ACCT 601: Applied Econometrics and Research Design

Professor: Clive Lennox
Office: HOH 802
Cell Phone: (626) 592-3544
Email: clennox@usc.edu
Office Hours: By appointment

Course Description
The goal of this course is to help you develop a strong intuition for the research design methodologies that are commonly used in archival research.

Learning Objectives
By the end of the course students will be able to:
   1. Discuss the limitations of the research design methodologies that have been used in empirical research.
   2. Use STATA to conduct data manipulation, creation of variables, univariate tests, and graphs.
   3. Explain the intuition for and - estimate using STATA - the following types of models:
      - Ordinary Least Squares regression (OLS)
      - Median regression
      - Logit and probit
      - Multinomial logit
      - Ordered logit
      - Poisson and negative binomial regression
      - Tobit
      - Duration models.
      - Fixed effects models.
      - Random effects models.
      - 2SLS and 3SLS.
      - Path analysis.
      - Heckman model
      - Treatment effects model
      - Propensity score matching and entropy balancing
      - Difference-in-difference designs
      - Regression discontinuity designs
**Required Materials**
Students need to download all the datasets in advance and make sure they have access to STATA on their computer.

**Recommended Preparation**
A graduate level Econometrics class is a pre-requisite for this course. For students that are completely new to STATA, I recommend the book by Ulrich Kohler and Frauke Kreuter, “Data Analysis Using Stata”, which is available for purchase from the STATA website. I also have some introductory slides that are available to any students that are not yet familiar with STATA.

**Course Notes**
Research articles, powerpoint slides, datasets, and pre-recorded videos are available via Blackboard.

**Grading**
Grading is based entirely on the student’s work on the exercises and presentations.

**Exercises and Paper Presentations**
Each topic has assigned exercises and student presentations. **All** students must complete the assigned exercises and submit their answers to me via email (clennox@usc.edu) no later than 6pm on the evening before the class.

Students will present their answers to the exercise questions and will also present individual papers in class. See below for the schedule of presenters for each exercise and paper presentation.

**COURSE CALENDAR - PRESENTATIONS OF CLASS EXERCISES AND PAPERS**

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
<th>Week 6</th>
<th>Week 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercise 1</td>
<td>Exercise 1</td>
<td>Exercise 1</td>
<td>Exercise 1</td>
<td>Exercise 1</td>
<td>Exercise 1</td>
<td>Exercise 1</td>
</tr>
<tr>
<td>Exercise 2</td>
<td>Exercise 2</td>
<td>Exercise 2</td>
<td></td>
<td>Exercise 2</td>
<td>Exercise 2</td>
<td></td>
</tr>
<tr>
<td>Exercise 3</td>
<td></td>
<td></td>
<td></td>
<td>Exercise 3</td>
<td>Exercise 3</td>
<td></td>
</tr>
<tr>
<td>Exercise 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paper 1</td>
<td>Paper 1</td>
<td>Paper 1</td>
<td>Paper 1</td>
<td>Paper 1</td>
<td>Paper 1</td>
<td>Paper 1</td>
</tr>
<tr>
<td>Paper 2</td>
<td>Paper 2</td>
<td>Paper 2</td>
<td>Paper 2</td>
<td>Paper 2</td>
<td>Paper 2</td>
<td>Paper 2</td>
</tr>
</tbody>
</table>
WEEK ONE
Title: OLS, heteroscedasticity, correlated errors and clustering

Students must first watch the following pre-recorded videos by Prof. Lennox

- 1.1 The importance of using .do files and .log files in STATA (24 mins)
- 1.2 The basic idea underlying OLS (66 mins)
- 1.3 Correctly interpreting the coefficients (38 mins)
- 1.4 Examining the residuals (24 mins)
- 1.5 Multiple regression (10 mins)
- 1.6 Heteroskedasticity (15 mins)
- 1.7 Correlated errors (31 mins)

Students must next complete the following exercises which are based on the material covered in the above video segments.

Exercise #1.
1.2 The basic idea underlying OLS
- What are the assumptions of OLS regression? What are the implications if one or more of the OLS assumptions do not hold?
- Open the Fees.dta file in Stata and run the following two regressions: (1) audit fees on total assets, (2) the log of audit fees on the log of total assets **
- What does the output of your regression mean? Which model appears to have the better “fit”?

** Hint: Here is the code **
```stata
use "D:\NETAC\phd\Fees.dta", clear
gen lnaf=ln(auditfees)
gen lnita=ln(totalassets)
egen miss=rmiss(lnaf lnita)
reg auditfees totalassets if miss==0
reg lnaf lnita if miss==0
```

Exercise #2.
1.3 Correctly interpreting the coefficients


Answer the following questions relating to the study by Blackwell et al. (JAR 1998):
Theoretically, how would you expect the interest rate to be affected by a firm’s decision to have an external audit? What is the impact of auditing on the interest rate in eq. (1) of Blackwell et al. (1998)?

