ECON 710B - Cheat Sheet

Pointwise Confidence Intervals

Suppose that $Y = m(X, \theta) + e$ with $E[e|X] = 0$, $\hat{\theta}$ is the NLLS estimator, and $\hat{V}$ the estimator of $var(\hat{\theta})$. You are interested in the conditional mean function $E[Y|X = x] = m(x)$ at some x. Find an asymptotic 95% confidence interval for $m(x)$.

The standard error is:

$$s(x, \hat{\theta}) = \sqrt{R(x, \hat{\theta})' \hat{V} R(x, \hat{\theta})}, \text{ where } R(x, \hat{\theta}) = \begin{pmatrix} \frac{\partial m}{\partial \theta_1}(x, \hat{\theta}) \\ \vdots \\ \frac{\partial m}{\partial \theta_k}(x, \hat{\theta}) \end{pmatrix}$$

So the confidence interval for $m(x)$ is: $[m(x, \hat{\theta}) \pm 1.96 s(x, \hat{\theta})]$.

GMM

If β can be identified from moment conditions $E[g_i(\beta)] = 0$, then we can propose a GMM estimator:

$$\hat{\beta}^{GMM} := \arg\min_{\beta} J_n(\beta) = \arg\min_{\beta} n\bar{g_n}(\beta)' W \bar{g_n}(\beta)$$

where $\bar{g_n}(\beta) = \frac{1}{n} \sum_i g_i(\beta)$ and W is a positive definite matrix.

Efficient GMM: $W = \Omega^{-1} = [E[g_i(\beta) g_i(\beta)']]^{-1} \implies \hat{\Omega} = \frac{1}{n} \sum_i g_i(\hat{\beta}) g_i(\hat{\beta})'$

Asymptotic variance of efficient GMM is $(Q'\Omega^{-1}Q)^{-1}$ for $Q = E[\frac{\partial}{\partial \beta'} g_i(\beta)]$.

For IV model,

$$\begin{aligned} g_i(\beta) &= z_i(y_i - x_i'\beta) \\ \hat{\Omega} &= \frac{1}{n} \sum_i z_i z_i' \hat{e}_i^2 \\ Q &= E[Z'X] \end{aligned}$$

Difference-in-Difference

	Treatment	Control
Before	20.43	23.38
After	20.90	21.10

Diff-in-diff estimator: $(20.90 - 20.43) - (21.10 - 23.38) = 2.75$.

$$y_{it} = 23.38 - 2.95 Treated_{it} - 2.28 After_{it} + 2.75 Treated_{it} * After_{it}$$

Two-way transformation: $\ddot{x}_{it} = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}$.

Kernel Density Estimation

Second order kernel properties:

$$
\begin{aligned}
0 \leq K(u) \geq \bar{K} < \infty \\
K(u) = K(-u) \\
\int K(u)du = 1 \\
\int uK(u)du = 0 \\
\int u^2K(u)du = 0
\end{aligned}
$$

Kernel density estimator: $\hat{f}(x) = \frac{1}{nh}\sum_i K(\frac{X_i - x}{h})$

If $f(\cdot)$ is continuous in a neighborhood of x, then $\hat{f}(x) \to_p f(x)$ as $h \to 0$ and $nh \to \infty$.

- Bias$(\hat{f}(x)) \approx \frac{1}{2}f''(x)h^2$.
- Var$(\hat{f}(x)) \approx \frac{R_k f(x)}{nh}$.
- MSE$(\hat{f}(x)) = E[(\hat{f}(x) - f(x))^2] \approx$ Bias^2+ Var
- IMSE $= \int$MSE$(\hat{f}(x))dx$.
- Optimal bandwidth $h = cn^{-1/5}$.

Kernal Regression

Nadaraya-Watson estimator (or local constant estimator):

$$
\hat{m}_{NW}(x) = \frac{\sum_i y_i K(\frac{x_i - x}{h})}{\sum_i K(\frac{x_i - x}{h})} = \min_c \sum_i (y_i - c)^2 K(\frac{x_i - x}{h})
$$

Local Linear estimator reduces boundary bias.

