrt{n}(\hat{\delta}(\widehat{W}) - \delta) = \left(S_{xz}'\widehat{W}S_{xz}\right)^{-1} S_{xz}'\widehat{W}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i\right) \xrightarrow{d} \left(\Sigma_{xz}'W\Sigma_{xz}\right)^{-1} \Sigma_{xz}'W \cdot N(0, S) \xrightarrow{d} N(0, \Omega)$$
$$\Omega = \left(\Sigma_{xz}'W\Sigma_{xz}\right)^{-1} \Sigma_{xz}'WSW\Sigma_{xz} \left(\Sigma_{xz}'W\Sigma_{xz}\right)^{-1}$$

14 Notes

$$\begin{bmatrix} x \\ y \end{bmatrix}' \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{dx^2 - (b+c)xy + ay^2}{ad - bc}$$

14.1 Couples' data

$$\hat{\beta}_{2\text{SLS}} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}z_{ij}'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2} x_{ij}\varepsilon_{ij}\right)$$
$$\sqrt{n}\left(\hat{\beta}_{2\text{SLS}} - \beta\right) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{2} x_{ij}z_{ij}'\right)^{-1}}_{A} \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1 j=1}^{n}\sum_{ij}^{2} x_{ij}\right)}_{B}$$

LLN: $A \xrightarrow{p} E\left[\sum_{i=1}^{2} x_{ij}z_{ij}'\right] = C$

$$B \xrightarrow{d} N\left(0, E\left[\sum_{j=1}^{2} x_{ij}\varepsilon_{ij}\left(\sum_{j=1}^{2} x_{ij}\varepsilon_{ij}\right)'\right]\right) = D$$

CLT: $\sqrt{n}\left(\hat{\beta}_{2\text{SLS}} - \beta\right) \xrightarrow{d} N\left(0, C^{-1}DC^{-1}\right)$

14.2 Normal distribution

$$\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$$
$$\mathbb{E}\left[X^3\right] = \mu^3 + 3\mu\sigma^2$$
$$\mathbb{E}\left[X^4\right] = \mu^1 + 6\mu^2\sigma^2 + 3\sigma^1$$
$$\mathbb{E}\left[X^2 | \mu = 0\right] = \sigma^2$$
$$\mathbb{E}\left[X^3 | \mu = 0\right] = 0$$
$$\mathbb{E}\left[X^4 | \mu = 0\right] = 3\sigma^4$$

1 Some Basic Time Series Concepts

1.1 Strict Stationarity

$$f_{y_1} = f_{y_2} = f_{y_3} = \dots$$

1.2 Covariance Stationarity

$$\mu_t = \mathbb{E}(Y_t) = \mu \; ; \; \lambda_{t,k} = \lambda_k \text{ for all } t$$

1.3 Martingale (Difference)

$$\mathbb{E}\left\{Y_t \mid \Omega_{t-1}\right\} = Y_{t-1} \qquad (M)$$

$$\mathbb{E}\left\{Y_t \mid \Omega_{t-1}\right\} = 0 \qquad (MDS)$$

1.4 AR(p) process

1.4.1 VAR(1)

$$Y_t = \Phi Y_{t-1} + \varepsilon_t = \Phi^t Y_0 + \sum_{i=0}^{t-1} \Phi^i \varepsilon_{t-i}$$

$$A(L)Y_t = (I - A_1 L - \dots - A_p L^p) Y_t = \varepsilon_t$$

For covariance stationarity, all eigenvalues of Φ are less than one in modulus, thus $\Phi^t \to 0$ as $t \to \infty$. OR $|A(z)|$ has roots outside unit circle.

1.4.2 AR(p)

AR(p) as VAR(1) by companion form:

$$Z_t = \Phi Z_{t-1} + e_t \; ; \; Z_t = \left[Y_t \; \dots \; Y_{t-p+1}\right]'$$

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_p \\ 1 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$e_t = [\varepsilon_t \; 0 \; \dots \; 0]'$$

1.5 MA(q) are stationary

$$Y_t = \varepsilon_t - \dots - \theta_q \varepsilon_{t-q} \; ; \; \varepsilon_t \sim iid\left(0, \sigma^2\right)$$

$$\mathbb{E}(Y_t) = 0 \; ; \; \mathrm{Var}(Y_t) = \sigma^2 \left(1 + \sum_{i=1}^{q} \theta_i^2\right)$$

$$\mathrm{cov}\left(Y_t Y_{t-k}\right) = \sigma^2 \left(-\theta_\kappa + \sum_{j=1}^{q-k} \theta_j \theta_{k+j}\right) \quad \text{for } k \le q$$

$$\mathrm{cov}\left(Y_t Y_{t-k}\right) = 0 \text{ for } |k| > q$$

1.5.1 Invertibility restriction

MA(1): $|\theta| < 1$ or roots of $\theta(z) = (1 - \theta z)$ greater 1 in modulus. MA(q): roots of $\theta(z) = \left(1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_q z^q\right)$ greater 1 in modulus.
Need the restriction for identification as we cannot be sure if we recover θ or $\tilde\theta = \theta^{-1}$, with $\tilde\sigma^2 = \sigma^2\theta^2$.

1.6 Autocovariance generating functions

$$\lambda(z) = \sum_{j=-\infty}^{\infty} \lambda_j z^j$$

$$\lambda(z) = \sigma^2 \theta(z) \theta\left(z^{-1}\right) \quad \text{for MA(q)}$$

1.7 ARMA(p,q) models

$$\phi(L) Y_t = \theta(L) \varepsilon_t$$

$$\lambda_0 = \sigma^2 \left(1 + \frac{(\phi - \theta)^2}{1 - \phi^2}\right)$$

$$\lambda_h = \sigma^2 \left((\phi - \theta)\phi^{h-1} + \frac{(\phi - \theta)^2 \phi^h}{1 - \phi^2}\right)$$

For the ACGF transform into MA(q) representation:

$$\lambda(z) = \sigma^2 c(z) c\left(z^{-1}\right)$$

$$= \sigma^2 \phi(z)^{-1} \theta(z) \phi\left(z^{-1}\right)^{-1} \theta\left(z^{-1}\right)$$