What is the main contribution of the study?

Using the coefficient estimates reported in Col. (1) of Table 4, calculate the values of total assets ($000) at which the effect of an audit on the interest rate is: (a) zero, (b) negative, (c) positive.

Page 58 of the study makes the following claim: “we estimate that a firm with $1 million of total assets saves 40 basis points from purchasing an audit, while one with $5 million of total assets saves 12 basis points.” Using the coefficient estimates reported in Col. (1) of Table 4, verify whether this claim is true or false.

Suppose the study had reported the estimated impact of an audit for firms with $11m in total assets or $15m in assets. What would have been the estimated impact of the audit on each firm’s basis points? Do you think the paper would have been published if the authors had made this alternate comparison?

Exercise #3.
1.4 Examining the residuals

- Open the “anscombe.dta” data file in Stata and run the following four OLS regressions: reg y1 x1, reg y2 x2, reg y3 x3, reg y4 x4
- What do you notice about the output from these four regressions?
- Draw scatter graphs for each of the four associations:
  - twoway (scatter y1 x1) (lfit y1 x1))
  - twoway (scatter y2 x2) (lfit y2 x2))
  - twoway (scatter y3 x3) (lfit y3 x3))
  - twoway (scatter y4 x4) (lfit y4 x4))
- From the scatter plots, which of the four regressions do you think are well specified (i.e., meet the OLS assumptions)?
- Which OLS assumptions are violated in the regressions?

Exercise #4.
1.5 Examining the residuals


Pong & Whittington (1994) estimate the following model of audit fees:

Audit fee = a + b Size + c. Size^2,
Where:

Audit fee = audit fees,

Size = total assets.

- What is the difference between the assumed functional form of their audit fee model in eq. (1) compared with prior literature? Do Pong & Whittington (1994) adequately justify their alternative functional form (see page 1075)?
- Estimate the same audit fee model as Pong and Whittington (1994) using the Fees.dta dataset. Obtain the OLS residuals and do a twoway scatterplot. Show a histogram of the residuals and compare the histogram to a normal distribution. **
- Based on your analysis of the residuals, do you consider the audit fee model to be well specified? Which OLS assumptions (if any) are violated? How will violation of the OLS assumptions affect the estimated coefficients and standard errors?

** Hint: Here is the code **

gen asset_sq= totalassets* totalassets
reg auditfees totalassets asset_sq
predict res, resid
twoway (scatter res totalassets, msize(tiny)) (lfit res totalassets)
sum res, detail
hist res if res>-40 & res<200, normal xlabel(-40(20)200)

Paper presentation #1


When presenting this paper, be sure to cover the following points:
- What is the typical approach used by studies in the “value relevance” literature to measure the value relevance of accounting variables?
- What are the three ways in which the R2 can vary across studies?
- What is the problem with trying to evaluate value relevance by comparing the R2s across studies or across time periods?
- Why does the author argue that residual variance is a better measure of model fit than the R2?
- Why does the author recommend using a scaling adjustment?
- Do you have any concerns with this paper or the author’s arguments?

Paper presentation #2

When presenting this paper, be sure to cover the following points:

- What do we mean by “cross-sectional dependence”? What types of study are most vulnerable to concerns about cross-sectional dependence?
- What do we mean by “time-series dependence”? What types of study are most vulnerable to concerns about time-series dependence?
- How are Fama-MacBeth standard errors calculated? Is the Fama-MacBeth procedure intended to address the problem of cross-sectional dependence or time-series dependence? What assumption does the Fama-MacBeth procedure make?
- What is the logic of the Newey-West approach? What is the main limitation of this approach?
- What is the major limitation with using firm fixed effects to address the problem of time-series dependence? What does it mean for a firm effect to be “temporary”?
- Suppose the independent variables and residuals for a given firm are perfectly correlated across time. What happens to the true (estimated) standard error as we add more years of data?
- Of the finance studies in Petersen’s survey, how many studies use the “wrong” approach to estimate the standard errors?
- Petersen describes the literature as a “teaching tool” – what does he mean by this? What is the implication for you as PhD students?
- Petersen concludes that “When both a firm and a time effect are present in the data, researchers can address one parametrically (e.g., by including time dummies) and then estimate standard errors clustered on the other dimension. Alternatively, researchers can cluster on multiple dimensions.” How would you choose between these two options?

Paper presentation #3


When presenting this paper, be sure to cover the following points:

- What do the authors mean when they say that “methods that correct for dependence in one direction typically assume independence in the other”?
- Of the accounting studies in their survey, how many studies use the “wrong” approach to estimate the standard errors?
- The study says “In contrast to Petersen's finding that "clustering standard errors by both firm and time appears unnecessary" in the finance applications he considers (2009, 473), we find that in a variety of accounting-specific applications, CL-2
is necessary to produce valid inferences.” Do you agree? Is there reason to doubt this recommendation based on what Petersen found?
- Do you have any reservations about the accounting applications selected by the authors of this study?

