Introduction to Financial Econometrics

Johnew Zhang

November 20, 2017

Contents

Background

There are three parts of the class, mathematical statistics, basic tools and application. 45% problem set, 15% short project and 40% final exam.

1 Limit Theorems and Asymptotic Inference

1.1 Estimator and Finite Sample Properties

An estimator is an inference value of the parameters in the assumed distribution over a sample distribution.

Definition 1. *$\hat{\theta}$ is unbiased if $E[\hat{\theta}] = \theta$. $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$.*

Definition 2. *Let $\hat{\theta}_1, \hat{\theta}_2$ be two estimators of θ. If $E[\hat{\theta}_1] = E[\hat{\theta}_2] = \theta$. If $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.*

Definition 3. *$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$*

1.2 Inequalities and Convergence

Chebyshev Inequality: if $X > 0, P(X > x) \leq \frac{E[X]}{x}$. Another form is $P(|X - \mu| > \varepsilon) \leq \frac{V[X]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$

1.3 LLN, CLT, and Delta Method

Law of Large Number

Suppose X_i are iid random variables with $E[X_i] = \mu$ and $E[X_i^2] < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\forall \varepsilon > 0, P(|\bar{X}_n - \mu| > \varepsilon) \to 0$, as $n \to \infty$, i.e., X_n is consistent for μ.

$X_n \to X$ or X_n converges in probability to $X \in \mathbb{R}$, if $\forall \varepsilon > 0, P(|X_n - X| > \varepsilon) \to 0$, as $n \to \infty$.

Properties of $\xrightarrow{p}$

- Additive
- Scalar multiplicative
- $X_n \xrightarrow{p} X$, g is continuous $\implies g(X_n) \xrightarrow{p} g(X)$

1.3.1 o_p and O_p notation

- $X_n = o_p(b_n)$ if $\frac{X_n}{b_n} \xrightarrow{p} 0$, as $n \to \infty$.
- $X_n = O_p(b_n)$ if $\forall \varepsilon > 0, \exists M$, s.t. $P(|\frac{X_n}{b_n}| > M) < \varepsilon, \forall n$.
- $X_n = o_p(a_n), Y_n = o_p(b_n) \implies X_n Y_n = o_p(a_n b_n)$.

X_n converges to a random variable X in distribution, i.e. $X_n \xrightarrow{d} X$, if $F_n(x) \to F(x)$ at all continuous points of $F(x)$.

1.3.2 Moment Generating Function

If $g_n(t) = E[e^{tX_n}], \forall t \in \mathbb{R}$ is the moment generating function of X_n; X is a random variable with mgf $g(t) = E[e^{tX}], \forall t \in \mathbb{R}$; $g_n(t) \to g(t), \forall t \in \mathbb{R}$ as $n \to \infty$, Then $X_n \xrightarrow{d} X$.

1.3.3 Central Limit Theorem

If $X_i \sim iid$, $E[X_i] = \mu$, $Var(X_i) = \sigma^2$, then $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$.

1.3.4 Slutsky's Theorem

If $X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c \in \mathbb{R}$, then

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n \cdot Y_n \xrightarrow{d} cX$
- $X_n / Y_n \xrightarrow{d} X/c$
- $g(X_n, Y_n) \xrightarrow{d} g(X, c)$, g is continuous function (Continuous Mapping Theorem)

1.3.5 Delta Method

If $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, then $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, [g'(\mu)]^2\sigma^2)$.

Lemma 1. $o(O_p(1)) = o_p(1)$

Proof. $a_n = o(b_n), \frac{a_n}{b_n} \to 0, n \to \infty$

$z_n = o_p(1), y_n = o(z_n)$. $\forall \gamma > 0, P(|Z_n| > M_\gamma) < \gamma$. $\forall \varepsilon, P(|X_n| > \varepsilon) = P(|\frac{X_n}{z_n}||z_n| > \varepsilon) \to .0$. □

Check the slides for matrix extension.

- Unbiased does not implies consistency.
- Convergences in distribution implies asymptotic in boundedness.
- CLT implies LLN.

  Proof. If CLT, then $X_n = O_p(\frac{1}{\sqrt{n}})$. Hence $\forall \delta > 0, \exists M$, such that $P(|X_n| < \frac{M}{\sqrt{n}}) < \delta$. $\forall \varepsilon > 0$, choose n such that $\frac{M}{\sqrt{n}} < \varepsilon$, $P(|X_n| > \varepsilon) \to 0$. □

- $\xrightarrow{d}$ does not imply $\xrightarrow{p}$.

1.4 Confidence Interval

A confidence interval $C(X)$ for a parameter θ is a set of possible values which contains θ with some specified probability.

Example

Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$\bar{X}_n \sim N(\mu, \sigma^2/n) \implies \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

$$\begin{aligned} Pr(|\bar{X}_n - \mu| \leq c_n) &= Pr(\bar{X}_n - c_n \leq \mu \leq \bar{X}_n + c_n) \\ &= Pr(|\sqrt{n}(\bar{X}_n - \mu)/\sigma| \leq \sqrt{n}c_n/\sigma) = 1 - \alpha \end{aligned}$$

1. $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)$
2. $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 \implies S^2/\sigma^2 \sim \chi^2(n-1)$
3. $\bar{X}_n$ and S^2 are independent.

   $\frac{\sqrt{n}(\bar{X}_n - \ mu)}{S} \sim \sqrt{N(0,1)}\sqrt{\chi^2(n-1)} \sim$ student-t distribution with degree of freedom of $n-1$

$$\implies c_n = t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}$$

Note that the confidence interval is the finite sample properties.

1.4.1 Large Sampling

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$
$$S^2 \xrightarrow{p} \sigma^2$$

By Slutsky's Theorem,

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

Thus

$$c_n = z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

2 Linear Models

2.1 Normal Linear Models

Suppose N realizations are sampled from a large population and Scalar output is Y_i. Then we can write $X_i = (X_{i0}, \cdots, X_{iK})'$ and $\forall i, X_{i0} = 1$. We need to explain the distribution of Y_i using X_i.

Let's use a normal linear model where

$$Y_i = X_i'\beta + \varepsilon_i$$

or

$$Y = X\beta + \varepsilon$$

Note ε is anything that cannot be captured by the model (or X_i).

Assumption

1. (X_i, Y_i) are i.i.d. with the first four moments finite and $E[X_i X_i']$ full rank.
   - Empirically, econometrist uses growth rate instead of GDP to model countries' growth prospect.
   - When modelling the impact of tax credit on firms' R&D decision, we may get correlated relationships between different firms because they may in the same industry.
2. $\varepsilon_i | X_i \sim N(0, \sigma^2)$
3. $\varepsilon_i \perp X_i$ with normalization $E[\varepsilon_i] = 0$
4. $E[\varepsilon_i | X_i] = 0$ is a weaker version of the 3
5. $E[\varepsilon_i \cdot X_i] = 0$

The ordinary least squares estimator solves

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - \beta' X_i)^2$$

The solution is

$$\hat{\beta}_{ols} = (X'X)^{-1}(X'Y)$$

Now based on assumption 2, then we will get

$$\hat{\beta}_{ols} | X \sim N(\beta^2, \sigma^2 \cdot (X'X)^{-1})$$

We know the following properties about $\hat{\beta}$.

1. $\hat{\beta}$ is unbiased

2. As $n \to \infty$, $Var(\hat{\beta}_1) \to 0$.
3. $\hat{\beta}_1$ is consistent.
4. The variance of $\hat{\beta}_1$ decreases as the variance of the errors, σ^2 decreases.
5. The variance of $\hat{\beta}_1$ decreases as the variance of the X's, $\sum_{i=1}^n (X_i - \bar{X})^2$ increases

2.1.1 Estimate of σ^2

We estimate the error ε_i by the residual $\hat{\varepsilon}_i$.

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\epsilon}_i^2}{n-2}$$

where in general, the degrees of freedom associated with a set of residuals is equal to the number of observations minus the number of parameters estimated. To understand the rule's application to simple linear regression, it is helpful to think about the case when $n = 2$: the least squares estimated line connects the two points, and both residuals are zero. There are no degrees of freedom for estimating the variance; only by adding a third or more observations there is any information about variability.