2 The Likelihood Function for Time Series

$$f(Y_{1:T}) = f\left(Y_T \mid Y_{1:T-1}\right) f\left(Y_{1:T-1}\right)$$

$$= f\left(Y_T \mid Y_{1:T-1}\right) f\left(Y_{T-1} \mid Y_{1:T-2}\right) \cdot f\left(Y_{1:T-2}\right)$$

$$= \prod_{t=2}^{T} f\left(Y_t \mid Y_{1:t-1}\right) f(Y_1)$$

3 The Kalman Filter

3.1 The Basic Linear Model

$$y_t = A' x_t + H' \xi_t + w_t \quad ; \; \mathbb{E}\left(w_t w_t'\right) = R$$

$$\xi_t = F \xi_{t-1} + v_t \quad ; \; \mathbb{E}\left(v_t v_t'\right) = Q$$

3.2 Signal extraction and the Kalman Filter

$$y_{1:t} = \{y_i\}_{i=1}^{t}$$

$$\xi_{t|k} = \mathbb{E}\left(\xi_t \mid y_{1:k}\right)$$

$$P_{t|k} = \mathrm{var}\left(\xi_t \mid y_{1:k}\right)$$

$$\begin{bmatrix} w_t \\ v_t \end{bmatrix} \sim \text{Niid}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} R & G \\ G' & Q \end{bmatrix}\right)$$

3.2.1 Kalman Filter equations

$$\xi_{t|t-1} = F \xi_{t-1|t-1} \; (\mu_1)$$

$$y_{t|t-1} = A' x_t + H' \xi_{t|t-1} \; (\mu_2)$$

$$P_{t|t-1} = F P_{t-1|t-1} F' + Q \; (\Sigma_{11})$$

$$h_t = H' P_{t|t-1} H + R + H'G' + GH \; (\Sigma_{22})$$

$$K_t = (P_{t|t-1} H + G) h_t^{-1} \; \left(\Sigma_{12} \Sigma_{22}^{-1}\right)$$

$$\eta_t = y_t - y_{t|t-1} \; (z_2 - \mu_2)$$

$$\xi_{t|t} = \xi_{t|t-1} + K_t \eta_t \; \left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (z_2 - \mu_2)\right)$$

$$P_{t|t} = P_{t|t-1} - K_t H' (P_{t|t-1} + G)$$

$$\left(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\right)$$

If ξ_t is covariance stationary, then $\xi_{0|0} = \mathbb{E}(\xi_0) = 0$, $P_{0|0} = \mathrm{Var}(\xi_0)$.

3.3 Hamilton

$$y_t = c_k + \beta_k x_t + \epsilon_{k,t}, \quad \epsilon_{k,t} \sim N(0, \sigma_k)$$

$$\xi_{i,t-1} = P\left(s_{t-1} = i \mid \tilde y_{t-1}; \theta\right)$$

$$P\left(s_t = j \mid s_{t-1} = i\right) = P_{ij} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

$$\eta_{jt} = f\left(y_t \mid s_t = j, \tilde y_{t-1}; \theta\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(y_t - c_j - \beta_j x_t\right)^2}{2\sigma_j^2}\right]$$

3.4 Likelihood function

Gaussian density for $y_{1:T}$

$$f\left(y_{1:T}\right) = \prod_{t=1}^{T} f\left(y_t \mid y_{1:t-1}\right) \; ; \; y_{1:0} = \{\emptyset\}$$

$$f\left(y_t \mid y_{1:t-1}\right) = \left(\frac{1}{\sqrt{2\pi |h_t|}}\right)^n \exp\left(-\frac{1}{2} \eta_t' h_t^{-1} \eta_t\right)$$

Then the likelihood is:

$$f\left(Y_{1:T}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^{nT} \left(\prod_{t=1}^{T} |h_t|^{-1/2}\right)$$

$$\exp\left(-\frac{1}{2} \sum_{t=1}^{T} \left(\eta_t' h_t^{-1} \eta_t\right)\right)$$

$$\eta_t = y_t - y_{t|t-1}$$

This is the same expression as in the last section with $h_t = \sigma_{t-1}^2$ and $y_{t|t-1} = \mu_{t-1}$.

4 The Linear model with Serially Correlated Data

4.1 Asymptotics for serially correlated processes

4.1.1 Ergodicity

A process is ergodic if its elements are asymptotically independent.
Suppose $\{z_t\}$ is stationary and ergodic with $E(z_t) = \mu$. Then $T^{-1} \sum_{i=1}^{T} z_t \xrightarrow{a.s} \mu$.
If z_t is stationary and ergodic, then so is $x_t = f(z_t)$ for arbitrary function f.

4.1.2 CLT for martingale difference sequences (MDS)

Let $\{g_t\}$ be a (possibly vector-valued) mds that is stationary and ergodic with $E\left(g_t g_t'\right) = \Sigma_{gg}$.

$$\sqrt{T} \bar g = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Sigma_{gg})$$

4.2 Linear and Serially Correlated Regressors

$$y_t = x_t' \beta + \varepsilon_t$$

4.2.1 Assumptions

(2) $\{y_t, x_t\}$ is a stationary and ergodic process
(3) $\mathbb{E}(\varepsilon_t x_t) = 0$, or letting $g_t = \varepsilon_t x_t$ then $E(g_t) = 0$
(4) $\mathbb{E}\left(x_t x_t'\right) = \Sigma_{xx}$ which is non-singular
(5) $\{g_t\}$ is a mds with $E\left(g_t g_t'\right) = \Sigma_{gg}$
If in addition to (2)-(5), $\mathbb{E}\left[\left(x_{t,i} x_{t,j}\right)^2\right]$ is finite for all i and j, and let $\hat g_t = \hat\varepsilon_t x_t = (y_t - x_t \hat\beta) x_t$, and $S_{\hat g \hat g} = \frac{1}{T} \sum \hat g_t^2 = \frac{1}{T} \sum \hat\varepsilon_t^2 x_t x_t'$. Then

$$S_{\hat g \hat g} \xrightarrow{p} \Sigma_{gg}$$

4.2.2 OLS

$$\hat\beta \xrightarrow{p} \beta$$

$$\sqrt{T}(\hat\beta - \beta) \xrightarrow{d} N\left(0, \Sigma_{xx}^{-1} \Sigma_{gg} \Sigma_{xx}^{-1}\right)$$