Past exam questions

**Q1. Model specification and the R²**
When presenting her empirical results during a research workshop, the presenter states that “The estimated OLS model has a high R² indicating that the model is well-specified.”
  a. Explain why the above statement is a non-sequitur.
  b. How should empirical researchers assess whether their OLS models are well-specified?

**Q2. Time-series dependence**
  a. Explain what is meant by time-series dependence. How is time-series dependence different from cross-sectional dependence?
  b. Explain why time-series dependence is particularly relevant for studies that use panel data. What are the main advantages and disadvantages of using panel data?
  c. Explain the merits of clustering by firm. Is it best to cluster by firm only or by both firm and year?
WEEK TWO
Title: Multicollinearity, outliers, and median regression

Students must first watch the following pre-recorded videos by Prof. Lennox

- 2.1 Multicollinearity (29 mins)
- 2.2 Outlying observations (18 mins)
- 2.3 Median regression (15 mins)
- 2.4 Looping (22 mins)
- 2.5 ANOVA (12 mins)

Students must then complete the following exercises which are based on the material covered in the above video segments.

Exercise #1.
2.1 Multicollinearity

• Open the international.dta dataset in STATA. Create dummy variables for each country. Run the following regressions in which importanceofequitymarket is the dependent variable and legalenforcement is the independent variable.

  o use "D:\NETAC\phd\international.dta", clear
  o tabulate countryid, gen(id_)
  o reg importanceofequitymarket legalenforcement id_1-id_24
  o reg importanceofequitymarket legalenforcement id_1-id_23
  o reg importanceofequitymarket legalenforcement id_1-id_22
  o reg importanceofequitymarket legalenforcement id_1-id_21

• How many country dummies can be included in your regression while still obtaining all the coefficient estimates and t-statistics? Explain.

• Are your results for the legalenforcement coefficient sensitive to your choice for which country dummies to exclude?

  o Hint: run the regression after dropping the dummy variables for countries 2, 9 and 16, and then re-run the regression after dropping the dummy variables for countries 6, 7 and 23

    ▪ reg importanceofequitymarket legalenforcement id_1 id_3-id_8 id_10-id_15 id_17-id_24
    ▪ reg importanceofequitymarket legalenforcement id_1-id_5 id_8-id_22 id_24

• Explain why your results are sensitive to which dummy variables are dropped. To answer this question, it will help to run the following scatter graph:
Exercise #2.
2.2 Multicollinearity


When presenting this paper, be sure to cover the following points:

- What are the two types of visible political events examined in the study? What does the study hypothesize?
- Explain the research design, variable definitions, and predictions in eq. (2) (on page 418) and in Panel A of Appendix B.
- How does the study measure crash risk? Does the crash risk variable capture a firm’s idiosyncratic crash risk or the risk of a market-wide crash?
- Discuss the results reported in Cols 1 and 2 of Table 4. Do you see any potential issues from the study’s inclusion of year fixed effects in Table 4?
- Do you have any other concerns with how the authors test their arguments?

Paper presentation #1


When presenting this paper, be sure to cover the following points:

- What do the following terms mean: “winsorization”, “truncation”, “Cook’s distance metric”, “robust regression”? What are the most commonly used approaches?
- Why do the authors say “It is important to stress at the outset that outliers identified by evaluating the distribution of each variable independently are not necessarily influential observations in a regression context”?
- Why do the authors say “OLS is known to be very sensitive to extreme observations”?
- Comment on the authors’ statement that “While OLS yields unbiased estimates, even in the presence of extreme observations or skewness in the distribution of the variables of interest, such estimates reflect the “average” effect, which is not typically the most appropriate statistic for hypothesis testing with skewed distributions.”
- What is the intuition for “robust regression”?
- Are there any other approaches the study could have considered for addressing outlier issues?
- Are there any disadvantages from using robust regression?
- Explain how you would use robust regression in Stata (Hint: type *help rreg* in the command window).

**Paper presentation #2**


When presenting this paper, be sure to cover the following points:
- What is meant by the term “rational expectations”?
- What research design do the authors use to test whether analysts process information efficiently?
- Why did the authors estimate a median regression (LAD) as an alternative to OLS? What assumptions do LAD and OLS make about the objective function of analysts?
- How does using LAD alter the inferences? What is the authors’ explanation?
- Can you think of any limitations in the study?

**Past exam questions**

a. What is the difference between “univariate” outliers and “multivariate regression” outliers?

b. Describe some alternative ways in which you might address outlier problems.
WEEK THREE
Title: Logit and probit models and extensions (multinomial and ordered models)

Students must first watch the following pre-recorded videos by Prof. Lennox
- 3.1 Why you should not use OLS when the dependent variable is binary (16 mins)
- 3.2 The basic idea underlying logit models (23 mins)
- 3.3 Estimating logit models (37 mins)
- 3.4 Multinomial models (26 mins)
- 3.5 Ordinal dependent variables (35 mins)

Students must then complete the following exercises which are based on the material covered in the above video segments.