2.1.2 Asymptotic Distribution of $\hat{\beta}_{ols}$

If we assume assumption 3,

$$\sqrt{N}(\hat{\beta}_{ols} - \beta) \xrightarrow{d} N(0, \sigma^2 \cdot E[X_i X_i']^{-1})$$

Proof.

$$\begin{aligned}
\hat{\beta} - \beta &= (\sum_i X_i X_i')^{-1} (\sum_i X_i Y_i) - \beta \\
&= (\sum_i X_i X_i')^{-1} (\sum_i X_i Y_i) - (\sum_i X_i X_i')^{-1} (\sum_i X_i X_i) \beta \\
&= (\sum_i X_i X_i')^{-1} [\sum_i X_i (X_i' \beta + \varepsilon_i)] - (\sum_i X_i X_i')^{-1} (\sum_i X_i X_i) \beta \\
&= (\sum_i X_i X_i')^{-1} \sum_i X_i \varepsilon_i \\
&= (\frac{1}{n} \sum_i X_i X_i')^{-1} (\frac{1}{n} \sum_i X_i \varepsilon_i) \\
&\xrightarrow{p} (\frac{1}{n} E[X_i X_i'])^{-1} N(E[X_i \varepsilon_i], Var(X_i \varepsilon_i)) \\
&= (\frac{1}{n} E[X_i X_i'])^{-1} \sqrt{n} N(0, E[X_i X_i'] \sigma^2) \\
\sqrt{n}(\hat{\beta} - \beta) &\xrightarrow{d} N(0, E[X_i X_i']^{-1} \sigma^2)
\end{aligned}$$

□

We will find that $\hat{\sigma}^2 = \frac{1}{N-K-1} \sum_i (Y_i - X_i' \hat{\beta}_{ols})^2$ is unbiased

2.1.3 Construct Confidence Interval

Hence the confidence interval for the $\hat{\beta}_{ols}$ is the following

$$Cl_{0.95} = (\hat{\beta}_k - 1.96\sqrt{\hat{V}_{kk}}, \hat{\beta}_k + 1.96\sqrt{\hat{V}_{kk}})$$

Testing whether $\beta_k = 0.1$. Then compare $t = \frac{\beta_k - .1}{\sqrt{\hat{V}_{kk}}}$ against 1.645 assuming 10% significant value.

2.1.4 Conditional Expectation and the Best Linear Predictor

What is the difference between assumption 4 and assumption 5? Let's start with assumption 4, then if $Y_i = E[Y_i|X_i] + \varepsilon_i$, $E[Y_i|X_i]$ is an unknown function of X_i, then $E[\varepsilon_i|X_i] = 0$ or $E[\varepsilon_i h(X_i)] = 0, \forall h$. Then under assumption 4, we know the correct functional form of the conditional expectation $E[Y_i|X_i]$.

Regarding assumption 5, $\hat{\beta}_{ols}$ is the best linear predictor. "Best" in the sense that it minimizes the mean squared error. Assumption 5 does not require that we know the correct functional form of $E[Y_i|X_i]$. If $E[Y_i|X_i]$ is non-linear, $\hat{\beta}_{ols}$ provides the best linear approximation of $E[Y_i|X_i]$.

2.2 Robust Variances

Under assumption 5, $E[\varepsilon_i \cdot X_i] = 0$, then

$$\sqrt{N}(\hat{\beta}_{ols} - \beta) \xrightarrow{d} N(0, V_{robust})$$

where

$$V_{robust} = E[X_i X_i']^{-1} E[\varepsilon_i^2 X_i X_i'] E[X_i X_i']^{-1}$$

Note robust variance allows ε_i can be dependent on X_i.

2.3 Bootstrap

Alternative way to estimate the variance of $\hat{\beta}_{ols}$ through bootstrapping. In other words, resample from the population distribution which can be approximated by the sample distribution. Suppose we have a random sample of size N from a distribution with cdf of $F_X(x)$. Suppose we know the cdf, then we can replace all expectations with integrals

$$\mu = \int x dF(x), \sigma^2 = \int (x - \mu)^2 dF(x)$$

Therefore,

$$\hat{V}(\bar{X}) = \frac{1}{N} \int (z - \int x dF(x))^2 dF(z)$$

We can estimate $F(x)$ using the empirical distribution function. $\hat{F}(x) = \sum_i 1_{x_i < x}/N$ where 1_A is the indicator function for the event A. Hence we can get

$$\tilde{\mu} = \frac{1}{N} \sum_i X_i$$

and

$$\tilde{\sigma}^2 = \sum_i (X_i - \bar{X})^2 / N$$

or

$$\tilde{V}(\bar{X}) = S^2(N-1)/N^2$$

which is close to the standard estimate S^2/N.

We can do this in much more complex settings. Best linear predictor: $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$. Given sample $N, \{(Y_i, X_i)\}_{i=1}^N, \hat{\beta}_{ols} = (\sum_i X_i X_i')^{-1} (\sum_i X_i Y_i)$. We can resample at random with replacement (Y_i, X_i), to get new sample $\{(X_{bj}, Y_{bj})\}_{j=1}^N$. Then we can compute our mean and variance.

2.3.1 Parametric Bootstrap

Instead of bootstrapping pairs (Y_i, X_i), we can bootstrap the residuals $\hat{\varepsilon}_i = Y_i - X_i' \hat{\beta}_{ols}$.

2.3.2 Jackknife

For each i, calculate the estimator leaving out the i^{th} observation $\bar{X}_{(i)} = \frac{1}{N-1}\sum_{j\neq i} X_j$. With those N new estimates μ, construct the variance estimate $\hat{V}(\bar{X}) = \frac{N-1}{N}\sum_i(\bar{X}_{(i)} - \bar{X})^2$

$$V(\bar{X}_{(i)} - \bar{X}) = \sigma^2(\frac{1}{N^2(N-1)} + \frac{1}{N^2}) \approx \sigma^2/N^2$$
$$E(\hat{V}(\bar{X})) = NE(\bar{X}_{(i)} - \bar{X})^2 \approx \sigma^2/N$$

To use Jackknife is to reduce potential computation complexity arising from using bootstrap since the B would be potentially way larger than N.

Example: NLSY

Using log transformation on earnings is to make sure prediction on earnings is positive. Using second order polynomial on work experience is because that the marginal returns of work experience should be diminishing.

2.3.3 Delta Method

Often we are interested in the average predicted value of the dependent variable given the independent variables. They can be complicated functions of the parameters. We use the delta method to compute the standard error for the predicted values.

What is the interpretation of $\hat{\beta}_{ols}$? This can be explained by the residual regression.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

$$X_k = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{k-1} X_{k-1} + u$$
$$\implies \beta_k = \frac{Cov(u, Y)}{Var(u)}$$

The residual variation of X_k is the β_k capturing.

What happens when we fail to include every explanatory variable? This might lead to biased estimation of the parametric estimation. When we are looking at non-linear cases, we can use delta method as well to estimate.

3 Clustering

Independence is a strong assumption and often may not hold so we need to relax this condition by clustered pairs of (Y_i, X_i). Inference: number of clusters $\to \infty$, number of observations within each cluster is fixed. Often the number of observations per group might be large compared to the number of clusters but increasing the number of observations does not give us consistent estimate. We will see the dependence structure greatly affects the standard errors.

Let Z be the $N \times S$ cluster indicator with element $Z_{is} = \mathbf{1}_{S_i = s}$.

Let L_N be the N dimensional vector with all elements equal to 1. $Z'L_N$ gives a S vector with the s^{th} element equal to N_s, the size of cluster s. $Z'Z$ be the $S \times S$ diagonal matrix with the s^{th} on diagonal element equal to the size of cluster s.

3.1 Homoskedastic Case

Consider the linear model $Y_i = X_i\beta + \varepsilon_i$.

$$E[\varepsilon\varepsilon'|X, Z) = \Omega = \sigma_\varepsilon^2((1-\rho)\cdot I_N + \rho \cdot ZZ')$$

where I_N is the $N \times N$ identify matrix.

In this model,

$$E[\varepsilon_i \varepsilon_j] = \begin{cases} \sigma_\varepsilon^2 & i = j \\ \rho \cdot \sigma_\varepsilon^2 & S_i = S_j, i \neq j \\ 0 & S_i \neq S_j \end{cases}$$

An alternative way to think about the error component structure is a "random effect" model:

$$E[\varepsilon_i \varepsilon_j] = \begin{cases} \sigma_\eta^2 + \sigma_\nu^2 & i = j \\ \sigma_\nu^2 & S_i = S_j, i \neq j \\ 0 & S_i \neq S_j \end{cases}$$

In general, the variance of the OLS estimator is

$$V(\hat\beta_{ols}) = (X'X)a^{-1} \cdot (X'\Omega X) \cdot (X'X)^{-1}$$

In the first model, we have

$$V(\hat\beta_{ols}) = V_{ols} \cdot [I_{K+1} + \rho \cdot (X'ZZ'X(X'X)^{-1} - I_{K+1}]$$

In the random effect model, we have

$$V(\hat\beta_{ols}) = \sigma_\eta^2 (X'X)^{-1} + \sigma_\nu^2 (X'X)^{-1}(X'ZZ'X)(X'X)^{-1}$$

Define $\tilde\varepsilon = (I_N - Z(Z'Z)^{-1}Z')\hat\varepsilon$ where $\hat\varepsilon$ is the residual of regressing $\hat\varepsilon$ on Z. Hence the variance estimator of $\tilde\varepsilon$ is $\tilde\sigma_\varepsilon^2 = \tilde\varepsilon'\tilde\varepsilon/(N-S-K-1)$

$$\begin{aligned} E[\tilde\varepsilon'\tilde\varepsilon/(N-S-K-1)] &= tr(E[\tilde\varepsilon\tilde\varepsilon'])/(N-S-K-1) \\ &= tr(E[I_N - Z(Z'Z)^{-1}Z'\hat\varepsilon\hat\varepsilon'(I_N - Z(Z'Z)^{-1}Z')])/(N-K-S-1) \\ &\approx tr(E[I_N - Z(Z'Z)^{-1}Z'\varepsilon\varepsilon'(I_N - Z(Z'Z)^{-1}Z')])/(N-K-S-1) \\ &= tr(I_N - Z(Z'Z)^{-1}Z')\sigma^2[(1-\rho)I_N + \rho ZZ'](I_N - Z(Z'Z)^{-1}Z')/(N-K-S-1) \\ &= \sigma^2(1-\rho)tr(I_N - Z(Z'Z)^{-1}Z')/(N-K-S-1) \\ &= \sigma^2(1-\rho)(N-S)/(N-S-K-1) \end{aligned}$$

Therefore, $\tilde\sigma_\varepsilon^2$ estimates $\sigma_\varepsilon^2(1-\rho)$. and $\hat\rho = \frac{\hat\sigma_\varepsilon^2 - \tilde\sigma_\varepsilon^2}{\hat\sigma_\varepsilon^2}$.