$$\xi_W = T(R\hat\beta - r)' \left(R \hat V_{\hat\beta} R'\right)^{-1} (R\hat\beta - r)$$

$$\xi_W \Rightarrow \chi_m^2 \; ; \; \frac{\xi_W}{m} \Rightarrow F_{m,\infty}$$

AR(1) example: $y_t = \phi x_{t-1} + \varepsilon_t$ and $\sqrt{T}(\hat\phi - \phi) \Rightarrow N\left(0, \sigma^2 \Sigma_{xx}^{-1}\right)$. Then use from AR(1): $\Sigma_{xx} = \frac{\sigma^2}{1 - \phi^2}$:

$$\hat\phi \overset{a}{\sim} N\left(\phi, \frac{1}{T}\left(1 - \phi^2\right)\right)$$

4.3 Let g_t not be a MDS

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Omega)$$

$$\Omega = \sum_{j=-T+1}^{T-1} \lambda_j - \frac{1}{T} \sum_{j=1}^{T-1} j\left(\lambda_j + \lambda_{-j}\right)$$

$$\to \sum_{j=-\infty}^{\infty} \lambda_j$$

$$\Omega = \lambda(z = 1)$$

4.3.1 OLS With Serially Correlated Errors

Let $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t \Rightarrow N(0, \Omega)$. Then OLS gives

$$\sqrt{T}(\hat\beta - \beta) \Rightarrow N\left(0, \Sigma_{XX}^{-1} \Omega \Sigma_{XX}^{-1}\right)$$

4.4 HAC and HAR inference

Let $\hat V_{\hat\beta} = S_{XX}^{-1} \hat\Omega S_{XX}^{-1}$ and $\xi_W = T(\hat\beta - \beta)' \hat V_{\hat\beta}^{-1} (\hat\beta - \beta)$. If $\hat\Omega \xrightarrow{p} \Omega$, then $\hat V_{\hat\beta} \xrightarrow{p} V_{\hat\beta}$, and $\xi_W \Rightarrow \chi_k^2$.

4.4.1 Estimators for Ω

With finite sample, impossible to consistently estimate Ω for all possible sequences $\left\{\lambda_j\right\}$. Sometimes it is: Suppose $\lambda_{|j|} = 0$ for $|j| > q$ (so g_t follows an MA(q) process). Only estimate the variance and first q autocovariances. These are consistent.

4.4.2 HAC Estimators for Ω

Truncated: $\hat\Omega^{\text{Trunc}} = \sum_{j=-k}^{k} \hat\lambda_j$ with $\hat\lambda_j = T^{-1} \sum_{t=1}^{T-j} g_t g_{t+j}$

Weighted Truncated: $\hat\Omega(w) = \sum_{j=-k}^{k} w_j \hat\lambda_j$ where w_j are weights.

$$\hat\Omega^{NW} = \sum_{j=-k}^{k} w_j \hat\lambda_j \; ; \; w_{|j|} = \frac{k + 1 - |j|}{k + 1}$$

These HAC estimators yield test statistics with good size/power properties in cases when there is limited autocorrelation.

4.5 OLS and HAC vs. GLS

OLS is perfect if $\mathrm{Var}(u \mid X) = \Lambda = \sigma^2 I$. When $\Lambda \neq \sigma^2 I$, use

$$\hat\beta^{GLS} = \left(X' \Lambda^{-1} X\right)^{-1} X' \Lambda^{-1} Y$$

If Λ is unknown, use feasible GLS:

$$\hat\beta^{FGLS} = \left(X' \hat\Lambda^{-1} X\right)^{-1} X' \hat\Lambda^{-1} Y$$

$$\hat\Lambda = \Lambda(\hat\theta)$$

4.6 OLS (with HAC inference) or GLS?

$$y_t = x_t' \beta + u_t$$

$$E(u_t \mid x_t) = 0 \Rightarrow E(u_t x_t) = 0$$

$$u_t = \rho u_{t-1} + \varepsilon_t \text{ where } \varepsilon_t \overset{iid}{\sim} \left(0, \sigma^2\right)$$

$$\tilde y_t = y_t - \rho y_{t-1} \text{ and } \tilde x_t = x_t - \rho x_{t-1}$$

$$\varepsilon_t = u_t - \rho u_{t-1}$$

For GLS where we regress $\tilde y_t$ on $\tilde x_t$, we need $\mathbb{E}(\varepsilon_t \tilde x) = 0$:

$$E\left[\left(u_t - \rho u_{t-1}\right)\left(x_t - \rho x_{t-1}\right)\right]$$

$$= E(u_t x_t) + \rho^2 E\left(u_{t-1} x_{t-1}\right)$$

$$- \rho E\left(u_t x_{t-1}\right) - \rho E\left(u_{t-1} x_t\right) = 0$$

Thus the following four must hold. The first two are implied by $\mathbb{E}(u_t \mid x_t) = 0$. The others need stronger assumptions.

$$E(u_t x_t) = 0 \; ; \; E\left(u_{t-1} x_{t-1}\right) = 0$$

$$E\left(u_t x_{t-1}\right) = 0 \; ; \; E\left(u_{t-1} x_t\right) = 0$$

Exogenous or predetermined: $E\left(u_t \mid x_t, x_{t-1}, \dots\right) = 0$. Strictly exogenous: $E\left(u_t \mid \dots x_{t+1}, x_t, x_{t-1}, \dots\right) = 0$. This is needed for GLS.

5 The Functional Central Limit

5.1 Wiener Process

$W(s)$ defined on $s \in [0, 1]$.
We have $W(0) = 0$.
$W(t_i) - W\left(t_{i-1}\right) \sim \mathbb{N}\left(0, t_i - t_{i-1}\right)$ are all iid.
Thus: $W(1) \sim N(0, 1)$. And realizations of $W(s)$ are continuous with probability 1.
Suppose $\varepsilon_t \overset{iid}{\sim} N(0, 1)$, and $\xi_T(t/T) = \frac{1}{\sqrt{T}} \sum_{i=1}^{t} \varepsilon_i$ is linear interpolation between the points.