Exercise #1.
3.1 Why you should not use OLS when the dependent variable is binary

- Open the Fees.dta dataset and do an OLS regression of the big6 dummy variable on the log of total assets. Obtain the predicted big6 values and the predicted residuals. Draw a scatterplot of the residuals against the predicted values of big6. **
- Do you notice a pattern between the residuals and the fitted values? Why does the pattern exist?

** Hint: Here is the code **
```
use "D:\NETAC\phd\Fees.dta", clear
gen lnta=ln(totalassets)
reg big6 lnta
predict big6hat
predict res, resid
scatter res big6hat
```

Exercise #2.
3.4 Multinomial models
3.5 Ordinal dependent variables

The following questions are related to the studies by Lau (1987) and Francis and Krishnan (1999):
- Which paper uses an ordinal model and which paper uses a multinomial model? What reasons (if any) did the authors give for estimating a multinomial model or ordinal model? Do you agree with those reasons?
- What are the pros and cons of estimating a multinomial model versus an ordinal model?
Paper presentation #1 (Logit / Probit versus OLS)


When presenting these two papers, be sure to cover the following points:
- Why is sample size an important issue to consider when estimating probit, logit and OLS models?
- What is the main reason put forward by Stone and Rasp (1991) for using probit or logit instead of OLS? How does their study contrast with the main reason put forward by Horrace and Oaxaca (2006) for using probit or logit instead of OLS?
- What is the reason for the statement on p. 172 of Stone and Rasp (1991) that OLS “can produce meaningless predicted probabilities”?
- What is the intuition for the finding of Horrace and Oaxaca (2006) that OLS produces biased and inconsistent estimates? Why do Horrace and Oaxaca (2006) suggest that a “trimming estimator” may help to reduce OLS bias? Do you think it would be feasible to use a trimming estimator in most datasets?

Paper presentation #2 (Simulated residuals)


When presenting this paper, be sure to cover the following points:
- Why can the residuals not be directly estimated in a nonlinear model such as probit or logit?
- How do Gourieroux et al. (1987) propose to generate simulated residuals?
- How can you use the simulated residuals to detect model misspecification problems, such as omitted variables, outliers and heteroscedasticity?

NB: You don’t need to worry about all the math in this paper. You just need to show that you understand how simulated residuals can be constructed in practice and how you would use the simulated residuals to evaluate model mis-specification.

Paper presentation #3 (The Ai and Norton critique)


When presenting these two papers, be sure to cover the following points:
- Why do Ai and Norton (2003) state that the magnitude of the interaction effect in nonlinear models does not equal the marginal effect of the interaction term?
• Do you have any observation about the formulation of eq. (1)? [Hint: Is it correct to say that the “dependent variable y is a dummy variable”?
• Explain how you can correctly calculate marginal effects in STATA using the inteff command. [Hint: You may need to install inteff to your version of STATA].
• Why does Puhani (2012) say that the Ai and Norton critique does not apply when researchers are using a diff-in-diff design?

Past exam questions
a. Why should you not use OLS when estimating models with categorical dependent variables?

b. Explain the basic logic for estimating a logit model.
**WEEK FOUR**

Title: Count data models, tobit models, and duration models

Students must first watch the following pre-recorded videos by Prof. Lennox

- **4.1 Count data models (35 mins)**
- **4.2 Tobit models (40 mins)**
- **4.3 Duration models (65 mins)**

Students must then complete the following exercise.

**Exercise #1.**

4.2 Tobit models

- Open the Fees.dta dataset and estimate models using OLS and tobit in which the dependent variable is the log of non-audit fees and the independent variable is the log of total assets **

- Is the Inta coefficient larger in the tobit model or the OLS model? Is the intercept larger in the tobit model or the OLS model?

- Draw the estimated relationships between lnnaf and Inta in the OLS and tobit models. Based on the drawn relationships, explain why the coefficients are different in the OLS and tobit models.

**Hint: Here is the code **

```bash
use "D:\NETAC\phd\Fees.dta", clear
sum nonauditfees, detail
gen lnnaf=ln(1+ nonauditfees)
sum lnnaf, detail
reg lnnaf Inta if miss==0
tobit lnnaf Inta if miss==0, ll
```

**Paper presentation #1 (Count data models)**


When presenting this paper, be sure to cover the following points:

- Why is OLS not well suited to models in which the dependent variable is a count-data variable?
- What’s the main difference between the empirical findings of Rock et al. (2001) versus Bhushan (1989)? Do you find the result of Rock et al. (2001) surprising? Do they provide a theoretical explanation for why they obtain a different result?
- Does the analyst following variable conform to all the properties of a count data process?
• What does the study mean when it refers to the “equidispersion” feature of the Poisson model and “unobserved heterogeneity”? How would you test the equidispersion assumption using STATA?
• The study says (p. 356) that variations of the Poisson model “may be appropriate for panel data but that is beyond the scope of this study”. Does STATA now provide panel data models for Poisson and negative binomial models [Hint: type *help xtpoisson* and *help xtnbreg* in the STATA command window.]
• What is the purpose of the Vuong test?

