Empirical Economics-Cheat Sheet

① Causal Effects, Experiments & Regression Analysis

Potential-Outcome for binary $D_i \in \{0;1\}$, outcome Y_i (potential outcome if treated: Y_{1i}, — " — untreated: Y_{0i})

$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i \Rightarrow \tau_i = Y_{1i} - Y_{0i}$ (causal effect)

Only one known ⇒ makes causal inference difficult

ATE: $E[Y_{1i} - Y_{0i}]$ ATT: $E[Y_{1i} - Y_{0i} | D_i = 1]$

Observed difference in means: $E[Y_i|D_i=1] - E[Y_i|D_i=0]$ = ATT + Selection bias, where Sbias $= E[Y_{0i}|D_i=1] - E[Y_{0i}|D_i=0]$

Under random assignment: $E[Y_{0i}|D_i=1] = E[Y_{0i}|D_i=0]$ ⇒ unbiased estimation. $Y_i = \alpha + \beta D_i + \varepsilon_i$ $E[Y_{1i} - Y_{0i}]$, $E[Y_{0i}]$

⇒ Control variables can be added to improve the model. | Heterogenuous TE: $Y_i = \alpha + (\beta + r_i) D_i + \varepsilon_i$

② as Recap: Gauss-Markov Assumptions given in exam

Effects of Assumptions: 1-6: $\hat\beta_j \sim N(\beta_j, \sigma^2_{\hat\beta_j})$

1-4: $E[\hat\beta_j|X] = \beta_j$ 1-5: $Var[\hat\beta_j|X] = \frac{\sigma^2}{SST_j(1-R_j^2)}$ ⇒ BLUE

Asymptotic properties: 1-4 $\text{plim}\, \hat\beta_j = \beta_j$ 1-5: $\hat\beta_j \overset{a}{\sim} N(\beta_j, \frac{\sigma^2}{n a_j^2})$

Threats to int. val.: Endogeneity (OVB, ME, simult.), HC (efficiency), multicollinearity

③ Heteroskedasticity: $Var(u_i|X) = \sigma_i^2$ ⇒ Wrong standard errors

⇒ No longer BLUE (efficiency) & unbiasedness, consistency remains!

Testing for Heteroskedasticity ① Breusch-Pagan-Test ⓘⓘ White Test

i) $\hat u^2 = \delta_0 + \sum \delta_i x_i + r_i \Rightarrow H_0: \delta_i = 0\ \forall i \geq 1$ $H_1: \exists i$ such that $\delta_i \neq 0$

Test statistic: $nR^2_{\hat u^2} \sim \chi_k$ or $F = \frac{R^2_{\hat u^2}/k}{(1-R^2_{\hat u^2})/(n-k-1)}$ $F_{k,n-k-1}$

ii) Use non-linear covariates/regressands: simplified:

$\hat u^2 = \delta_0 + \delta_1 \hat y + \delta_2 \hat y^2 + r_i \Rightarrow H_0: \delta_1 = \delta_2 = 0$ $\boxed{F_{2,n-k-1}}$

Test: F-test (see above)

Dealing with Heteroskedasticity: i) WLS/FGLS ii) robust SEs

i) WLS: $Var(u|X) = \sigma^2 h(x_i)$ ⇒ divide everything by $\sqrt{h(x)}$ s.t. $h > 0$

ii) FGLS: You don't know the contributing factor ⇒ estimate weight

⇒ $\ln \hat u_i^2 = \delta_0 + \sum_{j=1} \delta_j x_{ij} + r_i \Rightarrow \hat h_i = \exp(\hat\delta_0 + \sum \hat\delta_j x_{ij})$

⇒ follow same procedure with $\sqrt{\hat h_i}$ like in WLS.

ii) Robust SE: $Var(\hat\beta_j) = \frac{\sum \hat r_{ij}^2 \hat u_i^2}{SSR_j^2}$ ⇒ still inefficient though

HC & LPM: Assume $y_i \in [0,1]$ and $y_i = \alpha + \beta X + u$

⇒ $Var(y_i|x_i) = p(x_i)[1 - p(x_i)]$ values close to 0.5 large error; value close to 0 or 1 ⇒ small error