4 Maximum Likelihood

The estimator of the maximum likelihood for σ^2 is slightly different

$$\hat\sigma_{mle}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i\beta)^2$$

What is a maximum likelihood estimator? Consider an i.i.d. random sample $X_1, \cdots, X_n$ from $f(x|\theta)$. We observe data $X_i = x_i, i = 1, \cdots n$. The likelihood function is the joint density of the data as a function of the parameters, here θ

$$L(\theta|X_1 = x_1, \cdots, X_n = x_n) = Pr(X_1 = x_1, \cdots, X_n = x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

The Maximum Likelihood Estimator for θ is defined by $\hat\theta = \arg\max_\theta L(\theta|X = x)$

In general, it is easier to solve the log of likelihood function since it generally converts the likelihood functions into a sum instead of multiplication. Computationally, it is much more straightforward to solve.

Is the MLE consistent? Let $Q_n(\theta) = \frac{1}{n}\sum_i \ln f(X_i|\theta)$ and $Q_0(\theta) = E[\ln f(X_i|\theta)]$. Then we can show that $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$ and it is maximized at $\hat\theta_{mle}$. In addition, we know that $Q_0(\theta)$ is maximized at the true parameter value θ_0. Then $\hat\theta_{mle} \xrightarrow{p} Q_0(\theta)$.

In addition, we should expect an information inequality meaning $E[\ln f(X|\theta_0)] \geq E[\ln f(X|\theta)]$.

Proof. We can show the following

$$\begin{aligned}E[\ln f(X|\theta_0)] - E[\ln f(X|\theta)] = E[\ln \frac{f(X|\theta_0)}{f(X|\theta)}] = E[-\ln \frac{f(X|\theta)}{f(X|\theta_0)}] &\geq \ln(E[\frac{f(X|\theta)}{f(X|\theta_0)}]) \\ &\geq -\ln \int \frac{f(X|\theta)}{f(X|\theta_0)} f(X|\theta_0)dx \\ &\geq -\ln \int f(X|\theta)dx \geq 0\end{aligned}$$

Jensen's Inequality: if $g(z)$ is convex, $E[g(z)] \geq g(E[z])$. Then

$$\begin{aligned}g(z) &= g(E[z]) + g'(E[z])(z - E[z]) + g''(\tilde{z})(z - E[z])^2 \\ &\geq g(E[z]) + g'(E[z])(z - E[z]) \\ E[g(z)] &\geq g(E[z]) + g'(E[z])(z - E[z]) \geq g(E[z])\end{aligned}$$

□

The inequality is strict if $Pr\{f(X|\theta) \neq f(X|\theta_0)\} > 0$ for all $\theta \neq \theta_0$, which means that there is no other θ giving the same distribution as θ_0 (identifiability).

4.1 Asymptotic Distribution of MLE

The score function is defined as

$$S(\theta) = \frac{\partial}{\partial\theta} \ln f(X_i|\theta)$$

Define $H(\theta)$ as the derivative of score function and information matrix

$$I(\theta) = E\left[\frac{\partial}{\partial\theta} \ln f(X_i|\theta) \cdot \frac{\partial}{\partial\theta} \ln f(X_i|\theta)'\right]$$

4.1.1 Properties of the Score Function

- $E[S(X_i, \theta_0)] = 0$.

$$\begin{aligned}E[S(X_i, \theta_0)] &= \int \frac{\partial}{\partial\theta} \ln f(X, \theta)|_{\theta=\theta_0} f(X, \theta_0)dx \\ &= \int \frac{1}{f(X, \theta_0)} \frac{\partial f(X, \theta)}{\partial\theta}|_{\theta=\theta_0} f(X, \theta_0)dx \\ &= \frac{\partial}{\partial\theta} \int f(X, \theta)dx|_{\theta=\theta_0} = 0\end{aligned}$$

- $I(\theta_0) = -H(\theta_0)$.

- $\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \ln f(X_i|\theta) \implies \frac{1}{n} \sum_i \frac{\partial \ln f(X_i|\theta)}{\partial\theta} = 0$

$$0 = \frac{1}{\sqrt{n}} \sum S_i(\theta_0) + \frac{1}{n} \sum \frac{\partial S_i(\tilde{\theta})}{\partial\theta} \sqrt{n}(\hat{\theta}_{mle} - \theta_0)$$

- $\sqrt{n}(\hat{\theta}_{mle} - \theta_0) \xrightarrow{d} N(0, H^{-1}IH^{-1})$

$$Var(S_i(\theta))|_{\theta=\theta_0} = E[S_i(\theta)S_i'(\theta)] - [E[S_i(\theta)]]^2|_{\theta=\theta_0} = E[S_i(\theta)S_i'(\theta)]$$

$$\frac{1}{n} \sum \frac{\partial S(\tilde{\theta})}{\partial\theta} \xrightarrow{p} E[\frac{\partial S(\tilde{\theta})}{\partial\theta}] = H(\tilde{\theta}) \to H(\theta_0)$$

4.1.2 Variance Estimation

The asymptotic variance $\hat{V} = -H^{-1}$ depends on the hessian of the likelihood function at θ_0, which is a measure of the curvature of the likelihood at its maximum.

4.1.3 Cramer Rao

If $\hat{\theta} = g(X_1, \cdots, X_n)$ is an unbiased estimator of θ, then

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information defined by

$$I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} \ln f(X_i|\theta)]$$

Proof. Let $\hat{\theta} = g(x)$ be unbiased. Then

$$\int g(x) f(x, \theta) dx = \theta$$

$$\int g(x) \frac{\partial f(x, \theta)}{\partial \theta} dx = 1$$

$$S(x, \theta) \implies \int g(x) [\sum S(x_i, \theta)] f(x, \theta) dx = 1$$

$$E[g(x) S(x, \theta)] = 1$$

$$Cov(g(x), S(x, \theta)) = E[g(x) S(x, \theta)] - E[g(x)] E[S(x, \theta)] = 1$$

We know

$$Var(g(x)) \geq \frac{1}{Var(S(x, \theta))} \geq \left[E[S(x, \theta) S'(x, \theta) - E[S(x, \theta)^2]] \right]^{-1} \geq E[SS']^{-1} = [nI(\theta)]^{-1}$$

□

4.2 Duration Method

The duration model is initiated to model the unemployment by Lancaster (1979). The economic theory underlying Lancaster's analysis is job search theory. An unemployed individual is assumed to receive job offers, arriving according to some rate $\lambda(t)$, such that the expected number of job offers arriving in a short interval of length dt is $\lambda(t)dt$. Each offer consists of some wage rate w, drawn independently of previous wages, from some distribution with distribution function $F(w)$. The offer is compared to some reservation wage $\bar{w}(t)$, and if the offer is better than the reservation wage, that is with the probability $1 - F(\bar{w}(t))$, the offer is accepted. The reservation wage is set to maximize utility. Suppose the arrival rate is constant over time. In that case, the optimal reservation wage is also constant over time, and the probability of receiving an acceptable offer in a short period tof time dt is θdt, with $\theta = \lambda(1 - F(\bar{w}(t)))$.

The constant acceptance rate implies that the distribution for the unemployment duration is exponential with mean $1/\theta$ and probability density function $f(y) = \theta \exp(-y\theta)$. This distribution widely used for durations fo various types. The mean and variance for the exponential distribution are $1/\theta$ and $1/\theta^2$ respectively. The expected value of the remaining duration conditional on survival up to c is $E[Y - c|Y > c] = \frac{1}{\theta}$; this does not depend on the elapsed duration .This is known as the lack of memory property.