Series Estimation

Approximate $m(x) \approx m_K(x) = x_K(x)'\beta_K = \beta_1\tau_1(x) + ... + \beta_K\tau_K(x)$.

- Polynomial regression: $m_K(x) = \beta_0 + \beta_1 x + ... + \beta_K x^K$.
- Splines with knots at $\tau_1 < \tau_2 < ... < \tau_N$:

$$
m_K(x) = \beta_0 + \beta_1 x + ... + \beta_p x^k + \beta_{p+1}(x - \tau_1)^p \mathbb{1}\{x \geq \tau_1\} + ... + \beta_{p+N}(x - \tau_N)^p \mathbb{1}\{x \geq \tau_N\}
$$

Estimation with OLS: $\hat{m}_K(x) = x_k(x)'\hat{\beta}_K$.

Approximation error: $r_k(x) := m(x) - m_K(x) = m(x) - x_k(x)'\beta_K$. Define $\delta_K := (E[r_k^2(x_i)])^{1/2}$.

Under some regularity conditions, the ISE satisfy $ISE(K) = O_p(\delta_K^2 + \frac{K}{n})$. Choose K by cross-validation.

Asymptotic normality: Suppose that $\theta = a(m)$ is some linear functional of m. If $a(m_K) = a_K'\beta_K$, then under some regularity conditions,

$$
\frac{\sqrt{n}(\hat{\theta} - \theta + a(r_K))}{V_K^{1/2}} \to_d N(0,1)
$$

where $V_K = a_K' Q_K^{-1}\Omega_K Q_K^{-1} a_K$, $Q_K = E[x_{Ki}x_{Ki}']$, and $\Omega = E[x_{Ki}x_{Ki}'e_i^2]$.

Regression Discontinuity

Sharp RD designs has assignment: $D_i = 1\{x_i \geq c\}$.

- Treated outcome: $Y_i(1)$ and untreated outcome: $Y_i(0)$
- Observable outcome: $y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$
- Average treatment effect for the subpopulation with $x_i \approx c$: $\tau = E[Y_i(1)|x_i = c] - E[Y_i(0)|x_i = c]$.
- Identification: Denote $m_1(x) = E[Y_i(1)|x_i = x]$ and $m_0(x) = E[Y_i(0)|x_i = x]$. If both $m_1(x)$ and $m_0(x)$ are continuous at $x = c$, then $\tau = m_1(c+) - m_0(c-) = E[y_i|x_i = c+] - E[y_i|x_i = c-]$.

Fuzzy RD designs has assignment: $P(D_i = 1|x_i = c-) \neq P(D_i = 1|x_i = c+)$.

- Average casual effects: $\tau = \frac{E[y_i|x_i=c+]-E[y_i|x_i=c-]}{E[D_i|x_i=c+]-E[D_i|x_i=c-]}$

M-Estimators

The parameter of interest is the θ_0 that minimizes $S(\theta) = E[\rho_i(\theta)]$:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} S_n(\theta) = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_i \rho_i(\theta) = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_i \rho(y_i, x_i, \theta)$$

Asymptotic normality: Under some regularity conditions, $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, H^{-1}\Omega H^{-1})$, where $H = E[\frac{\partial^2}{\partial\theta\partial\theta'}\rho_i(\theta_0)]$ and $\Omega = E[\frac{\partial}{\partial\theta}\rho_i(\theta_0)\frac{\partial}{\partial\theta}\rho_i(\theta_0)']$.

[Regularity conditions include θ_0 is in the interior of Θ and consistency $\hat{\theta} \to_p \theta_0$, etc.]

Nonlinear Least Squares

$E[y_i|x_i] = m(x_i, \theta)$ has a known functional form where θ is the parameter of interest and $m(x_i, \theta)$ is non-linear in parameter θ.

$$\hat{\beta}_{NLLS} = \arg\min_{\beta} \frac{1}{n}\sum_i (y_i - m(x_i, \beta))^2$$

Asymptotic normality (follows from M-estimators): $\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, H^{-1}\Omega H^{-1})$. Here $\Omega = E[e_i^2 m_{i\theta} m_{i\theta}']$ and $H = E[m_{i\theta} m_{i\theta}']$.