Paper presentation #2 (Parametric hazard models)

When presenting this paper, be sure to cover the following points:
• Why are hazard models useful for studying the duration of relationships, such as the relationship between an auditor and audit client?
• Comment on the potential differences between organizational relationships versus interpersonal relationships.
• Levinthal and Fichman (1988) are primarily interested in showing how the baseline hazard varies over time (see their Fig. 1). The baseline hazard is immediately recoverable from parametric survival models such as the Weibull.
  o Explain how you can estimate parametric survival models such as the Weibull in STATA. [Hint: see *help streg*].
  o The baseline hazard is not immediately recoverable from a non-parametric model such as the Cox proportional hazards. However, one can estimate the baseline hazard using a Poisson model. Explain how one can do this. [Hint: see page 93 of *help stcox*].
• Do you have any reservations with the Levinthal and Fichman (1988) study?

Paper presentation #3 (Hazard models)

When presenting this paper, be sure to cover the following points:
• According to Shumway (2001), what are the three econometric advantages of using a hazard model to forecast bankruptcy?
• What’s the difference between using a static vs a dynamic estimator?
• How does the Shumway (2001) study evaluate forecast accuracy?
• Do you have any reservations about the Shumway (2001) study?

NB: You don’t need to worry about all the math in this paper. You just need to show that you understand the intuition for the arguments.
Past exam questions
For which types of dependent variables would you estimate:
  i) multinomial models,
  ii) ordered models,
  iii) count-data models,
  iv) tobit models, and
  v) duration models.

For each dependent variable give one example from accounting research.

Identify the commands in STATA that can be used to estimate these five types of models.
WEEK FIVE
Title: Panel data models

Students must first watch the following pre-recorded videos by Prof. Lennox

- 5.1 The basic idea underlying panel data models (35 mins)
- 5.2 Linear regressions with fixed effects and random effects (65 mins)
- 5.3 Fixed effects and random effects models for logit and probit (24 mins)
- 5.4 Fixed effects and random effects models with discrete dependent variables (10 mins)

Students must then complete the following exercise.

Exercise #1.
5.2 Linear regressions with fixed effects and random effects

- Open the Fees.dta dataset and estimate OLS, fixed effects, and random effects models in which the dependent variable is the log of audit fees and the independent variable is the log of total assets. How do the coefficient estimates vary across the different models? Using a Hausman test and a Lagrange Multiplier test, decide which of the three models is preferable? **

**Hint: Here is the code**

```stata
use "D:\NETAC\phd\Fees.dta", clear
gen lnaf=ln(auditfees)
gen Inta=ln(totalassets)
reg lnaf Inta
xtreg lnaf Inta, fe i(companyid)
estimates store fixed_effects
xtreg lnaf Inta, re i(companyid)
estimates store random_effects
hausman fixed_effects random_effects
xttest0
```

Paper presentation #1 (Fixed effect models)

When presenting this paper, be sure to cover the following points:
- Why does using FE often come at the expense of efficiency?
- What does the author mean by type 2 errors and type 1 errors?
- What problems can arise when including FE for groups that lack any variation in X? What problems can arise when including FE eliminates a lot of the variation in X?
- Some studies report Firm × Year fixed effects. What does this notation mean (and not mean)?
- How are FE models related to diff-in-diff designs? Why are staggered diff-in-diff designs generally estimated by including fixed effects?
- What issues arise when the researcher is interested in the determinants of a specific regression coefficient (e.g., the earnings response coefficient)?
- Under which three conditions is it not necessary to include FE in order to obtain unbiased coefficient estimates? What are the potential advantages and disadvantages of including FE when their inclusion is not necessary to obtain unbiased coefficient estimates?
- How can you check how many firms in your sample have zero variation in X or Y?
- What would you consider doing if you find significant differences between firms with (without) variation in X?
- “The wisdom of using FE in OLS versus non-linear models is an active debate, but to date most accounting and finance studies have defaulted to OLS.” Comment on what you think the major issues are in this debate.
- Why can including FE induce a look-ahead bias?