5.1.1 Theorem 1 (Weak Convergence of random functions on $C[0,1]$)

Function cannot go too crazy as T grows and at the origin.

5.1.2 Theorem 2 (CMT)

$$g : C[0, 1] \to \mathbb{R} \text{ and } \xi_T(.) \Rightarrow \xi(.)$$

$$g(\xi_T) \Rightarrow g(\xi)$$

5.1.3 Theorem 3 (Functional CLT)

Suppose ε_t is a MDS with σ_ε^2 and bounded $2 + \delta$ moments. Then any function $\xi_T(s)$ that linearly interpolates between the points $\xi(t/T) = \frac{1}{\sqrt{T}} \sum_{i=1}^{t} \varepsilon_i(t/T)$ converges in distribution to a Wiener process:

$$\xi_T \Rightarrow \sigma_\varepsilon W$$

$$\nu_T = \frac{1}{T^{3/2}} \sum_{t=1}^{T} x_t = \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{T^{1/2}} \sum_{i=1}^{t} \varepsilon_i\right]$$

$$= \sigma_\varepsilon \int_0^1 \xi_T(s) ds \Rightarrow \sigma_\varepsilon \int_0^1 W(s) ds = \nu$$

5.2 Application: Testing for a break

Null and alternative: $H_0 : \delta = 0$ vs. $H_a : \delta \neq 0$

$$y_t = \beta_t + \varepsilon_t, \text{ where } \varepsilon_t \sim iid\left(0, \sigma_\varepsilon^2\right)$$

$$\beta_t = \begin{cases} \beta \text{ for } t \le \tau \\ \beta + \delta \text{ for } t > \tau \end{cases}$$

5.2.1 Chow Test (known break)

$$\hat\delta = \bar Y_2 - \bar Y_1$$

$$\bar Y_1 = \frac{1}{\tau} \sum_{t=1}^{\tau} y_t \text{ and } \bar Y_2 = \frac{1}{T - \tau} \sum_{t=\tau+1}^{T} y_t$$

$$\hat\delta \overset{a}{\sim} \mathbb{N}\left(\delta, \sigma_\varepsilon^2 \left(\frac{1}{\tau} + \frac{1}{T - \tau}\right)\right)$$

$$\xi_W = \frac{1}{\hat\sigma_\varepsilon^2} \frac{\hat\delta^2}{\left(\frac{1}{\tau} + \frac{1}{T - \tau}\right)} \Rightarrow \xi \sim \chi_1^2$$

5.2.2 Quandt Test (unknown break)

Compute Chow statistic for many possible values of τ and use largest.

5.3 Application: Unit root AR(1) model

$\phi = 1$. Note that the following distribution is only negative, when the numerator is: $P(\phi < 0) \approx 65\%$

$$\hat\phi = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2} \; ; \; T(\hat\phi - 1) \Rightarrow \frac{\frac{1}{2}\left[X_1^2 - 1\right]}{\int_0^1 W(s)^2 ds}$$

$$t = \frac{\int_0^1 W(s) dW(s)}{\left[\int_0^1 W(s)^2 ds\right]^{\frac{1}{2}}}$$

6 VARs and Related Topics

6.1 Basic Concepts and Notation

6.1.1 VAR and MA representation

$$A(L) Y_t = \eta_t$$

$$A(L) = I - A_1 L - \dots - A_p L^p$$

Where η_t is a MDS with Σ_η. Y_t is covariance stationary. Invert $A(L)$ for MA process:

$$Y_t = C(L)\eta_t = \eta_t + C_1 \eta_{t-1} + C_2 \eta_{t-2} + \dots$$

$\eta_t = Y_t - \mathbb{E}\left(Y_t \mid Y^{t-1}\right)$ are the one-period-ahead forecast errors (or Wold shocks).

6.1.2 SVAR and SMA representation

ε_t is mds vector of STRUCTURAL shocks. Then:

$$\eta_t = H \varepsilon_t \; ; \; \mathrm{Var}(\varepsilon_t) = \Sigma_\varepsilon$$

$$B(L) Y_t = H^{-1} A(L) Y_t = \varepsilon_t \text{ (SVAR)}$$

$$Y_t = C(L) H \varepsilon_t = D(L) \varepsilon_t \text{ (SMA)}$$

6.1.3 Objects of interest

Impulse Responses Write the SMA as $Y_t = \sum_{k=0}^{\infty} D_k \varepsilon_{t-k}$.

$$SIRF_{ij,h} = \frac{\partial Y_{i,t}}{\partial \varepsilon_{j,t-h}} = \frac{\partial Y_{i,t+h}}{\partial \varepsilon_{j,t}} = D_{ij,h}$$

Forecast Error Var Decomp Suppose the structural shocks are mutually uncorrelated. Then

$$\mathrm{Var}\left(Y_{i,t+h} - E\left(Y_{i,t+h} \mid Y_t\right)\right) = \sum_{j=1}^{n} \sum_{k=0}^{h-1} D_{ij,k}^2 \sigma_{\varepsilon_j}^2$$

and the fraction that is explained by the jth shock is

$$FEVD_{ij,h} = \frac{\sum_{k=0}^{h-1} D_{ij,k}^2 \sigma_{\varepsilon_j}^2}{\sum_{j=1}^{n} \sum_{k=0}^{h-1} D_{ij,k}^2 \sigma_{\varepsilon_j}^2}$$

6.2 Invertibility

If $n_\varepsilon > n_Y$ we cannot recover H for structural shocks. Also: Can I determine ε_t from current and lagged Y.

6.3 Identification of H

We can estimate Σ_η from data and $\Sigma_\eta = H \Sigma_\varepsilon H'$. We have $n(n+1)/2$ elements in Σ_η and in Σ_ε, and n^2 in H. Thus, we have n^2 too many unknowns.

6.3.1 Restrictions

1. Uncorrelated structural shocks: make Σ_ε diagonal. Still $n(n+1)/2$ too many.
2.1 Scale normalization: drop ε_t unit: $\eta_t = H \varepsilon_t$.
2.2 Scale normalization: Set $\mathrm{Var}(\varepsilon_t) = I$ or $H_{ii} = 1$ for $i = 1, \dots, n$. Still $n(n-1)/2$ restrictions short.
3. other restrictions, e.g.: timing restriction. Set upper triangle of H to zero. I.e. ε_2 does not affect Y_1, and ε_3 not Y_1, Y_2 etc.