④ Specification & Data Issues: Misspecification using RESET

Include higher order terms of dependent variable & test:

$y = \beta X + \delta_1 y^2 + \delta_2 y^3 \Rightarrow F = \frac{(R_U^2 - R_R^2)/q}{(1-R^2)/(n-k-q-1)}$ $F_{q, n-k-q-1}$ (added terms, independent)

Nested vs. Non-Nested Models: 1: βx_1 & $\beta_1 x_1 + \beta_2 x_2$; 2: βx_i & $\beta \ln x_i$

Testing Non-Nested Models:

CMA: model$_1$ + model$_2$ Test model$_1$: coefficients of model 2 ⇒ 0 and Use F-test vice versa

Davidson-MacKinnon Test:

⇒ T-test: y = model$_1$ + fitted$_{\text{model}_2}$ y = model$_2$ + fitted$_{\text{model}_1}$

↯ • They may be no clear winner (use R^2 as a decision)
• Rejecting does not validate the other.

Proxy variables (unobservables): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u_y$

Conditions of a good proxy: $x_3^* = \delta_0 + \delta_3 x_3 + v_3$

$\rho(u, x_i) = 0\ \forall i$ $\rho(u, x_3^*) = 0$ $\rho(v_3, x_i) = 0\ \forall i$

Measurement Error: CMEA: $E(e_0) = 0$, $\rho(y^*, e_0) = 0$

⇒ Dependent variable: Remains unbiased & consistent but inefficient

⇒ Independent variable: Attenuation bias: $\beta_1 \cdot \frac{Var(x_1^*)}{Var(x_1^*) + Var(e_1)}$

⇒ biased toward zero

Missing Data & Sample Selection:

Exogenous: Based on X ⇒ OLS remains consistent if $y \sim x$ omitted equal for all observations

Endogenous: Based on Y ⇒ OLS biased $E(u|X) \neq 0$ ↯

LAD: $\min_\beta \sum |y_i - x_i'\beta|$ ⇒ deal with outliers ⇒ centered around median

⑤ Simple Panel Models: Pooled Cross Sections combines multiple cross-sectional datasets sampled at different points in time

⊕ more observations for precise estimation, allows testing for effects over time

DiD estimation: $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 D_i + \delta_1 (D_i \cdot d2_t) + u_{it}$ ↳ post treatment indicator

Assumptions: Parallel trends; treatment effect ATT $= (y_{1,post} - y_{1,pre}) - (y_{0,post} - y_{0,pre})$

Fixed effect: $y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$ a_i: time-invariant ind effect

First differences: $\Delta y_{it} = y_{it} - y_{it-1} = \beta_1 \Delta x_{it} + \Delta u_{it}$ consistent if $\rho(\Delta x_{it}, \Delta u_{it}) = 0$

Balanced vs. Unbalanced Panel: same number of obs. across time; " varying number across individuals

Estimation in First Differences: Conditions for consistency: strict exogeneity $E[u_{it}|x_{iT}] = 0\ \forall t$, no residual autocorrelation.

↯ cannot estimate time-invariant variables, variation in Δx_{it}

↳ address by interaction term: $\Delta y_{it} = \beta_1 \Delta x_{it} + \gamma_i (T_{time} \cdot a_i) + \Delta u_{it}$

Serial Autocorrelation: Regress $\hat u_{it}$ on their lags, use FGLS if significant

⑥ Advanced Panel Data Methods: Fixed Effect Estimation

i) within transformation: $y_{it} - \bar y_i = \beta_1 (x_{it} - \bar x_i) + (u_{it} - \bar u_i)$

ii) LSDV: $y_{it} = \beta_0 + \beta x_{it} + \sum_{i=1} \gamma_i D_i + u_{it}$ ⇒ removes time-invariant heterogeneity

Random Effects Estimator: $y_{it} = \beta_1 x_{it} + a_i + u_{it}$ with $\rho(x_{it}, a_i) = 0$

FGLS transformation: $y_{it} - \theta \bar y_i = \beta_1 (x_{it} - \theta \bar x_i) + (1-\theta)\beta_0 + (u_{it} - \theta \bar u_i)$

$\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}}$ ~~Variance estimation~~ Hausman-Test for FE & RE

Model comparisons:

