This review material provides you a list of important concepts we've learned from lectures since the midterm. You can take it as a "table of content" with which we guide you to specific pages in lecture slides or section notes to review these concepts in more details.

This note does not include material covered before the midterm, but that will be part of what can be asked in the final: the properties of the OLS estimators, the five assumptions of the simple and the multiple linear regression models, omitted variable bias, interpretation of regression coefficients (including significance), single and multiple hypotheses tests, nonlinear regression functions, threats to internal and external validity, etc. Therefore, it would be best if you also read review notes for midterm as you prepare for the exam.

1 Panel Data, Fixed Effects

1.1 Introduction

- Panel data basics
  - **Lecture Notes:** *Handout 7 (pp.2-5)* **Section Notes:** Section 8 *(pp.1)*
- Two-period panels, first differences
  - **Lecture Notes:** Handout 7 *(pp.6-16)* **Section Notes:** Section 8 *(pp.2-3)*
- Longer panels
  - **Lecture Notes:** Handout 7 *(pp. 17-22)* **Section Notes:** Section 8 *(pp. 5-6)*

1.2 Fixed Effects (FE) Estimation

- Three estimation methods: (1) "n-1 binary regressors" OLS regression; (2) "Entity-demeaned" OLS regression; (3) "Changes" specification (only works for $T = 2$).
  - **Lecture Notes:** Handout 7 *(pp. 23 - 35)* **Section Notes:** Section 8 *(pp. 6-9)*

1.3 FE Assumptions

Consider the following model:

$$Y_{it} = \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + \beta_k X_{itk} + \alpha_i + u_{it}$$

- Assumption 1: $\mathbb{E}[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, \alpha_i] = 0$. This assumption says that u_{it} has mean zero, given the state fixed effect and the entire history of the X's for that entity. There are no omitted lagged effects and there is not feedback from u to future $\mathbf{X}$.

- Assumption 2: $\{\mathbf{W}_i\}_{i=1}^n$ are i.i.d., where $\mathbf{W}_i = \{Y_{i1}, \ldots, Y_{iT}, \mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{iT}\}$ is the data for unit i. This is satisfied if units are randomly sampled from their population by simple random sampling, then data for those units are collected over time. This does not require observations to be i.i.d. over time for the same entity.

- Assumption 3: $\mathbb{E}[\sum_{t=1}^T \tilde{\mathbf{X}}_{it}\tilde{\mathbf{X}}'_{it}]$ has rank k, where $\tilde{\mathbf{X}}_{it} = \mathbf{X}_{it} - \frac{1}{T}\sum_{t=1}^T \mathbf{X}_{it}$ are the demeaned regressors. This is the no multicollinearity assumption for the fixed effect case. Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables. This is important: we can't use fixed effects to analyze impacts of independent variables that don't change over time.

- Assumption 4: $\left(\mathbf{X}'_{it}, u_{it}\right)$ have nonzero finite fourth moments. Large outliers are unlikely.
  - **Lecture Notes:** Handout 7 *(pp. 36-40)* **Section Notes:** Section 8 *(pp. 4)*

1.4 Large Sample Theory

- From Assumptions $A1 \to A3$, we get that $\hat{\beta}_{FE}$ is consistent and asymptotically normal.
  - **Lecture Notes:** Handout 7 *(pp. 41-50)* **Section Notes:** Section 8 *(pp. 4)*
- Heteroscedasticity-and-autocorrelation-consistent asymptotic variance (HAC) and clustered standard errors
  - **Lecture Notes:** Handout 7 *(pp. 51-62)* **Section Notes:** Section 9 *(pp. 1)*

1.5 Advantages and limitations of fixed effects

- What types of OVB do fixed effects address and what do they not address?
  - **Lecture Notes:** Handout 7 *(pp.63-71)* **Section Notes:** Section 9 *(pp.2-3)*

2 Instrumental Variables

2.1 Introduction

- Treatment effect in the presence of noncompliance
  - **Lecture Notes:** Handout 8 *(pp.2-9)*
- IV estimator $\hat{\beta}_{IV} \xrightarrow[n\to\infty]{p} \frac{Cov(Y,Z)}{Cov(X,Z)}$, relevance and exogeneity