6.4 Local Projections

Use companion form of the VAR:

$$Z_t = \Phi Z_{t-1} + e_t \quad \text{(AR)}$$

$$Z_t = e_t + \Phi e_{t-1} + \Phi^2 e_{t-2} + \dots \quad \text{(MA)}$$

$$Z_{t+k} = \Phi^k Z_t + v_{t+k} \quad \text{(Forecast)}$$

Let $J = [I_n \; 0 \; \dots \; 0]$, then $Y_t = J Z_t$ and $e_t = J' \eta_t$. Thus

$$Y_t = \eta_t + J \Phi J' \eta_{t-1} + J \Phi^2 J' \eta_{t-2} + \dots$$

$$C_k = J \Phi^k J'$$

$$Y_{t+k} = C_k Y_t + \sum_{i=1}^{p-1} W_i Y_{t-i} + u_{t+k}$$

u_{t+k} has mean zero conditional on $(Y_t, \dots, Y_{t-p+1})$. Thus this is a regression and C_k are the coefficients on Y_t from Y_{t+k} onto $(Y_t, \dots, Y_{t-p+1})$.
While the LP estimator is inefficient relative to estimators of IRS from a VAR, the LP estimators are simple to construct, and potentially more robust to miss-specification than the VAR-plugin estimators.

6.5 Examples of Identification Schemes

6.5.1 IV instrument

Split H: H_1 is first column, and H^* the rest.

$$Y_t = H_1 \varepsilon_{1,t} + \sum_{k=1} C_k H_1 \varepsilon_{t-k} + H^* \varepsilon_t^*$$

$$+ \sum_{k=1} C_k H^* \varepsilon_{t-k}^*$$

$$\frac{\partial Y_{t+k}}{\partial \varepsilon_{1,t}} = C_k H_1$$

Use instrument z_t that is only correlated with $\varepsilon_{1,t}$:

$$\eta_t = H \varepsilon_t = H_1 \varepsilon_{1,t} + H^* \varepsilon_t^*$$

$$\mathbb{E}(\eta_t z_t) = \Sigma_{\eta z} = H \mathbb{E}(\varepsilon_t z_t) = H_1 \mathbb{E}\left(\varepsilon_{1,t} z_t\right)$$

This recovers H_1 up to a scale factor. Set $H_{1,1} = 1$ and done.

$$H_{j,1} = \frac{\mathbb{E}\left(\eta_{j,1} z_t\right)}{\mathbb{E}\left(\eta_{1,1} z_t\right)} \qquad (1)$$

6.5.2 iid shocks

Let $\varepsilon_{j,t}$ be iid over t and independent over j. If the distribution is not Gaussian, identification is possible. Look for H^{-1} that generates iid variables in ε_t.

7 Discrete Choice Models

7.1 Linear Probability Model

$$V[\varepsilon_i \mid x_i] = P(y_i = 1 \mid x_i) - P(y_i = 1 \mid x_i)^2$$

7.2 Nonlinear Approaches

$$P(y_i = 1 \mid x_i) = F(x_i'\beta)$$

$$F(\eta) = \Phi(\eta) \qquad \text{(Probit)}$$

$$F(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} \qquad \text{(Logit)}$$

7.2.1 MLE

$$\mathcal{L} = \prod_i F(x_i'\beta)^{y_i}\left(1 - F(x_i'\beta)\right)^{1-y_i}$$

$$\ln \mathcal{L} = \sum y_i \ln F(x_i'\beta) + (1-y_i)\ln\left(1 - F(x_i'\beta)\right)$$

7.3 Marginal Effects

$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_{i\ell}} = \phi(x_i'\beta)\beta_\ell \qquad \text{(Probit)}$$

$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_{i\ell}} = \frac{\exp(x_i'\beta)}{\left(1+\exp(x_i'\beta)\right)^2}\beta_\ell \qquad \text{(Logit)}$$

8 Non-Linear Least Squares

$$y_i = f(x_i, \beta) + \varepsilon_i \quad \text{with} \quad E[\varepsilon_i \mid x_i] = 0$$

$$S_n(b) = \sum_{i=1}^n (y_i - f(x_i, b))^2$$

8.1 Asymptotic Distribution

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, A^{-1}BA^{-1}\right)$$

$$A = E\left[\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)'\right]$$

$$B = E\left[\varepsilon_i^2\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)'\right]$$

$$= E\left[E\left[\varepsilon_i^2 \mid x_i\right]\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)\left(\frac{\partial f(x_i,\beta)}{\partial\beta}\right)'\right]$$

9 Quantile Regression

$$y_i = x_i'\beta + \varepsilon_i \quad \text{with} \quad P(\varepsilon_i \le 0 \mid x_i) = \alpha$$

$$\rho_\alpha(\eta) = \begin{cases} -(1-\alpha)\eta & \text{if } \eta < 0 \\ \alpha\eta & \text{if } \eta \ge 0 \end{cases}$$

$$\widehat{\beta} = \arg\min_b \frac{1}{n}\sum_{i=1}^n \rho_\alpha\left(y_i - x_i'b\right)$$

$$= \arg\min_b \sum_{i=1}^n \rho_\alpha\left(y_i - x_i'b\right)$$

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N\left(0, \alpha(1-\alpha)\Gamma^{-1}V\Gamma^{-1}\right)$$

$$V = E\left[x_i x_i'\right] \;;\; \Gamma = E\left[f_{\varepsilon|x}(0)x_i x_i'\right]$$

9.1 Quantile IV

$$y_i = x_i'\beta + u_i, \quad P(u_i \le 0 \mid z_i) = \alpha$$

$$E\left[(1-\alpha)1\left\{y_i \le x_i'\beta\right\} - \alpha 1\left\{y_i \ge x_i'\beta\right\} \mid z_i\right] = 0$$

$$E\left[\left((1-\alpha)1\left\{y_i \le x_i'\beta\right\} - \alpha 1\left\{y_i \ge x_i'\beta\right\}\right)g(z_i)\right] = 0$$