**Paper presentation #2 (Testing the statistical significance of fixed effects)**


When presenting these papers, be sure to cover the following points:
- How do Bertrand and Schoar (2003) test for managerial style effects? Can you think of any other research design the authors could have used to test for managerial style effects?
- How do Bertrand and Schoar (2003) test whether there are overarching patterns in managerial decision-making?
- How do Bertrand and Schoar (2003) address the concern about endogenous matching of managers to companies?
- What is Fee et al.’s (2013) concern with respect to endogeneity? Do you have any concerns with their characterization of “exogenous” CEO departures?
- What is Fee et al.’s (2013) concern with respect to the F-statistics?
- Notwithstanding the issues raised by Fee et al. (2013), do you have any other concerns with the Bertrand and Schoar (2003) study?
**Paper presentation #3 (Lagged dependent variables in fixed effects models)**


When presenting these papers, be sure to cover the following points:

- Why is there a potential problem of bias if a researcher includes a lagged dependent variable in a regression that also incorporates fixed effects?
- What are the implications of this bias for the tests and conclusions in the studies of Hilary and Menzly (2006) and Hilary and Hsu (2011)?
WEEK SIX
Title: Instrumental variables

Students must first watch the following pre-recorded videos by Prof. Lennox

- **6.1 The problem of endogeneity bias and the basic idea underlying IV** (33 mins)
- **6.2 2SLS** (37 mins)
- **6.3 3SLS** (10 mins)

Students must then complete the following exercises.

**Exercise #1.**

*6.2 2SLS*

- Consider eqs. (1) and (2)
  - \[ \ln af = a_0 + a_1 \ln ta + a_2 \text{big6} + u \]  
  - \[ \ln ta = b_0 + b_1 \ln \text{age} + b_2 \text{listed} + v \]
  - where \( \ln af \) is the log of audit fees, \( \ln ta \) is the log of total assets, \( \ln \text{age} \) is the log of the company’s age in years, \( \text{listed} \) is a dummy variable indicating whether the company’s shares are publicly traded on a market.

- Is the instrumented \( \ln ta \) variable in (1) over-identified, just-identified, or under-identified? Explain.

- Estimate Eq. (1) using 2SLS. Test the validity of the chosen instrumental variables. Test whether the \( \ln ta \) variable is affected by endogeneity bias. **

- Verify that the test for instrument validity is not available if you change the model so that it is just-identified.

**Hint: Here is the code**

```bash
use "D:\NETAC\phd\Fees.dta", clear
gen fye=date(yearend, "MDY")
gen year=year(fye)
gen \text{age} = year-\text{incorporationyear}
gen ln_age=ln(\text{age})
gen listed=0
replace listed=1 if \text{companytype}==2 | \text{companytype}==3 | \text{companytype}==5
gen lnaf=ln(auditfees)
gen ln_ta=ln(totalassets)
ivregress 2sls lnaf \text{big6} (ln_ta=ln_age listed)
estat overid
estat endogenous
```
Exercise #2.
6.3 3SLS
The following questions are about the study by Whisenant et al. (JAR 2003).

• Are the models for audit fees and non-audit fees under-identified, just-identified, or over-identified?

• What are the exclusion restrictions in Whisenant et al. (2003)? How much justification did the study give for the chosen exclusion restrictions? Is it reasonable to suppose that the study’s exclusion restrictions are valid?

Exercise #3.
6.3 3SLS
The following questions are about the study by Ball et al. (JAE 2012).

• Are the models for audit fees and disclosure under-identified, just-identified, or over-identified?

• What are the study’s chosen exclusion restrictions? What justification did the authors give for their chosen exclusion restrictions? Is it reasonable to suppose their exclusion restrictions are valid?

Paper presentation #1 (IV models)

When presenting this paper, be sure to cover the following points:
• What do the authors find in their survey of 42 papers that have IV applications?
• What is meant by the terms “weak instruments” and “over-identifying restrictions”?
• Why do the authors emphasize the importance of economic theory in justifying instrumental variables? How practical do you find this suggestion?
• Why can IV estimation be “even less reliable” than OLS?
• Why do the authors highlight that “researchers frequently comment that their results are robust to endogeneity if the IV and OLS results produce similar estimates”?
• Why is it problematic to report only the R2 from the first-stage model without also reporting the partial explanatory power of the instruments that are unique to that model? How can we determine whether the chosen instruments are “weak” or “strong”? 
• The authors state that the β coefficient in eq. (1) is consistent only if δ = 0 and Corr(x, u) = 0. Why are both conditions necessary for β coefficient to be consistent?
Why is it not necessary for $\text{Corr}(x, \varepsilon) = 0$? If the $x$ variable is endogenous, does this automatically imply that the $\beta$ coefficient is inconsistent? [Hint: You can refer to eq. (4) in your answer.]

- With reference to eq. (5) what are the two potential problems with IV estimation?
- Why can the validity of the chosen instruments only be tested when the system is over-identified rather than just-identified? What are the limitations of the instrument validity test and the Hausman test?
- Explain the suggested approach for IV estimation.
- Explain the alternative approach to examining the potential impact of unobserved omitted variables?
- Larcker and Rusticus (2010) focus on two main issues: 1) whether the chosen instruments are weak or strong, 2) whether the chosen instruments are exogenous. Is there a third important issue that their study largely overlooked?