Let Y_i be the ith man's unemployment spell and X_i be ith characteristics. Assume the conditional density of Y_i given X_i is exponential with arrival or hazard rate $\theta = \exp(\beta_0 + \beta_1 X)$. The conditional density $f(y|x, \beta_0, \beta_1) = h(y) \cdot S(y) = \exp(\beta_0 + \beta_1 x) \exp(-y \exp(\beta_0 + \beta_1 x))$. This is an extension of the exponential distribution allowing the arrival rate to depend on covariates. The log likelihood function is

$$L(\beta_0, \beta_1) = \sum_{i=1}^{N} \log f(y_i|x_i, \beta_0, \beta_1) = \sum_{i=1}^{N} \beta_0 + \beta_1 x_i - y_i \exp(\beta_0 + \beta_1 x)$$

. The maximum likelihood estimator chooses the $\theta = (\beta_0, \beta_1)$ that maximizes the log likelihood function.

5 Sampling

For a random sample of people entering unemployment, we observe the exact unemployment duration y_i. However, this may not always the case. Suppose we observe a number of people all becoming unemployed at the same time but we only observe if they exited unemployment, d_i, be fixed point in time c. Hence our likelihood function has to incorporate the survival.

$$L(\theta) = \prod_i F(c|\theta)^{d_i} \cdot (1 - F(c|\theta)^{(1-d_i)} = \prod_i S(c|\theta)^{(1-d_i)}(1 - S(c|\theta))^{d_i}$$

Suppose we know the exact exit time if occurs before c but only an indicator if after c, then the likelihood function is as following

$$L(\theta) = \prod_i f(y_i|\theta)^{d_i} \cdot S(c|\theta)^{(1-d_i)} = \prod_i h(y_i|\theta)^{d_i} \cdot S(y_i|\theta)^{d_i} \cdot S(c|\theta)^{(1-d_i)}$$

or

$$L(\theta) = \prod_i h(t_i|\theta)^{d_i} \cdot S(t_i|\theta)$$

where $t_i = \min(y_i, c)$.

By solving the MLE, we get $\hat{\theta}_{MLE} = \bar{d}/\bar{t}$. If we ignore the censoring and assume the ones censored at c exited at c, this leads to an over-estimate of θ, $\tilde{\theta} = 1/\bar{t}$. If we discard the observations that did not exit before c, this also leads to an over-estimate of θ, $\bar{\theta} = (\sum_i d_i)/(\sum_i d_i \cdot t_i)$.

5.1 Stock Sampling

So far we have assumed that we start observing the individuals at the time of entering the unemployment, which is called flow sampling. If we sample from the stock of unemployed, i has been unemployed for $s_i = 4$ and finds a job after 6 weeks, i.e., $y_i = 10$. In this case, let s_i be the incomplete duration of unemployment spell for i before he or she was first observed in the sample

$$L(\theta) = \prod_i f(y_i|\theta/S(s_i|\theta) = \prod_i h(y_i|\theta) \cdot \frac{S(y_i|\theta)}{S(s_i|\theta)}$$

However we will need to use numerical methods to find the maximum likelihood estimator.

One can do a regression on this instead of the maximum likelihood estimation. It is easy to show that the MLE estimator has lower variance than the one of OLS. Using least squares instead of maximum likelihood in this case is equivalent to throwing away 40% of the observations.

6 Hypothesis Testing

So far we have assumed that the hazard function is constant over time, $h(y|x, \theta) = \exp(x'\beta)$. This means that for individual i, the chance of finding a job on the first day of unemployment is the same as that of the 50th day conditional on that i has been unsuccessful finding a job in the first 49 days. We let the hazard function vary with y by using the Weibull distribution

$$f(y|x, \beta, \alpha) = (\alpha + 1) \cdot y^{\alpha} \exp(x'\beta) \exp(-y^{\alpha+1} \exp(x'\beta))$$

We can test if constant hazard rate over time is a reasonable assumption.

$$H_0 : \alpha = 0 \text{ vs } H_1 : \alpha \neq 0$$

Consider three classical tests: likelihood ratio, the Wald, and the Rao's score test. Next we introduce general notation and specific formula for testing the scale parameter in the Weibull model.

Suppose we have a model for random variable Z with the density function $f(z|\theta_0, \theta_1)$ where we split the parameter θ into two parts. The dimension of the full parameter vector θ is K while the dimension of θ_0 and θ_1 is L_0 and K_1. Note here we only model the conditional distribution of $Y_i|X_i$, but we can think of the marginal distribution of X_i as known.

6.1 Likelihood Ratio Test

Now, Let $\hat{\theta}_u$ denote the unrestricted MLE and $\hat{\theta}_r$ denote the estimator based on the restricted model such that $\hat{\theta}_{1r} = \arg\max_{\theta_1} L(0, \theta_1)$. The likelihood ratio test statistic is

$$LR = 2(L(\hat{\theta}_{u0}, \hat{\theta}_{u1}) - L(0, \hat{\theta}_{r1}))$$

where $LR \sim \chi^2(K_0)$ under H_0.

Why is it reasonable test statistics?

$$\begin{aligned} L(\theta) &\approx L(\hat{\theta}) + \frac{\partial L}{\partial \theta}(\hat{\theta})(\theta - \hat{\theta}) + 0.5\frac{\partial^2 L}{\partial \theta^2}(\hat{\theta})(\theta - \hat{\theta})^2 \\ LR = 2(L(\hat{\theta}) - L(\theta)) &\approx -\frac{\partial^2 L}{\partial \theta^2}(\hat{\theta})(\theta - \hat{\theta})^2 \approx NI(\hat{\theta})(\theta - \hat{\theta})^2 \sim \chi^2(K_0) \\ \sqrt{N}(\theta - \hat{\theta}) \xrightarrow{d} N(0, I^{-1}(\theta)) \end{aligned}$$

Here K_0 is the dimension of the θ_0.

6.2 Lagrange Multiplier Test

If the limiting log likelihood function is maximized at $\theta_0 = 0$. The Lagrange multiplier test statistic is defined as

$$LM = \frac{1}{N}\sum S(z_0, 0, \theta_{1r})' \cdot \hat{I}^{-1}\sum S(z_i, 0, \theta_{1r}) \sim \chi^2(K_0)$$

under H_0.

$$\begin{aligned} \frac{\partial L}{\partial \theta}(\hat{\theta}) &\approx \frac{\partial L}{\partial \theta}(\theta) + \frac{\partial^2 L}{\partial \theta^2}(\theta)(\hat{\theta} - \theta) = 0 \\ (\hat{\theta} - \theta) &= -\frac{\partial^2 L}{\partial \theta^2}(\theta)^{-1}\frac{\partial L}{\partial \theta}(\theta) \\ \sqrt{n}(\hat{\theta} - \theta) &= I^{-1}(\theta)\frac{1}{\sqrt{n}}\sum S(z_i, \theta) \to N(0, I^{-1}(\theta)) \\ \frac{1}{n}\sum(S(z_i, \theta))'I^{-1}(\theta)\sum S(z_i, \theta) \sim \chi^2(K_0) \end{aligned}$$

There is a small advantage of LM test because it avoids calculating the unrestricted estimator which can be complicated.

6.3 Wald Test

If the null hypothesis is correct, $\hat{\theta}_{u0}$ should be close to 0. Let the inverse of the information matrix be

$$I^{-1} = \begin{pmatrix} l^{00} & l^{01} \\ l^{10} & l^{11} \end{pmatrix}$$

The Wald test is defined as

$$W = N\hat{\theta}_{u0}(\hat{l}^{00})^{-1}\hat{\theta}_{u0} \sim \chi^2(K_0)$$

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &\xrightarrow{d} N(0, I^{-1}) \\ \sqrt{n}\begin{pmatrix} \hat{\theta}_{0u} - \hat{\theta}_{0r} \\ \hat{\theta}_{1u} - \hat{\theta}_{1r} \end{pmatrix} &\xrightarrow{d} N(0, \begin{pmatrix} l^{00} & l^{01} \\ l^{10} & l^{11} \end{pmatrix}) \\ \sqrt{n}(\hat{\theta}_{0u} - 0) &\xrightarrow{d} N(0, I^{00}) \\ N\hat{\theta}_{u0}(\hat{l}^{00})^{-1}\hat{\theta}_{u0} &\sim \chi^2(K_0) \end{aligned}$$

7 Bayesian Inference

We have a random variable X which has probably density or probability mass function $f(x, \theta)$. θ is an unknown parameter and we would like to know what the plausible values are. In other words, we believe there are some prior and our experience influencing the outcome of our belief, posterior distribution. Suppose θ is the prior and the we can write the posterior distribution using Bayes' theorem

$$f_{\theta|X}(\theta|x) = \frac{f_{X|\theta}(x|\theta) \cdot f_\theta(\theta)}{\int f_{X|\theta}(x|\theta) \cdot f_\theta(\theta) d\theta}$$

When assuming observations are independent, then the posterior distribution can be written as proportional of likelihood function such as

$$p(\theta|X_1, X_2, \cdots, X_N) \propto L(\theta|X_1, \cdots, X_N) \cdot p(\theta)$$

7.1 Berstein-Von Mises theorem

Suppose we have N observations. In general, the posterior distribution of θ is approximately

$$\sqrt{N}(\theta - \theta_{mle})|X_1, \cdots, X_N \sim N(0, I^{-1})$$

when N is large.