This is GMM with discontinuous objective function. Note, that if we knew β, one could try to find γ as the quantile regression estimator:

$$y_i - x_i'\beta = z_i'\gamma + u_i, \quad P(u_i \le 0 \mid z_i) = \alpha$$

$$\widehat{\beta} = \arg\min_b \widehat{\gamma}(b)'W\widehat{\gamma}(b)$$

10 Extremum Estimators

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} Q_n(\theta) = \arg\max_{\theta\in\Theta} n^{-1}Q_n(\theta)$$

$$Q_n(\theta) = \sum_{i=1}^n q(z_i, \theta) \;;\; Q(\theta) = E[q(z_i, \theta)]$$

$$0 = Q_n'(\widehat{\theta}) = \sum_{i=1}^n q'\left(z_i, \widehat{\theta}\right) \qquad \text{(FOC)}$$

Using a Taylor approximation (where $\tilde{\theta}$ lies between θ_0 and $\widehat{\theta}$), one can show that:

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -\left[\frac{1}{n}\sum_{i=1}^n q''\left(z_i, \tilde{\theta}\right)\right]^{-1} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^n q'(z_i, \theta_0)$$

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, A^{-1}V\left[q'(z_i,\theta_0)\right]A^{-1}\right)$$

$$A = E\left[q''(z_i, \theta_0)\right]$$

10.0.1 MLE

$$\sqrt{n}\left(\widehat{\theta}_{\text{MLE}} - \theta_0\right) \xrightarrow{d} N\left(0, A^{-1}V\left[q'(z_i,\theta_0)\right]A^{-1}\right)$$

$$A = E\left[q''(z_i, \theta_0)\right]$$

$$q(z_i, \theta) = \log(f(z_i, \theta))$$

$$q'(z_i, \theta) = \frac{\partial \log(f(z_i,\theta))}{\partial\theta}$$

$$q''(z_i, \theta) = \frac{\partial^2 \log(f(z_i,\theta))}{\partial\theta\partial\theta'}$$

If correctly specified:

$$-E\left[\frac{\partial^2 \log\left(f(z_i,\theta)\right)}{\partial\theta\partial\theta'}\right] = V\left[\frac{\partial \log(f(z_i,\theta))}{\partial\theta}\right] = \mathcal{I}$$

$$\sqrt{n}\left(\widehat{\theta}_{\text{MLE}} - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1}\right)$$

10.0.2 Clustering

$$Q_n(\theta) = \sum_{i=1}^n \sum_{t=1}^{T_i} q(z_{it}, \theta)$$

$$0 = Q_n'(\widehat{\theta}) = \sum_{i=1}^n \sum_{t=1}^{T_i} q'\left(z_{it}, \widehat{\theta}\right)$$

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, ABA)$$

$$A = E\left[\sum_{t=1}^{T_i} q''(z_{it}, \theta_0)\right]^{-1}$$

$$B = V\left[\left(\sum_{t=1}^{T_i} q'(z_{it}, \theta_0)\right)\right]$$

11 Generalized MoM (GMM)

$$0 = E[f(z_i, \theta_0)]$$

$$\widehat{\theta} = \arg\min_\theta\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\theta)\right)' W_n\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\theta)\right)$$

11.0.1 Asymptotics

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Sigma)$$

$$\Sigma = A^{-1}B'W_0SW_0BA^{-1}$$

$$A = E\left[\frac{\partial f(z_i,\theta_0)}{\partial\theta}\right]' W_0 E\left[\frac{\partial f(z_i,\theta_0)}{\partial\theta}\right]$$

$$B = E\left[\frac{\partial f(z_i,\theta_0)}{\partial\theta}\right]$$

$$S = V[f(z_i, \theta_0)]$$

$$\text{if } W_0 = S^{-1} \quad \text{efficient GMM}$$

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \left(G'S^{-1}G\right)^{-1}\right)$$

$$G = E\left[\frac{\partial f(z_i,\theta_0)}{\partial\theta}\right]$$

11.0.2 MoM (just identified)

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, A^{-1}S(A')^{-1}\right)$$

$$A = E\left[\frac{\partial f(z_i,\theta_0)}{\partial\theta}\right]$$

12 Sequential Estimators

$$0 = \sum_{i=1}^{n} q\left(x_i, \widehat{\theta}_1\right) \;;\; 0 = \sum_{i=1}^n r\left(x_i, \widehat{\theta}_1, \widehat{\theta}_2\right)$$

$$f(x_i, \theta) = \begin{pmatrix} q(x_i, \theta_1) \\ r(x_i, \theta_1, \theta_2) \end{pmatrix} \quad \text{GMM}$$

$$Q_1 = E\left[\frac{\partial q(\theta_{10}, \theta_{20})}{\partial\theta_1'}\right]$$

$$R_1 = E\left[\frac{\partial r(\theta_{10}, \theta_{20})}{\partial\theta_1'}\right] \;:\; R_2 = E\left[\frac{\partial r(\theta_{10}, \theta_{20})}{\partial\theta_2'}\right]$$

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\theta}_1 \\ \widehat{\theta}_2\end{pmatrix} - \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}\right)$$

$$\xrightarrow{d} N\left(0, \begin{pmatrix} Q_1^{-1}V_{11}Q_1^{-1} & \text{mess} \\ \text{mess} & \text{mess}\end{pmatrix}\right)$$

$$\sqrt{n}\left(\widehat{\theta}_2 - \theta_2\right) \xrightarrow{d} N\left(0, R_2^{-1}V_{22}R_2^{-1}\right) \quad \text{if } R_1 = 0$$

13 Treatment Effects and Selection Models

13.1 Treatment Heterogeneity

If effect only varies with observable covariates, let $\varepsilon_1 = \varepsilon_0 = \varepsilon$. If the effect is even common, additionally use $X'\beta_0 = \alpha + X'\beta_1$.

$$Y_0 = X'\beta_0 + \varepsilon_0$$

$$Y_1 = X'\beta_1 + \varepsilon_1$$

$$Y = X'\beta_0 + D\left(X'(\beta_1 - \beta_0) + \varepsilon_1 - \varepsilon_0\right) + \varepsilon_0$$

$$TE = Y_1 - Y_0 = X'(\beta_1 - \beta_0) + \varepsilon_1 - \varepsilon_0$$

Unobservable. Focus on average effect instead.