Choosing between FE & FD: FE preferred if u_{it} serially auto-correlated. FD preferred if u_{it} follows a random walk

⑦ Instrumental variable estimation: Model $y = \beta x + u$ exhibits $\rho(x,u) \neq 0$

⇒ Instrumental variable: $cov(z,x) \neq 0$ (relevance), $cov(z,u) = 0$ (Exogeneity) ⇒ $\hat\beta_1^{IV} = \frac{cov(z,y)}{cov(z,x)}$

Wald Estimator: $\hat\beta_{IV} = \frac{\bar y_{z=1} - \bar y_{z=0}}{\bar x_{z=1} - \bar x_{z=0}}$

Two Stage Least Squares

Multiple Instruments: First stage: $x = \pi_0 + \pi_1 z_1 + \dots + \pi_m z_m + v$ (all exogenous & instrumental variables)

Second stage: $y = \beta_0 + \beta_1 \hat x$ + other exog. vars + e

Variance and Inference: $Var(\hat\beta^{IV}) = \frac{\sigma^2}{n Var(x) \rho[z,x]^2}$ $\sigma^2 = Var(u)$

Weak Instruments: If $\rho(z,x)$ small ⇒ large SE

R^2 is different as OLS and can be negative.

⑧ Regression Discontinuity Design:

Sharp RD: deterministic $D_i = \begin{cases} 1 & \text{if } x_i > x_0 \text{ cutoff} \\ 0 & \text{otherwise} \end{cases}$

Model: $Y_i = \alpha + \beta X_i + \rho D_i + \eta_i$, where $\rho = Y_{1i} - Y_{0i}$

Identification: $\rho = \lim_{x \downarrow x_0} E[Y_i|x_i = x] - \lim_{x \uparrow x_0} E[Y_i|x_i = x]$

Fuzzy RD: $D_i = \gamma_0 + \gamma_1 I(x_i \geq x_0) + g(x_i) + r_i$ (probabilistic)

Wald Estimator: $\rho = \frac{\text{Reduced Form Discontinuity}}{\text{First Stage Discontinuity}} = \frac{E[Y|x_0^+] - E[Y|x_0^-]}{E[D|x_0^+] - E[D|x_0^-]}$

2SLS specification: $\alpha + \beta X_i + \rho \hat D_i + \eta_i$ (with $\hat D_i$ instrumented by $I(x_i \geq x_0)$)

Estimation Approaches: Parametric: polynomial approximation: $Y_i = \alpha + \sum \beta_k (x_i - x_0)^k + \rho D_i + \varepsilon_i$; Nonparametric: local linear reg: $\hat\rho = \arg\min \sum K\left(\frac{x_i - x_0}{\Delta}\right)(Y_i - \alpha - \beta(x_i - x_0) - \rho D_i)^2$ ↳ kernel function

Validity Checks: Continuity of covariates check: Pre-treatment covariates should not jump at x_0. Placebo Tests: no discontinuity in outcomes at false cutoffs

⑨ Limited Dependent Variable Models & Sample Selection Correction

LPM: $P(y=1|x) = \beta_0 + x\beta$ Logit Model: $P(y=1|x) = \Lambda(x\beta) = \frac{\exp(x\beta)}{1 + \exp(x\beta)}$ Probit Model: $P(y=1|x) = \Phi(x\beta) = \int_{-\infty}^{x\beta} \phi(t)\,dt$

Latent Variable Framework: $y^* = x\beta + \varepsilon$ $y = I(y^* > 0)$ where $\varepsilon \sim$ logistic (logit) or $\varepsilon \sim N(0,1)$ (probit)

Partial Effects: $\frac{\partial P(y=1|x)}{\partial x_j} = g(x\beta)\beta_j$, where $g(\cdot)$ PDF (logistic/normal). For discrete x_j, compute: $\Delta P = G(x\beta + \beta_j) - G(x\beta)$ MLE: $\ell(\beta) = \sum [y_i \log G(x_i\beta) + (1-y_i)\log(1 - G(x_i\beta))]$ Other Models: Poisson $\lambda_i = \exp(x_i\beta)$, $P(y=k) = e^{-\lambda}\lambda^k/k!$ Assumptions & Tests: Logit/Probit ⇒ correct distributio