**Paper presentation #2 (IV models)**


When presenting this paper, be sure to cover the following points:
- What is the study’s main research idea? Why is there a potential endogeneity concern?
- Why do the authors choose extreme weather as their exogenous instrument? Do the authors test the power of the weather instrument? Is this instrument a powerful determinant of site visits?
- Is the weather variable likely to be exogenous? Is the exclusion restriction likely to be valid? Do the authors test the validity of the weather instrument – why or why not?
- In addition to 2SLS, what other approaches do the authors use in an attempt to address the endogeneity problem?

**Past exam questions**

- Why are exclusion restrictions required when estimating 2SLS or 3SLS models?
- Describe the “test of overidentifying restrictions”. What is the purpose of this test? When can it be used? What does the test assume? How reliable is the test for drawing inferences?
WEEK SEVEN
Title: Heckman models, treatment effect models, PSM, diff-in-diff, and regression discontinuities

Students must first watch the following pre-recorded videos by Prof. Lennox
- 7.1 The basic intuition underlying the Heckman model (38 mins)
- 7.2 One-sided and two-sided Heckman models (22 mins)
- 7.3 Treatment effects models (36 mins)
- 7.4 Propensity score matching (42 mins)
- 7.5 Diff-in-diff models (39 mins)
- 7.6 Regression discontinuity models (35 mins)

Students must then complete the following exercises.

Exercise #1.
7.2 One-sided and two-sided Heckman models

The following questions are about the study by Chaney et al. (TAR 2004).

- What are the three methodological differences between Chaney et al. (2004) and prior audit fee research? What are the study’s justifications for these three methodological changes? Given that there are three key differences between Chaney et al. (2004) and prior research and their findings are different from prior research, what would you have asked if you had been the reviewer?

- What are the exclusion restrictions employed in Chaney et al. (2004)? What does the study say about the need for exclusion restrictions? Are the coefficients on the inverse Mills ratio identifiable in the absence of exclusion restrictions?

- How do the study’s results compare between the OLS model and the Heckman model of audit fees?

- Do the results in Table 5 make sense? Does anything look odd?

Exercise #2.
7.2 One-sided and two-sided Heckman models

- Consider eqs. (1) and (2) which are two models for audit fees.
  - \( \ln af = a_0 + a_1 \ln ta + u \) \hspace{1cm} (1)
  - \( \ln af = a_0 + a_1 \ln sales + u \) \hspace{1cm} (2)

  - where \( \ln af = \log \) of audit fees, \( \ln ta = \log \) of total assets, \( \ln sales = \log \) of sales
• Suppose you want to estimate the two models separately for Big 6 and Non-Big 6 audit clients (i.e., four equations in total). At the same time, you want to control for potential selection bias given that Big 6 and Non-Big 6 audit firms are not randomly chosen. You are going to control for selection bias by estimating two-sided Heckman models for audit fees.

• Your first stage auditor choice models are as follows:
  o \( \text{big6} = b_0 + b_1 \ln \text{sales} + b_2 \ln \text{ta} + v \)
  o \( \text{nbig6} = c_0 + c_1 \ln \text{sales} + c_2 \ln \text{ta} + w \)
  o where \( \text{big6} = 1 \) (\( \text{big6} = 0 \)) if the company chooses a Big 6 (Non-Big 6) auditor; and \( \text{nbig6} = 1 \) (\( \text{nbig6} = 0 \)) if the company chooses a Non-Big 6 (Big 6) auditor.

• What exclusion restrictions are being assumed in equations (1) and (2)? Is there significant evidence of selection bias when you estimate the Heckman models? **

• Based on your results:
  o what are the signs of the MILLS coefficients?
  o what are the signs of rho?

** Hint: Here is the code **

use "D:\NETAC\phd\Fees.dta", clear
gen \( \ln \text{sales} = \ln \text{(sales)} \)
gen \( \ln \text{af} = \ln \text{(auditfees)} \)
gen \( \ln \text{ta} = \ln \text{(totalassets)} \)
egen miss = rmiss(\( \ln \text{af} \) \( \ln \text{ta} \) \( \ln \text{sales} \))
drop if miss! = 0
gen \( \text{nbig6} = 0 \)
replace \( \text{nbig6} = 1 \) if \( \text{big6} = 0 \)
heckman \( \ln \text{af} \) \( \ln \text{ta} \), select (\( \text{big6} = \ln \text{ta} \) \( \ln \text{sales} \)) twostep
heckman \( \ln \text{af} \) \( \ln \text{sales} \), select (\( \text{big6} = \ln \text{ta} \) \( \ln \text{sales} \)) twostep
heckman \( \ln \text{af} \) \( \ln \text{ta} \), select (\( \text{nbig6} = \ln \text{ta} \) \( \ln \text{sales} \)) twostep
heckman \( \ln \text{af} \) \( \ln \text{sales} \), select (\( \text{nbig6} = \ln \text{ta} \) \( \ln \text{sales} \)) twostep

Exercise #3.
7.3 Treatment effects models

The following questions are about the study by Leuz and Verrecchia (JAR 2000).