7.2 Discrete Choice Models

We are interested in models where the dependent variable is discrete.

Let's consider the decision to go to college. Linear model is $Y_i = X_i'\beta + \varepsilon_i$. Model has to have heteroskedasticity. Fit using a sigmoid function.

8 Generalized Methods of Moment

In GMM estimation we are going to look at estimation problems of the following type. We are given a function $\psi(\cdot)$, such $E[\psi(Z, \theta)] = 0$ at some unknown true value θ_0, and this expectation differs from zero for all other values θ. The vector $\psi(Z, \theta)$ is a known function of the observed random variable Z and the unknown parameter θ. ψ is the moment function and has dimension M. θ is the K dimensional parameter of interest. We are interested in two cases: $M = K$ the just identified case, and $M > K$ the over identified case. If $M < K$, typically we can't point identify the parameters but can derive bounds on them.

8.1 Examples

8.1.1 Maximum Likelihood

$$\max_\theta \sum_i \log f(Y_i|X_{i,\theta}) = \hat{\theta}_{MLE}$$

Thus

$$E[S(y, x, \theta) = E\left[\frac{\partial \log f}{\partial \theta}(y|x, \theta)\right] = 0|_{\theta=\theta_0}$$

$$\psi(y, x, \theta) = S(y, x, \theta)$$

8.1.2 Instrumental Variables

Consider

$$Y_i = X_i'\beta + \varepsilon$$

such that $E[X_i\varepsilon] \neq 0$

Here

$$E[Z_i(Y_i - X_i'\beta)] = 0$$

where Z_i satisfies $E[Z_i\varepsilon_i] = 0$ and $E[X_iZ_i] \neq 0$ then $\psi(X, Y, Z, \beta) = Z(Y - X'\beta)$

8.1.3 Panel Data with fixed effects

$$Y_{it} = X_{it}'\beta + \alpha_i + \varepsilon_{it}, t = 1, 2$$
$$E[X_i(\alpha_i + \varepsilon_{it})] \neq 0$$

since $E[X_i\alpha_i] \neq 0$

$$Y_{i2} - Y_{i1} = (X_{i2} - X_{i1})'\beta + (\varepsilon_{i2} - \varepsilon_{i1})$$

where $X_{it} \perp \varepsilon_{is}$

$$E[h(X_{i1}, X_{i2})(Y_{i2} - Y_{i1} - (X_{i2} - X_{i1})'\beta)] = 0$$

8.1.4 Regression with estimated regressors

$$Y_i = X_i'\beta + \gamma(w_i - E[w_i|Z_i]) + \varepsilon_i$$

where $w_i = Z_i\delta + \eta_i$

- $\hat{\delta} = (\sum_i Z_i Z_i')(\sum_i Z_i w_i')$
- $\hat{\eta} = w_i - Z_i'\hat{\delta}$
- $Y_i = X_i'\beta + \gamma\hat{\eta}_i + \varepsilon_i$

Then we can consider the following ψ

$$\psi_1(y, x, w, z, \delta, \beta, \gamma) = z(w - z'\delta)$$
$$\psi_2(y, x, w, z, \delta, \beta, \gamma) = \begin{pmatrix} x \\ w - z'\delta \end{pmatrix} (y - x'\beta - (w - z'\delta)x)$$

8.2 GMM: Example I

- $$E[Y_{i1} - X_i'\beta]^2 = E[w_i + \varepsilon_{i1} + \eta_{i1}]^2 = \sigma_w^2 + \sigma_\eta^2\sigma_\nu/(1 - \alpha^2)$$
  $$\psi_1 = (y_{i1} - X_i'\beta) - \sigma_w^2 - \sigma_\nu^2/(1 - \alpha^2) - \sigma_\eta^2$$

- $$E[(Y_{it} - X_i'\beta)(Y_{it+s} - X_i'\beta)] = \sigma_w^2 + \sigma_\eta^2 + \alpha^s\sigma_\nu^2/(1 - \alpha^2)$$
  Hence
  $$\psi_2 = (y_{it} - x_i'\beta)(y_{it+s} - x_i'\beta) - (\sigma_w^2 + \sigma_\eta^2 + \alpha^s\sigma_\nu^2/(1 - \alpha^2))$$

8.3 GMM Example II

$$E_t\left[\frac{\beta\frac{\partial U}{\partial c}(c_{t+1}, \gamma)}{\frac{\partial U}{\partial c}(c_t, \gamma)}\frac{p_{t+1} + d_{t+1}}{p_t} - 1\right] = 0$$

8.4 GMM: consistency

Consider the identified case and the estimator

$$\frac{1}{N}\sum_{i=1}^{N}\psi(Z_i,\hat{\theta})=0$$

Under a set of regularity conditions, $\hat{\theta}$ is a consistent estimator of θ_0. That is

$$\sqrt{N}(\hat{\theta}-\theta_0)\overset{d}{\to}N(0,(\Gamma')^{-1}\Delta\Gamma^{-1})$$

where

$$\Gamma=E\left[\frac{\partial\psi'}{\partial\theta}(Z_i,\theta_0)\right]$$

and

$$\Delta=E[\psi(Z,\theta_0)\cdot\psi(Z,\theta_0)']$$

Consider the following

$$\begin{aligned}0&=\frac{1}{N}\sum_{i=1}^{N}\psi(Z_i,\hat{\theta})\\&=\sum_{i=1}^{N}\psi(Z_i,\theta)+\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi'}{\partial\theta}(Z_i,\hat{\theta})(\hat{\theta}-\theta_0)\\\sqrt{N}(\hat{\theta}-\theta)&\approx-(\Gamma')^{-1}\sum_{i=1}^{N}\psi(Z_i,\theta)\end{aligned}$$

MLE We know that $\frac{\partial\ln f}{\partial\theta}(Z,\theta)=\psi(Z,\theta)$. Then we know that

$$\begin{aligned}(\Gamma'\Delta^{-1}\Gamma)^{-1}&=\left\{E\left[\frac{\partial\psi}{\partial\theta}(Z,\theta_0)'\right]'E[\psi(Z,\theta_0)\psi(Z,\theta_0)']^{-1}E\left[\frac{\partial\psi}{\partial\theta}(Z,\theta_0)'\right]\right\}^{-1}\\&=\left\{E\left[\frac{\partial^2\ln f}{\partial\theta\partial\theta'}(Z,\theta_0)'\right]'E\left[\frac{\ln f}{\partial\theta}(Z,\theta_0)\frac{\ln f}{\partial\theta}(Z,\theta_0)'\right]^{-1}E\left[\frac{\partial^2\ln f}{\partial\theta\partial\theta'}(Z,\theta_0)'\right]\right\}^{-1}\\&=I(\theta_0)^{-1}\end{aligned}$$

OLS $\psi(y,x,\beta)=x(y-x'\beta)$

$$\begin{aligned}\beta_{blp}&=E[x_ix_i']^{-1}E[x_iy_i]\\\Gamma&=E[x_ix_i']\\\Delta&=E[x_i\varepsilon_i^2x_i']\\\varepsilon_i&=y_i-x_i'\beta_{blp}\end{aligned}$$

Regression with estimated regressor $\psi(y,x,w,z,\delta,\beta,\gamma)=\begin{pmatrix}z(w-z'\delta)\\x(y-x'\beta-(w-z'\delta)\gamma)\\(w-z'\delta)(y-x'\beta-(w-z'\delta)\gamma)\end{pmatrix}$

$$\Gamma=\begin{pmatrix}-z_iz_i'&0&0\\x_iz_i\gamma_0&-x_ix_i'&-x_i\eta_i\\-z_i(y_i-x_i'\beta_0-(w_i-z_i\delta_0)\gamma_0)+\eta_iz_i'x_0&-\eta_ix_i'&-\eta_i^2\end{pmatrix}$$

where $\varepsilon_i=y_i-x_i'\beta_0-(w_i-z_i\delta_0)\gamma_0$.

$$\Delta=E\left[\begin{pmatrix}\eta_i^2z_iz_i'&z_i\eta_ix_i'\varepsilon_i&z_i\eta_i\varepsilon_i\\x_iz_i'\eta_i\varepsilon_i&\varepsilon_i^2x_ix_i^2&x_i\eta_i\varepsilon_i\\z_i\eta_i\varepsilon_i&x_i'\eta_i\varepsilon_i^2&\eta_i^2\varepsilon_i^2\end{pmatrix}\right]$$

Suppose $E[\varepsilon_i|x_i,z_i]=E[\eta_i|x_i,z_i]$, or $\varepsilon_i\perp\eta_i$. Then $\Gamma(3,1)=\Gamma(3,2)=\Gamma(2,3)=0$.