  - **Lecture Notes:** Handout 8 *(pp.10-11)* **Section Notes:** Section 9 *(pp.3-6)*

- Two Stage Least Squares (TSLS or 2SLS)
  - **Lecture Notes:** Handout 8 *(pp.12-15)* **Section Notes:** Section 10 *(pp.2)*

2.2 IV Model

Consider the following model:

$$Y = \beta_0 + \beta_1 \tilde{X}_1 + ... + \beta_k \tilde{X}_k + \beta_{k+1} \tilde{W}_1 + ... + \beta_{k+r} \tilde{W}_r + U$$

- Y is the outcome variable
- $\left(\tilde{X}_1, ..., \tilde{X}_k\right)$ are the **endogenous** regressors (potentially correlated with U)
- $\left(\tilde{W}_1, ...\tilde{W}_r\right)$ are the included **exogenous** regressors (uncorrelated with U)
- $\left(\tilde{Z}_1, ...\tilde{Z}_m\right)$ are the excluded exogenous regressors (uncorrelated with U) or instruments
- $(\beta_1, ..., \beta_{k+r})$ are the parameters of interest

Define $(r+m+1) \times 1$ dimensional vector $\mathbf{Z} = \left(1,, \tilde{Z}_1, ..., \tilde{Z}_m, \tilde{W}_1, ..., \tilde{W}_r\right)'$ This vector collects all the exogenous variables (whether in the regression or not). Define $(r+k+1) \times 1$ dimensional vector $\mathbf{X} = \left(1, \tilde{X}_1, ..., \tilde{X}_k, \tilde{W}_1, ..., \tilde{W}_r\right)'$ collecting all the included regressors (whether exogenous or not).

- Assumption 1: Error term U is mean independent of the exogenous regressors $\mathbf{Z}$ $\mathbb{E}\left(U|\mathbf{Z}\right) = 0$. Recall, this implies $\mathbb{E}\left(\mathbf{Z}U\right) = 0$ and since $\mathbf{Z}$ includes a constant, this means that the exogenous regressors are uncorrelated with U.

- Assumption 2: We observe an i.i.d. sample

$$\left\{Y_i, \tilde{X}_{1i}, ..., \tilde{X}_{ki}, \tilde{W}_{1i}, ...\tilde{W}_{ri}, \tilde{Z}_{1i}, ...\tilde{Z}_{mi}\right\}_{i=1}^{n}$$

- Assumption 3: The $(r+m+1) \times (r+k+1)$ dimensional matrix $\mathbb{E}\left(\mathbf{ZX}'\right)$ has rank $(r+k+1)$. The square matrix $\mathbb{E}\left(\mathbf{ZZ}'\right)$ has rank $(r+m+1)$. This assumption requires that $k \leq m$, which means that "we need as many instruments as endogenous regressors". When there is only one endogenous variable, if $m = 1$ and $r = 0$, this assumption requires $Cov(X, Z) \neq 0$, if $m > 1$, this assumption requires that at least one of the coefficients on the instruments from the regression of X on $\mathbf{Z}$ not equal to zero. When $k = m$, the model is **just identified**. When $k < m$, the model is **overidentified** (can do some instrument checking). When $k > m$ the model is not identified (need more instruments).

- Assumption 4: The matrix $\mathbf{S} \equiv \mathbb{E}\left(\mathbf{Z}U^2\mathbf{Z}'\right)$ is strictly positive definite.

- Assumption 5: The elements of $\left\{Y, \tilde{X}_1, ..., \tilde{X}_k, \tilde{W}_1, ...\tilde{W}_r, \tilde{Z}_1, ...\tilde{Z}_m\right\}$ all have finite fourth moments (large outliers are therefore unlikely).
  - **Lecture Notes:** Handout 8 *(pp.23-29 and pp.41)* **Section Notes:** Section 10 *(pp.1-2)*

2.3 Asymtotic Theory

- Under Assumption $A1$ through $A4$, $\hat{\beta}_{2SLS}$ is consistent for β, and $\hat{\beta}$ is approximately normally distributed with mean β and population variance matrix $V/\sqrt{n}$ in large samples. With $A1$ through $A5$, we also get an expression we can estimate for the asymptotic variance of the 2SLS estimator $Avar(\hat{\beta})$.
  - **Lecture Notes:** Handout 8 *(pp.30-43)* **Section Notes:** Section 10 *(pp.1-2)*