Under homoskedasticity, $E[e_i^2|x_i] = \sigma^2 \implies H^{-1}\Omega H^{-1} = \sigma^2 H^{-1}$

Quantile Regression

τth quantile $Q_\tau[X] : P(X \leq Q_\tau[X]) = \tau$ for continuous random variable x.

τth quantile $Q_\tau[Y|X] : P(Y \leq Q_\tau[Y|X]|X) = \tau$ for continuous random variable x.

Properties of medians

- $M[X] = \arg\min_m E[|X - m|]$
- $M[Y|X = x] = \arg\min_m E[|Y - m||X = x]$
- $M[Y|X] = \arg\min_{m(\cdot)} E[|Y - m(X)|]$

Check function: $\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\})$

Properties of quantile: $Q_\tau[X] = \arg\min_q E[\rho_\tau(X - q)]$

Estimation of linear model $y_i = x_i'\beta + e_i$:

- Median if $M[e_i|x_i] = 0$: $\hat{\beta}_{LAD} = \arg\min_\beta \frac{1}{n}\sum_i |y_i - x_i'\beta|$.
- Quantile if $Q_\tau[e_i|x_i] = 0$: $\hat{\beta}_\tau = \arg\min_\beta \frac{1}{n}\sum_i \rho_\tau(y_i - x_i'\beta)$.
- Under regularity conditions, $\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \to_d N(0, V_\tau)$.

Binary Choice Models

Binary choice: $y_i \in \{0, 1\}$ with $E[y_i|x_i] = P(y_i = 1|x_i) = p(x_i) \implies y_i = p(x_i) + e_i$.

Heteroskedasticity:

$$e_i = \begin{cases} 1 - p(x_i) & \text{with probability } p(x_i) \\ -p(x_i) & \text{with probability } 1 - p(x_i) \end{cases}$$

$E[e_i|x_i] = 0$ and $E[e_i^2|x_i] = p(x_i)(1 - p(x_i))$.

Linear probability model: $p(x_i) = x_i'\beta$. Estimate with OLS. Pros: Simple and easy to interpret. Cons: $p(x_i) = x_i'\beta \in \mathbb{R}$.

Index model $p(x_i) = G(x_i'\beta)$.

Probit:

- G is standard normal CDF.
- The log-likelihood function is: $\ell_n(\beta) = \sum_i [y_i \log \Phi(x_i'\beta) + (1 - y_i) \log(1 - \Phi(x_i'\beta))]$

Logit:

- G is logistic CDF: $\Lambda(x) = \frac{\exp(x)}{1+\exp(x)}$.
- The log-likelihood function is: $\ell_n(\beta) = \sum_i [y_i x_i'\beta - \log(1 + \exp(x_i'\beta))]$

Latent utility/profit interpretation: $y_i = \mathbb{1}\{y_i^* > 0\}$ with $y_i^* = x_i'\beta + e_i$.

Marginal effects: $\delta(x) = \frac{\partial}{\partial x} p(x) = \beta g(x'\beta)$.

Average marginal effects $E[\delta(x_i)]$

Multiple Choice Models

Multiple choice: $y_i \in \{1, 2, 3, ..., J\}$ with $p_j(x_i) = P(y_i = j|x_i)$.

Multinomial Logit/Probit:

- Latent variable is utility of option j: $U_j = x'\beta_j + \varepsilon_j$ where x_i represents an individual's characteristic. Option j is chosen if $U_j > U_k$ for all $k \neq j$.
- If ε_j is iid EV1, then we get multinomial logit model: $P(y = 1|x) = \frac{\exp(x'\beta_j)}{\sum_k \exp(x'\beta_k)}$
- Implies the independence of irrelevant alternatives: $\frac{P_j(x)}{P_k(x)} = \frac{\exp(x'\beta_j)}{\exp(x'\beta_k)}$
- ε_j could also be iid $N(0, 1)$ or jointly normal.
- Conditional logit: $U_j = W'\beta_j + x_j'\gamma + \varepsilon_j$ (includes alternative specific characteristics X_j).
- Mixed logit: $U_j = W'\beta_j + x_j'\eta + \varepsilon_j$ with $\eta \sim F(\eta|\alpha)$.
- Nested logit and ordered response.