8.5 GMM: over-identified case

We have a function $\psi : \mathbb{R}^K \to \mathbb{R}^M$ such that $E[\psi(Z_i, \theta)] = 0 \iff \theta = \theta_0$. How do we construct our estimator when $M > K$. The idea is to minimize the quadratic form

$$Q_C(\theta) = \left(\frac{1}{N}\sum_i \psi(Z_i, \theta)\right)' C \left(\frac{1}{N}\sum_i \psi(Z_i, \theta)\right)$$

where C is a $M \times M$ positive semi-definite matrix.

Alternatively, we can use the method called empirical likelihood to get an estimator for θ. The idea is the following

- Consider the joint distribution of Z
- Compute the empirical probability distribution using discrete approximation and the restrictions given by the moment conditions
- Compute the parameter estimate $\hat{\theta}$ as the solution to the moment conditions at the estimated probabilities.

8.5.1 Two step estimator

We want to choose the weight matrix C to minimize the asymptotic variance of the estimator.

$$\hat{\theta} = \arg\min_{\theta} Q_C(\theta)$$

Consider a 2×2 matrix C, then $\hat{\theta}_\lambda = \lambda\bar{X} + (1-\lambda)\bar{Y}$ where

$$\lambda = \frac{2C_{11} + C_{12} + C_{21}}{2(C_{11} + C_{12} + C_{21} + C_{22})}$$

The asymptotic variance of $\hat{\theta}_\lambda$ is

$$\lambda^2\sigma_x^2 + (1-\lambda)^2\sigma_Y^2 + 2\lambda(1-\lambda)\sigma_{XY}$$

Choose λ by minimizing the asymptotic variance

$$\lambda = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Therefore,

$$C = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}^{-1} = \Delta^{-1}$$

This result is true in general, thus we look for $\hat{\theta}$ that minimizes $Q_{\Delta^{-1}}(\theta)$.

8.5.2 Large Sample Properties

$$g(\hat{\theta}) = 2\left[\frac{1}{N}\sum_i \frac{\partial\psi'}{\partial\theta}(Z_i, \hat{\theta})\right] C_N \left[\frac{1}{\sqrt{N}}\sum_i \psi(Z_i, \hat{\theta})\right] = 0$$

$$0 = \Gamma' C_0 \left[\frac{1}{\sqrt{N}}\sum_i \psi(Z_i, \theta_0)\right] + \Gamma' C_0 \left[\frac{1}{N}\sum \frac{\partial\psi}{\partial\theta'}(Z_i, \hat{\theta})\right] \sqrt{N}(\hat{\theta} - \theta_0)$$

$$\sqrt{N}(\hat{\theta} - \theta_0) \approx -(\Gamma' C_0 \Gamma)^{-1} \left(\Gamma' C_0 \frac{1}{\sqrt{N}}\sum \psi(Z_i, \theta)\right)$$

8.5.3 Optimal Weight Matrix

$C_0 = \Delta^{-1}$ is the optimal choice within the class of estimators minimizing a quadratic form of the type $Q_C(\theta)$.

$$\begin{aligned}
C_0 &= \Delta^{-1} + aA \\
V &= (\Gamma' C_0 \Gamma)^{-1}(\Gamma' C_0' \Delta C_0 \Gamma)(\Gamma' C_0 \Gamma)^{-1} \\
&= \frac{\Gamma' C_0 \Delta C_0 \Gamma}{(\Gamma' C_0 \Gamma)^2} = \frac{\Gamma'(\Delta^{-1} + aA)\Delta(\Delta^{-1} + aA)\Gamma}{(\Gamma'(\Delta^{-1} + aA)\Gamma)^2} \\
&= \frac{\Gamma'(I + aA\Delta)(\Delta^{-1} + aA)T}{(\Gamma'\Delta^{-1}\Gamma + a\Gamma' A\Gamma)^2} = \frac{\Gamma'\Delta^{-1}\Gamma + 2a\Gamma' A\Gamma + a^2\Gamma A\Delta A\Gamma}{(\Gamma'\Delta^{-1}\Gamma + a\Gamma A\Gamma)^2} \\
\frac{\partial V}{\partial a} &= 0, \text{ when } a = 0, \frac{\partial^2 V}{\partial a^2} > 0 \\
\frac{\partial V}{\partial a} &= \frac{1}{()^3}[2(\Gamma' A\Gamma)(\Gamma'\Delta^{-1}\Gamma) - 2(\Gamma'\Delta^{-1}\Gamma)(\Gamma' A\Gamma)] = 0
\end{aligned}$$

8.5.4 Empirical Likelihood

We use a very similar idea to demonstrate semi-parametric efficiency, due to Chamberlain.

FOC $\frac{\sum_i \delta_{ik}}{\pi_k} - \lambda\gamma_{Y_k} - \mu = 0, \forall k$. Therefore,

$$x\pi_k, \sum_k : \sum_i \sum_k \delta_k - \lambda \sum_k \gamma_{Y_k}\pi_k - \mu \sum_k \pi_k = 0 = N - \mu \implies \mu = N$$

Hence

$$\lambda\gamma_{Y_k} + N = \frac{\sum \delta_{ik}}{\pi_k} \implies \pi_k = \frac{\sum_i \delta_{ik}}{\lambda\gamma_{Y_k} + N} = \frac{\hat{p}_k}{1 + \lambda\gamma_{Y_k}/N}$$

Since $\lambda \sum_k \gamma_{Y_k}\pi_k = 0$, then

$$= \sum_k \frac{\hat{p}_k \gamma_{Y_k}}{1 + \gamma_{Y_k}\lambda/N} = \sum_k \frac{1}{N} \sum_i \frac{\delta_{ik} \cdot \gamma_{Y_k}}{1 + \gamma_{Y_k}\lambda/N} = \frac{1}{N} \sum_i \sum_k \frac{\delta_{ik} \cdot \gamma_{Y_k}}{1 + \gamma_{Y_k}\lambda/N} = \frac{1}{N} \sum_i \frac{Y_i}{1 + Y_i\lambda/N} = 0$$

$$\begin{aligned}
\theta &= \sum_k \pi_k \gamma_{X_k} \\
\sum_k \pi_k \gamma_{x_k} &= \sum_k \frac{\hat{p}_k \gamma_{X_k}}{1 + \gamma_{Y_k}\lambda/N} = \sum_k \frac{\hat{p}_k \gamma_{X_k}}{1 + \gamma_{Y_k}\lambda/N} = \sum_k \frac{1}{N} \sum_i \frac{\delta_{ik} \cdot \gamma_{X_k}}{1 + \gamma_{Y_k}\lambda/N} = \frac{1}{N} \sum_i \sum_k \frac{\delta_{ik} \cdot \gamma_{X_k}}{1 + \gamma_{Y_k}\lambda/N} \\
&= \frac{1}{N} \sum_i \frac{X_i}{1 + Y_i\lambda/N} = 0 \\
0 &= \sum_i \psi(Y_i, X_i, t, \theta)
\end{aligned}$$

$$\begin{aligned}
\psi(y, x, t, \theta) &= \begin{pmatrix} (\theta - x)/(1 + ty) \\ y/(1 + ty) \end{pmatrix} \\
t &= \lambda/N \\
\theta &= \theta_0, t = t_0, E[\psi] = 0
\end{aligned}$$

$$
\begin{aligned}
V &= (\Gamma'\Delta^{-1}\Gamma)^{-1} \\
\Gamma &= \begin{pmatrix} 1 & \sigma_{XY} \\ 0 & -\sigma_Y^2 \end{pmatrix}\Bigg|_{\theta=\theta_0,t=0} \\
\Delta &= E[\psi\psi']\big|_{\theta=\theta_0,t=0} = \begin{pmatrix} \sigma_X^2 & -\sigma_{XY} \\ -\sigma_{XY} & \sigma_Y^2 \end{pmatrix} \\
\Delta^{-1} &= \frac{1}{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}\begin{pmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{pmatrix} \\
V &= \begin{pmatrix} \sigma_X^2 - \sigma_{XY}/\sigma_Y^2 & 0 \\ 0 & 1/\sigma_Y^2 \end{pmatrix} \\
\text{2-step estimate } \tilde{\psi} &= \begin{pmatrix} \theta - x \\ y \end{pmatrix} \\
V(\hat{\theta}) &= \sigma_X^2 - \sigma_{XY}^2/\sigma_Y^2 = ((0 \quad 1)\,\Delta^{-1}\begin{pmatrix} 0 \\ 1 \end{pmatrix})^{-1}
\end{aligned}
$$

9 Causal Inference

9.1 Estimating Treatment Effects: potential outcomes

Action/treatment and potential outcomes. Causal effect: comparison of potential outcomes on a single unit $Y(1)-Y(0)$ where $Y(1)$ is the potential outcomes of the treatment and $Y(0)$ is the outcome without the treatment. fundamental problem of causal inference for a single unit: only one of the potential outcomes can be revealed. Rely on before and after observations or comparisons across multiple units.