13.2 Parameters of Interest

$$E[TE] \text{ or } E[TE \mid X] \qquad \text{(ATE)}$$

$$E[TE \mid D = 1] \text{ or } E[TE \mid D = 1, X] \qquad \text{(ATET)}$$

13.2.1 Bounds

Assume Y_k for $k \in \{0,1\}$ is bounded, so $y^\ell \le Y_k \le y^u$. Then $y^\ell \le E[Y_k \mid D = 0] \le y^u$. Then we can find $E[TE] = E[Y_1 - Y_0]$ by using

$$\Pr(D=k)E[Y_k \mid D=k] + (1-\Pr(D=k))y^\ell$$

$$\le E[Y_k]$$

$$\le \Pr(D=k)E[Y_k \mid D=k] + (1-\Pr(D=k))y^u$$

13.2.2 Matching

Assume that conditional on X, (Y_1, Y_0) is independent of D, and that there are actually observations to match across treatment groups $1 > \Pr(D=1 \mid X) > 0$.

ATE

$$E[Y_1 - Y_0] = E[E[Y_1 - Y_0 \mid X]] \quad \text{(ATE)}$$

$$= E[E[Y_1 \mid X, D=1]$$

$$- E[Y_0 \mid X, D=0]]$$

ATET

- construct average for each X, and D
- difference each average across Ds
- Average the differences. Weight by appearance in $D = 1$

13.2.3 Propensity Score Matching

If (Y_1, Y_0) is independent of D conditional on X, then (Y_1, Y_0) is independent of D conditional on $P(X) = \Pr(D = 1 \mid X)$. Thus, if it is valid to match on X, then one can alternatively match on $P(X)$. Very difficult to justify from an economic perspective.

13.2.4 Differences-in-Differences Estimator

$$\widehat{\beta}_1^{\text{diff-in-diff}} = \left(\left(\bar{Y}^{\text{treat,after}} - \bar{Y}^{\text{treat,before}}\right)\right.$$

$$\left. - \left(\bar{Y}^{\text{control,after}} - \bar{Y}^{\text{control,before}}\right)\right)$$

13.3 Randomized Experiments with Imperfect Compliance

Let Z be 1 if assigned to treatment, and 0 if assigned to control. Also let D_1 be the treatment status if $Z = 1$, and D_0 the treatment status if $Z = 0$. Also, D_1, D_0 are binary. Must assume

- Independence: (Y_0, Y_1, D_0, D_1) is independent of Z (random assignment)
- First Stage: $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \ne P(D_0 = 1)$
- Monotonicity: $D_1 \ge D_0 \longrightarrow$ (no defiers)

Then we have for the compliers (Local average TE = LATE):

$$\alpha_{\text{LATE}} = E[Y_1 - Y_0 \mid D_1 > D_0]$$

$$= \frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{E[D \mid Z=1] - E[D \mid Z=0]} = \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)}$$

Effectively, Z acts as an instrument for the treatment, and one can run 2SLS of Y on a constant and D, using Z as instrument (one may include other controls X).

13.3.1 Parameter Heterogeneity

Every individual has own paremeter.

$$y_i = x_i'\beta_i + \varepsilon_i \quad E[x_i\varepsilon_i] = 0$$

$$\widehat{\beta} \xrightarrow{p} E[\beta_i]$$

Assume β_{1i} and δ_{1i} are distributed independently of (u_i, v_i, z_i). And $E[u_i \mid z_i] = 0$, $E[v_i \mid z_i] = 0$, and $E[\delta_{1i}] \ne 0$:

$$\widehat{\beta}_1^{2SLS} \xrightarrow{p} \frac{\text{cov}(y_i, z_i)}{\text{cov}(x_i z_i)} = \frac{E[\delta_{1i}\beta_{1i}]}{E[\delta_{1i}]}$$

2SLS estimates the causal effect for individuals for whom Z_i is most influential (those with large δ_{1i}).

13.4 Regression Discontinuity

$$P(D = 1 \mid X = x) = \begin{cases} 0 & \text{for } x < c \\ 1 & \text{for } x \ge c \end{cases}$$

$$E[Y \mid X = x] = \begin{cases} E[Y_0 \mid X = x] & \text{for } x < c \\ E[Y_1 \mid X = x] & \text{for } x \ge c \end{cases}$$

$$\lim_{x\searrow c} E[Y \mid X = x] - \lim_{x\nearrow c} E[Y \mid X = x]$$

$$= E[Y_1 - Y_0 \mid X = c]$$

14 Nonparametrics

14.1 Kernel Density Estimator

$$\widehat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x - x_i}{h_n}\right)$$

$$E[\widehat{f}(x)] = f(x) + \frac{1}{2}h^2 f''(x)\int v^2K(v)dv + O\left(h^4\right)$$

$$V[\widehat{f}(x)] = \frac{1}{nh}f(x)\int K(v)^2dv + O\left(n^{-1}\right)$$

14.1.1 Epanechnikov kernel

$$K_{\text{opt}}(t) = \frac{3}{4\cdot 5^{1/2}}\left(1 - \frac{1}{5}t^2\right)1\left(t^2 \le 5\right)$$

15 Machine Learning

15.1 Trees

Highly intuitive, easy to explain, highly flexible BUT hard to interpret, discrete step function (even for continuous data), and might need a lot of leaves.

Uses regression sample split algorithm:

$$Y_i = \mu_1 1\{X_{di} \le \gamma\} + \mu_2 1\{X_{di} > \gamma\} + \varepsilon_i$$

$$\mathbb{E}[\varepsilon_i \mid X_i] = 0$$

- The parameters are d, γ, μ_1, and μ_2
- d and γ are estimated by grid search
- The estimates produce a sample split
- need $N_{\min}$ for stopping criteria

15.2 Bagging (Bootstrap Aggregating)

You generate a large number B of bootstrap samples. Estimate your regression model on each bootstrap sample. The average of the bootstrap estimates is the bagging estimator.

15.3 Random Forests

Random forests are a modification of bagged regression trees. The modification is to reduce estimation variance.