Censored Regression

Censored model with unobserved latent variable y_i^* and observed (y_i, x_i):

$$\begin{aligned} y^* &= x'\beta + e \\ e &\sim N(0,\sigma^2) \perp x \\ y &= y^* \mathbb{1}\{y^* > 0\} \end{aligned}$$

Estimation with MLE. The pmf is:

$$\Phi\left(-\frac{x_i'\beta}{\sigma}\right)^{\mathbb{1}\{y_i=0\}} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]^{\mathbb{1}\{y_i>0\}}$$

CLAD estimator: Assume $M[e_i|x_i] = 0$ (instead of $e_i \sim N(0,\sigma^2)$), then

$$\hat{\beta}_{CLAD} = \arg\min_{\beta} \frac{1}{n}\sum_i |y_i - \max\{0, x_i'\beta\}|$$

Selection Models

Sample selection model:

$$\begin{aligned} y_i &= x_i'\beta + e_{1i} \\ S_i &= \mathbb{1}\{z_i'\gamma + e_{0i} > 0\} \\ (e_{1i}, e_{0i})' &\sim N(0,\Sigma) \perp (x_i, z_i) \end{aligned}$$

where y_i is observable if $S_i = 1$ and $\Sigma = \begin{bmatrix} \sigma^2 & \rho \\ \rho & 1 \end{bmatrix}$

Heckit: This model implies $y_i = x_i'\beta + \rho\lambda(z_i'\gamma) + \varepsilon_i$ with $E[\varepsilon_i|S_i = 1, x_i, z_i] = 0$.

Traditional estimation approach:

1. Estimate probit regression of S_i on z_i, derive estimator $\hat{\gamma}$.
2. OLR regress the observed data y_i on x_i and $\lambda(z_i'\hat{\gamma})$.

Alternatively, use MLE.

Model Selection

Three different information criterion to help to choose models:

- Cross-Validation (CV) is based on leave-one-out estimation.
- $BIC = -2\log L_n(\hat{\theta}) + K\log(n)$.
- $AIC = -2\log L_n(\hat{\theta}) + 2K$.

In linear model,

- $CV = \frac{1}{n}\sum_i (y_i - x_i'\hat{\beta}_{-i})^2$
- $BIC = n\log\hat{\sigma}^2 + K\log(n)$
- $AIC = n\log\hat{\sigma}^2 + 2K$

Choose the model with the smallest value of CV, AIC, and/or BIC.

Properties: Model selection consistent (AIC, BIC), over selection (AIC, CV), parsimonious (BIC), and asymptotically optimal (AIC, CV).

Shrinkage Methods

Idea: Exploit trade-off between variance and bias in estimator.

James-Stein Shrinkage Estimator

- Let $\hat{\theta} \sim N(0, V)$, then $\tilde{\theta} = (1 - \frac{c}{\hat{\theta}' V^{-1} \hat{\theta}})\hat{\theta}$.
- When $0 < c < 2(K-2)$, then $\tilde{\theta}$ has a weighted MSE less than $\hat{\theta}$.

Ridge Regression

$$\hat{\beta}_{ridge} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

$$\implies \hat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'Y$$

LASSO Regression

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

Elastic net uses a linear combination of L_1 (Lasso) and L_2 (ridge) penalties.

Machine Learning

Regression Trees: Branches, nodes or leaves, growing a tree. Disadvantage is large variance.

Bagging (Bootstrap Aggregation): Use bootstrap sample to grow regression tree, repeat and then take the average. Disadvantage is bootstrap estimates from each tree are correlated.

Random Forests: Decorrelate the regression tree in Bagging by randomly selecting the variables allowed to be branched on at each node.

Treatment effects and double ML: $Y = D\theta + X'\beta + e$.

- Post-model selection
- Post-regularizarion lasso
- Double debias