9.2 Neyman's Idea

Neyman's two questions: average outcome if all units were exposed to treatment/control? Idea: construct estimator for the average treatment effects and derive its distribution by sampling repeatedly from the known distribution of W, the assignment vector. Suppose we have a population consisting N units. There exist two potential outcomes for each unit, $Y_i(0)$ and $Y_i(1)$, both are fixed instead of random. He was interested in the population average treatment effect

$$\frac{1}{N}\sum_i (Y_i(1) - Y_i(0)) = \tau$$

Suppose that we observed data from a completely randomized experiment in which M units were assigned to treatment and $N-M$ assigned to control. The intuition estimator for the average treatment effect is

$$\hat{\tau} = \frac{1}{M}\sum_{W_i=1} Y_i^{obs} - \frac{1}{N-M}\sum_{W_i=0} Y_i^{obs} = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$$

Hence $\hat{\tau}$ is unbiased estimator. To see this, consider the statistic $T_i = \frac{W_i \cdot Y_i^{obs}}{M/N} - \frac{(1-W_i)\cdot Y_i^{obs}}{(N-M)/N}$

$$E[T_i|Y_i(1), Y_i(0)] = \frac{E[W_iY_i(1)]}{M/N} - \frac{(1-E(W_i))Y_i(0)}{(N-M)/N} = Y_i(1) - Y_i(0)$$

$$E[\frac{1}{N}\sum_i T_i|Y(0), Y(1)] = \frac{1}{N}\sum_i E[T_i|Y_i(1), Y_i(1)] = \frac{1}{N}(Y_i(1) - Y_i(0)) = \tau$$

Given the assumption of a completely randomized experiment with the number of treated units fixed at M, the observations are not independent.

$$E[W_i \cdot W_j] = Pr(W_i = 1)Pr(W_j = 1|W_i = 1) = M/N \cdot (M-1)/(N-1)$$

$$E[W_i^2] = M/N$$

We can show that

$$V(\hat{\tau}) = \frac{S_0^2}{N-M} + \frac{S_1^2}{M} - \frac{S_{01}^2}{N}$$

where

$$S_W^2 = \frac{N-1}{\sum}_i (Y_i(W) - \bar{Y}(W))^2$$

$$S_{01}^2 = \frac{1}{N-1} \sum_i (Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0)))^2$$

Estimating Average Treatment Effects We have a random sample of size N drawn from a large population/ Each unit is characterized by a pair of potential outcomes, $Y_i(0)$ under the control treatment and $Y_i(1)$ under the active treatment. Each unit has a vector of characteristics, referred to as covariates, pretreatment variables or exogenous variables, and denoted by X_i. Each unit is exposed to a single treatment W_i, $W_i = 0$ if control treatment and $W_i = 1$ is active treatment. For each unit, we observe the tuple (Y_i, W_i, X_i). Distributions of (W_i, Y_i, X_i) refer to the distribution induced by the random sampling from the super population. The propensity score is the conditional probability of receiving the treatment

$$e(X_i) = Pr(W_i = 1|X_i = x) = E[W_i|X_i = x]$$

Define the conditional expectation and variance functions $\mu_W(X_i) = E[Y_i(W)|X_i = x]$ and $\sigma_W^2(X_i) = V(Y_i(W)|X_i = x)$, where $W \in \{0,1\}$.

Population average treatment effect $E[Y_i(1) - Y_i(0)]$. Population average treatment effect for the treated $E[Y_i(1) - Y_i(0)|W_i = 1]$. Alternatively, we might be interested in the sample version of these estimands. Sample average treatment effect $\frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$. Sample average treatment effect for the treated

$$\frac{1}{N} \sum_{i:W_i=1} (Y_i(1) - Y_i(0))$$

where $N_t = \sum_{i=1}^{N} W_i$

In the sample, we still don't observe both $Y_i(0)$ and $Y_i(1)$ for each i. Even if we do, we can't estimate $E[Y_i(1) - Y_i(0)]$ without error. We can estimate $\frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$ at least as accurately as $E[Y_i(1) - Y_i(0)]$ and typically more so. The difference between the two variance is the variance of the treatment effect. We can't learn more about the population than what we can learn about the observed sample.

9.2.1 Identification assumption

- Assumption 1 (Unconfoundedness)

  $$(Y_i(0), Y_i(1)) \perp W_i|X_i$$

  Suppose $\tau = Y_i(1) - Y_i(0)$ is constant.

  $$Y_i(0) = \alpha + x_i'\beta + \varepsilon_i, \varepsilon_i \perp x_i$$

  $$Y_i = \alpha + \tau W_i + x_i'\beta + \varepsilon_i$$

  Given constant τ, unconfoundedness $\iff W_i \perp \varepsilon_i|X_i$.

- Assumption 2 (Overlap)

  $$0 < Pr(W_i = 1|X_i) < 1$$

  The distribution of X_i between the treatment and the control group should overlap. For many of the formal results one will also need smoothness assumptions on the conditional regression functions and the propensity score ($\mu_W(X_i)$ and $e(X_i)$), and moment conditions on $Y_i(W)$.

- Assumption 3 (Mean Independence)

$$E[Y_i(W)|W_i, X_i] = E[Y_i(W)|X_i]$$

- Assumption 4 (Unconfoundedness for controls):

$$Y_i(0) \perp W_i|X_i$$

- Assumption 5 (Weak overlap):

$$Pr(W_i = 1|X_i) < 1$$

In reality, adjusting for the observables may not be sufficient: W_i is typically an outcome of agent's optimization behaviour. Using these assumptions can provide guidance to a first attempt to understand the evidence regarding treatment effect. Given the real difficulty in the analysis the unconfoundedness assumption merely asserts that all variables that need to be adjusted for are observed by the researcher (empirical question). Even when agents optimally choose their treatment, two agents with the same values for observed characteristics may differ in their treatment choices without invalidating the unconfoundedness assumption if the difference in their choices is driven by differences in unobserved characteristics that are themselves unrelated to the outcomes of interest.

9.2.2 Example

Adoption of new technology and firm output, evaluation of job training programs.

$$Y_i = g(W_i, \varepsilon_i)$$

$$\pi_i = Y_i - c_i W_i$$

$$W_i = \arg\max_W E[\pi(W)|c_i] = \arg\max_W E[g(W, \varepsilon_i) - c_i W|c_i] = \mathbf{1}_{E[g(1,\varepsilon)-g(0,\varepsilon)\geq c_i|c_i]} = h(c_i)$$

If $c_i \perp \varepsilon_i$, $W_i \perp [g(0, \varepsilon_i), g(1, \varepsilon_i)]$

Given the two key assumptions, unconfoundedness and overlap, one can identify the average treatment effects

$$\mu_W(X_i) = E[Y_i(W)|X_i = x] = E[Y_i(W)|W_I = W, X_i = x] = E[Y_i|W_i = w, X_i = x]$$

$$\begin{aligned}\tau(X_i) &= E[Y_i(1) - Y_i(0)|X_i = x] = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]\\ &= E[Y_i(1)|X_i = x, W_i = 1] - E[Y_i(0)|X_i = x, W_i = 0]\\ &= E[Y_i|X_i, W_i = 1] - E[Y_i|X_i, W_i = 0]\end{aligned}$$

9.2.3 Identification Assumptions: Unconfondedness

An important result building on the unconfoudnedness assumption shows that one need not condition simultaneously on all covariates.

$$\begin{aligned}Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) &= e(X_i)\\ Pr(W_i = 1|e(X_i)) &= e * X_i)\\ Pr(W_i = 1|Y_i(1)Y_i(0)e(X_i)) &= E[W_i|Y_i(1)Y_i(0)e(X_i)]\\ &= E[E[W_i|Y_i(0), Y_i(1), e(X_i), X_i]|Y_i(0), Y_i(1), e(X_i)]\\ &= E[E[W_i|Y_i(0), Y_i(1), X_i]|Y_i(0), Y_i(1), e(X_i)]\\ &= E[E[W_i|X_i]|Y_i(0), Y_i(1), e(X_i)] = E[e(X_i)|Y_i(0), Y_i(1), e(X_i)] = e(X_i)\\ Pr(W_i = 1|e(X_i)) &= E[W_i|e(X_i)] = E[E[W_i|X_i]|e(X_i)] = E[e(X_i)|e(X_i)] = e(X_i)\end{aligned}$$

Suppose

$$Y_i = \beta_0 + \beta_1 W_1 + \beta_2 X_i + \varepsilon_i$$

$$Y_i = \alpha_0 + \alpha_1 W_i + e_i$$

$$\alpha_1 = \beta_1 + \beta_2 \delta, W_i = \delta_0 + \delta X_i + \nu_i$$

By conditioning on $e(X_i) = E[W_i|X_i]$, therefore

$$W_i \perp X_i | e(X_i)$$

9.2.4 Identification assumption:f distributional and quantile treatment effects

In may cases one may wish to estimate other features of joint distribution of outcomes. Assumption 1 and 2 also allow for identification of the full marginal distribution of $Y_i(0)$ and $Y_i(1)$.