2.4 Instrument Validity

- Instrument **exogeneity**, J-test of overidentifying restrictions
  - **Lecture Notes:** Handout 8 *(pp.50-54)* **Section Notes:** Section 10 *(pp.4-5)*

- Instrument **relevance**, checking for weak instruments with a single X
  - **Lecture Notes:** Handout 8 *(pp.55-62)* **Section Notes:** Section 10 *(pp.3-4)*

3 Experiments and Quasi-Experiments

3.1 Treatment Effects in Heterogeneous Populations

- Setup
  - **Lecture Notes:** Handout 9 *(pp.3)* **Section Notes:** Section 11 *(pp.1)*

- Effects depending upon variables that are observed in the data
  - **Lecture Notes:** Handout 9 *(pp.4)* **Section Notes:** Section 11 *(pp.1)*

- Effects depending upon variables that are not observed in the data: average treatment effects (ATE) and local average treatment effects (LATE)
  - **Lecture Notes:** Handout 9 *(pp.5-16)* **Section Notes:** Section 11 *(pp.2-5)*

3.2 Randomized Experiments

- Experiments and quasi-experiments
  - **Lecture Notes:** Handout 9 *(pp.17-19)*

- Idealized experiments and causal Effects, DID and panel data
  - **Lecture Notes:** Handout 9 *(pp.20-24)*

- Including additional subject characteristics (W's)
  - **Lecture Notes:** Handout 9 *(pp.25-26)*

- Partial compliance
  - **Lecture Notes:** Handout 9 *(pp.27)*

3.3 Problems with Randomized Experiments in Practice

- Threats to internal validity
  - **Lecture Notes:** Handout 9 *(pp.28-30)*
- Threats to external validity
  - **Lecture Notes:** Handout 9 *(pp.31)*

3.4 Control Variables

- Introduction
  - **Lecture Notes:** Handout 9 *(pp.40-42)*
- Conditional mean independence
  - **Lecture Notes:** Handout 9 *(pp.43-44)*

- Implications of Conditional mean independence
  - **Lecture Notes:** Handout 9 *(pp.45-47)*

- Definition of control variable
  - **Lecture Notes:** Handout 9 *(pp.48-49)*

3.5 Quasi-Experiments

- Definition
  - **Lecture Notes:** Handout 9 *(pp.59-61)*
- Threats to internal validity
  - **Lecture Notes:** Handout 9 *(pp.62)*
- Threats to external validity
  - **Lecture Notes:** Handout 9 *(pp.63)*

4 Comparing Different Methods

- Comparisons should be performed on a case-by-case basis. Some hints:
  - Always start by comparing the sign, size, and the significance of key estimators under different settings may give you some intuition.
  - Always check which method's key assumptions are most likely to hold in the particular context of the question?
    - Always check whether you can directly infer the validity of the assumption (either based on the details given about the context, or from the data or from the output result table given).
    - If you cannot check/test for the assumptions directly, discuss the procedure/method you would follow to test them and how you would proceed depending on whether or not they hold.
  - Always check which method would have the least/most amount of bias (i.e. how much of the unobservable omitted variables each method corrects for)
  - When you compare a method to OLS, which OLS assumption is violated, and how does the alternative method deal with this violation?
  - Always check whether you can directly infer the validity discuss

5 Time Series

- Introduction
  - **Lecture Notes:** Handout 10 *(pp. 3-9)*
- Stationarity
  - **Lecture Notes:** Handout 10 *(pp. 10-13)*

- Autocorrelation and autocovariance
  - **Lecture Notes:** Handout 10 *(pp. 14-19)*
- Autoregression
  - **Lecture Notes:** Handout 10 *(pp. 20-26)*
- Autoregressive Distributed Lag (ADL)
  - **Lecture Notes:** Handout 10 *(pp. 27-30)*
- Autoregressive Moving Average (ARMA)
  - **Lecture Notes:** Handout 10 *(pp. 31-35)*