1. Draw a nonparametric bootstrap sample.
2. Grow a regression tree on the bootstrap sample using m variables chose at random from the p regressors

$$\widehat{m}_{\text{rf}}(x) = B^{-1}\sum_{b=1}^B \widehat{m}_b(x)$$

15.4 Elastic Net (Ridge / Lasso)

$$y_i = \sum_{j=1}^k x_{ij}\beta_j + \varepsilon_i \text{ (many regressors)}$$

For Lasso, set $\alpha = 0$. For Ridge, set $\alpha = 1$. Get the parameters via m-fold cross validation.

$$\min_{b_j}\sum_{i=1}^n\left(y_i - \sum_{j=1}^k x_{ij}\beta_j\right)^2$$

$$+\lambda\left((1-\alpha)\sum_{j=1}^k\left|b_j\right| + \alpha\sum_{j=1}^k b_j^2\right)$$

15.5 Double Selection Lasso (IV)

Use Lasso to estimate

$$D_i = x_i'\gamma + v_i$$

Let x_1 be the selected variables.

Use Lasso to estimate

$$Y_i = x_i'\delta + v_i$$

Let x_2 be the selected variables.

Let $\tilde{x} = x_1 \cup x_2$ and regress (OLS)

$$y_i = D_i\theta + \tilde{x}_i'\beta + \varepsilon_i$$

to get the estimator of θ.

16 Notes

16.1 Binomial

$$p(k) = \binom{n}{k}p^k(1-p)^{n-k}$$

$$E[X] = np$$

$$Var(X) = np(1-p)$$

$$M(t) = (1 - p + pe^t)^n$$

Note: if $n = 1$, it's a Bernoulli distribution.

16.2 Poisson

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E[X] = \lambda$$

$$Var(X) = \lambda$$

$$M(t) = e^{\lambda(e^t - 1)}$$

16.3 Uniform

$$f(x) = \frac{1}{b-a}$$

$$E[X] = \frac{1}{2}(a+b)$$

$$Var(X) = \frac{1}{12}(b-a)^2$$

$$M(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$$

16.4 Univariate Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$

$$M(t) = e^{\mu t}e^{\frac{\sigma^2 t^2}{2}}$$

χ^2 Distribution

$$U := \sum_{i=1}^n Z_i^2$$

$$U \sim \chi_n^2$$

$$E[U] = n, \forall n \ge 1$$

$$Var(U) = 2n, \forall n \ge 1$$

t Distribution

$$T := \frac{Z}{\sqrt{U/n}}$$

$$T \sim t_n$$

$$E[T] = 0, \forall n \ge 2$$

$$Var(T) = \frac{n}{n-2}, \forall n \ge 3$$

F Distribution

$$W := \frac{U/m}{V/n}$$

$$W \sim F_{m,n}$$

$$E[W] = \frac{n}{n-2} \qquad \forall n \ge 3$$

$$Var(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \forall n \ge 5$$

Marginal Distribution

DRV	CRV
$p_x(x_i) = \sum_j p(x_i, y_j)$	$f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy$

Conditional Distribution

DRV	CRV
$P(X = x_i \mid Y = y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} = \frac{P_{XY}(x_i,y_j)}{P_Y(y_j)}$	$f_{X\mid Y}(x\mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}$

Theorems

A	Linear functions of X are normally distributed $Y = \eta + BX \sim N_k\left(\eta + B\mu, B\Sigma B'\right)$
B	X has density given by $f_X(x) = \frac{1}{(2\pi)^{p/2}\|\Sigma\|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$
C	Independent normally distributed RVs are jointly normal. $X = (X_1', X_2')' \sim N_{p+q}(\mu, \Sigma)$ $\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}$ and $\Sigma = \begin{pmatrix}\Sigma_1 & 0 \\ 0 & \Sigma_2\end{pmatrix}$
D	Conditional normal distribution $(X_1 \mid X_2 = x_2) \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$
E	Suppose $X_2 \sim N(\mu_2, \Sigma_{22})$ and $X_1 \mid X_2 = x_2 \sim N(A + Bx_2, \Omega)$. Then $X = (X_1', X_2')'$ has a multivariate normal distribution $\begin{pmatrix}X_1 \\ X_2\end{pmatrix} \sim N\left(\begin{pmatrix}A + B\mu_2 \\ \mu_2\end{pmatrix}, \begin{pmatrix}B\Sigma_{22}B' + \Omega & B\Sigma_{22} \\ \Sigma_{22}B' & \Sigma_{22}\end{pmatrix}\right)$
F	Sums of independent normals $X_1 + X_2 \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$

Let $X \sim N_p(\mu, \Sigma)$. Also let $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$, and $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$.

G	The marginal distribution of X_1 is $N_k(\mu_1, \Sigma_{11})$
H	For a normal, a zero correlation implies independence.
I	Characterizing independence of linear combinations of normal variables. If $X \sim N_p(\mu, \Sigma)$, B is a $p \times k$ matrix, and C is a $p \times m$ matrix, then $B'X$ and $C'X$ are independent iff $B'\Sigma C = 0$. Note that $B'X$ and $C'X$ are jointly normal and $B'\Sigma C$ is the covariance.

Quadratics: Assume A is symmetric, then $Y'AY$ is a quadratic form.

J	If $X \sim N_p(\mu, \Sigma)$ where Σ has rank p, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi_p^2$.
K	Let M denote an idempotent $p \times p$ matrix with rank k, then $Z'MZ \sim \chi_k^2$ $M = P\Lambda P'$, where Λ contains the eigenvalues of M on the diagonal, and the rows of P are the orthonormal eigenvectors. Then $M = [P_1\ P_2]\begin{bmatrix}I_k & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}P_1' \\ P_2'\end{bmatrix} = P_1P_1'$ Thus, $P_1'Z \sim N(0, P_1'P_1)$, where $P_1'P_1 = I_k$
L	Let $X = PZ$, and $Q = Z'AZ$, where $PA = 0$, then X and Q are independent.
M	Let $Q_1 = Z'A_1Z$, and $Q_2 = Z'A_2Z$, where $A_1A_2 = 0$. Then Q_1 and Q_2 are independent.