9.3 Estimate ATE: regression

$\hat{\mu}_0(x)$ used to predict $\bar{x}_t$

$$\bar{Y}_c + \hat{\beta}'(\bar{x}_t - \bar{x}_c)$$

$\bar{X}_t$ and $\bar{X}_c$ are different.

9.4 Estimate ATE: propensity score

$$\hat{\tau} = \frac{\sum w_i Y_i}{\sum w_i} - \frac{\sum (1 - w_i) Y_i}{\sum (1 - w_i)}$$

Weight each i by $\frac{1}{e(X_i)}$

$$E\left[\frac{w_i Y_i}{e(X_i)}\right] = E\left[\frac{w_i Y_i(1)}{e(x_i)}\right] = E\left[E\left[\frac{w_i Y_i(1)}{e(X_i)} | X_i\right]\right] = E\left[\frac{E[w_i|x_i) Y_i(1)}{e(X_i)}\right] = E[Y_i(1)]$$

It is similar for the other side. Therefore,

$$\hat{\tau} = \frac{1}{N} \sum \left[\frac{w_i Y_i}{e(X_i)} - \frac{(1 - w_i) Y_i}{1 - e(X_i)}\right]$$

9.5 Linear LV with constant coefficients

Let Y_i be the outcome of interest for unit i, W_i the endogenous regressor and Z_i the instrument. We are interested in the causal relationship between Y_i and $W_i : Y_i = \beta_0 + \beta_1 W_i + \varepsilon_i$. The concern here is W_i is endogenous. The solution for this is to use Z_i as an instrument. Instrumental variable Z_i is correlated with W_i and uncorrelated with the unobserved component ε_i.

9.5.1 Local Average Treatment Effects

Here let

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & W_i = 1 \\ Y_i(0) & W_i = 1 \end{cases}$$

Here W_i is the endogenous regressor. Here is a few assumption for the instrumental variable.

- Independence $Z_i \perp (Y_i(0), Y_i(1), W_i(0), W_i(1))$
- Random Assignment
- Exclusion Restriction: $Y_i(z, w) = Y_i(z', w)$

9.5.2 Local Average Treatment Effects

$$\hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

$$\hat{\beta}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{W}_i$$

$$\hat{\beta}_i = \frac{\frac{1}{N}\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{N}\sum(W_i - \bar{W})(Z_i - \bar{Z})}$$

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_i$$

$$\hat{\beta}_1^{IV} = \frac{\hat{\alpha}_1}{\hat{\pi}_i}$$

First we know that $\alpha_1 = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$ and

$$\begin{aligned} E[Y_i|Z_i = 1] &= E[Y_i|Z_i = 1, \text{complier}]Pr(\text{complier}|Z_i = 1) \\ &+ E[Y_i|Z_i = 1, \text{never}]Pr(\text{never}|Z_i = 1) \\ &+ E[Y_i|Z_i = 1, \text{always}]Pr(\text{always}|Z_i = 1) \\ &= E[Y_i(1)|\text{compiler}]\pi_c + E[Y_i(0)|\text{never}]\pi_n + E[Y_i(1)|\text{always}]\pi_a \end{aligned}$$

$$E[Y_i|Z_i = 0] = E[Y_i(0)|\text{compiler}]\pi_c + E[Y_i(0)|\text{never}]\pi_n + E[Y_i(1)|\text{always}]\pi_a$$

Therefore,

$$\alpha_1 = (E[Y_i(1)||]\text{complier}] - E[Y_i(0)|\text{complier}])\pi_c = E[Y_i(1) - Y_i(0)|\text{complier}]\pi_c$$

$$\pi_1 = E[W_i|Z_i = 1] - E[W_i|Z_i = 0]$$

$$\begin{aligned} E[W_i|Z_i = 1] &= E[W_i|Z_i = 1, \text{complier}]Pr(\text{complier}|Z_i = 1) \\ &+ E[W_i|Z_i = 1, \text{never}]Pr(\text{never}|Z_i = 1) \\ &+ E[W_i|Z_i = 1, \text{always}]Pr(\text{always}|Z_i = 1) \\ &= 1 \cdot \pi_c + 0 + 1 \cdot \pi_a \end{aligned}$$

Similarly

$$E[W_i|Z_i = 0] = \pi_a$$

$$\pi_1 = \pi_c$$

Lastly, we can get

$$\beta_1^{IV} = E[Y_i(1) - Y_i(0)|\text{complier}]$$

10 Non-Parametric Estimation

10.1 Nonparametric Density Estimation

Nonparametric methods to estimate distributions of random variables. Let $X_1, \cdots, X_N$ be a random sample from a distribution with density function $f_X(x)$. We are interested in estimating the density at a point c.

10.2 Histograms

Let $[a, b]$ be the support of X (in practice we may take the min and max of the data, although for the formal properties we need to have bounded support). Divide the range into K equal sized intervals,

$$(a + (k-1)(b-a)/K, a + k(b-a)/K]$$

Let N_k be the number of observations in interval k. The expectation of the proportion of observations in the kth interval.

$$\begin{aligned} E[N_k|N] &= Pr(a + (k-1)(b-a)/K < X_i < a + k(b-a)/K) \\ &= \int_{a+(k-1)(b-a)/K}^{a+k(b-a)/K} f_X(x)dx \\ &\approx f_X(a + (k-1/2)(b-a)/K)\frac{b-a}{K} \end{aligned}$$

Suppose the intervals are narrow, i.e., $(b-a)/K$ is small. Therefore, the density function is approximately constant in the interval. We can write the expectations as

$$f_X(a + (k-1/2)(b-a)/K) \cdot (b-a)/K$$

so we can get an estimate of the density function at a point c in interval k: $\hat{f}_X(c) = \frac{N_k}{N(b-a)K}$

Before considering modification to the basic method, let's think about the precision and bias of the histogram based estimate. For the bias, we know

$$E[\hat{f}_X(c)] = \int_{a+(k-1)(b-a)/K}^{a+k(b-a)/K} f_X(x)dx \cdot K/(b-a) = f_X(\tilde{a})$$

Therefore, the bias is

$$f_X(c) - E[\hat{f}_X(c)] = f_X(c) - f_X(\tilde{a})$$

, where $\tilde{a}$ is in the interval.

The prevision is

$$\begin{aligned} Var(\hat{p}) = p(1-p)/N &= f(\tilde{a})\frac{K}{b-a}(1 - f(\tilde{a})\frac{K}{b-a})\frac{1}{N} \\ Var(\hat{f}(a)) = Var((b-a)K/\hat{p}) &= f(\tilde{a})\frac{b-a}{K}(1 - f(\tilde{a})\frac{K}{b-a})\frac{1}{N} \end{aligned}$$

10.3 Centered Histograms

One problem is that the estimates at adjacent points can be very different if they fall on either side of an interval boundary. For example, if the density function is linear within an interval, the estimate in the middle is unbiased while the estimate at the boundaries have the largest bias.

10.4 Nearest Neighbor Estimators

10.5 Kernel Estimation

10.6 Bandwidth Selection

$$\begin{aligned} E[\hat{f}(x) - f(x)] = E\left[\frac{1}{Nh}\sum K\left(\frac{X_i - K}{h}\right) - f(x)\right] &= \frac{1}{h}\int k\left(\frac{y-K}{h}\right) f(y)dy - f(x) \\ &= \int k(u)f(x+h+hu)du - f(x) && u = \frac{y-x}{h} \\ &= \int k(u)f(x)dx + \int k(u)\cdot u \cdot hf'(x)du + \frac{1}{2}\int k(u)u^2h^2f''(x)du - f(x) \\ &= \frac{1}{2}f''(x)h^2k_2 && k_2 = \int z^2 \cdot K(z) \end{aligned}$$

$$
\begin{aligned}
V(\hat{f}(x)) &= V\left(\frac{1}{Nh}\sum K\left(\frac{X_i - K}{h}\right)\right) = \frac{1}{Nh^2}V\left(K\left(\frac{X_i - K}{h}\right)\right) \\
&= \frac{1}{Nh^2}E\left[K\left(\left(\frac{X_i - K}{h}\right)^2\right) - E\left[K\left(\frac{X_i - K}{h}\right)\right]^2\right] \\
&= \int K\left(\frac{y - K}{h}\right) f(y)dy \\
&= h\int K(u)^2 f(K + hu)du \\
&= h\int K(u)^2 f(x)du + o(h) \\
\int V(\hat{f}(x)) &= \int \frac{1}{Nh^2}h\int K(u)^2 du
\end{aligned}
$$

Solve the FOC, we get $h^* = 1.06\sigma N^{-1/5}$