• What are the exclusion restrictions employed by Leuz and Verrecchia (2000)? Are the exclusion restrictions likely to be valid?
• Are their instruments likely to be exogenous? Are their instruments powerful?
• Comment on the results for IR97 and Inv. Mills Ratio in Panel B of Table 4.
• Comment on the functional form of the Free Float variable in the first and second stage models.
• Why would it have been helpful if Leuz and Verrecchia (2000) could have reported OLS models as well as Heckman models? What do you think the OLS models would have shown?

**Paper presentation #1 (Heckman)**


When presenting this topic, be sure to cover the following points:

• How are Heckman models similar to IV models? In what way are Heckman and IV models different?
• How are Heckman models supposed to work in theory? What is the inverse Mills’ ratio and what does the coefficient of the inverse Mills’ ratio capture?
• Why do the Z variables need to be: 1) exogenous, 2) powerful predictors of D, and 3) validly excluded from the second stage model for Y?
• What’s the difference between a treatment effects model (treatreg) and a sample selection model (heckman)?
• Why is the inverse Mills’ ratio highly correlated with the endogenous regressor (D) and the independent variables from the second stage model (X)?
• What are the two sources of identification for the inverse Mills’ ratio?
• Why is multicollinearity an issue for estimating coefficients?
• How has the Heckman model been implemented in many accounting studies? Why is the implementation a cause for concern?
• What suggestions do the authors offer for future research?

**Paper presentation #2 (PSM)**


When presenting this topic, be sure to cover the following points:

• How do studies often oversell the advantages of PSM?
• What is the advantage of PSM over traditional regression?
• Why does PSM tend to reduce the effect size and reduce testing power?
• Why does PSM raise concerns about sampling criteria and sampling representativeness?
• Why does the degree of explanatory power in the first-stage model not necessarily indicate the effectiveness of PSM?
• What are the authors’ suggestions for improved applications of PSM?

**Paper presentation #3 (Staggered diff-in-diff)**


NB: Both papers cover the same topic and you don’t need to present them separately. Instead, you can focus on the commonalities of the two papers. The Barrios (2021) study is probably easier to understand so I suggest you read that one first. The Baker et al. (2001) study is more comprehensive and a bit more sophisticated but you don’t need to discuss everything in that paper (e.g., you don’t need to discuss the replications of prior studies).

When presenting this topic, be sure to cover the following points:

- How does a traditional diff-in-diff work?
- How is a staggered diff-in-diff different from a traditional diff-in-diff? How many studies in accounting and finance use staggered diff-in-diff versus traditional diff-in-diff?
- Why is a staggered diff-in-diff usually estimated with fixed effects for each cross-sectional unit and fixed effects for each time period?
- What does it mean to say that the already-treated group acts as a control group even though it has already been subjected to treatment?
- Why is the estimated treatment effect biased in a staggered diff-in-diff in which the treatment effects vary across units or across time (e.g., see Simulation 6 in Baker et al. (2001))?
- Why are time-varying treatment effects problematic even when the treatment effects are identical across units? (e.g., see Figure 5 of Barrios 2021).
- Why do groups in the middle of the panel receive more weight than those at the end?
- How does the “stacked regression” design work?

**Past exam questions**

a. Why are Heckman (or treatment effect) models technically identifiable even when the researcher does not impose any exclusion restriction?

b. Would you advocate estimating Heckman (or treatment effect) models without any exclusion restrictions? Explain why (why not).

c. Describe three accounting studies that poorly implement instrumental variables techniques. What are the econometric problems in those studies?
d. A student informs you that mandatory audit firm rotation was recently introduced for public interest entities in the UK. Companies traded on the London Stock Exchange are classed as public interest entities and hence they are now subject to the mandatory audit firm rotation requirement. Companies traded on the Alternative Investment Market are not public interest entities and hence they are not subject to the mandatory audit firm rotation requirement. The student proposes to examine the impact of mandatory audit firm rotation using the following research design:

\[ Y = \alpha_0 + \alpha_1 TREAT + \alpha_2 POST + \alpha_3 TREAT \times POST + u \]

where \( Y \) = a set of dependent variables that could be affected by mandatory audit firm rotation (e.g., audit fees, audit quality); \( TREAT = 1 \) if the company is a public interest entity, 0 otherwise; \( POST = 1 \) in the years following mandatory audit firm rotation, 0 in the years before mandatory rotation.

Record the expected values of \( Y \) in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Before the introduction of mandatory audit firm rotation</th>
<th>After the introduction of mandatory audit firm rotation</th>
<th>Difference (after vs. before)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-public interest entities</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Public interest entities</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Difference (PIE vs. non-PIE)</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

\[ Y = \alpha_0 + \alpha_1 TREAT + \alpha_2 POST + \alpha_3 TREAT \times POST + u \]

where \( Y \) = a set of dependent variables that could be affected by mandatory audit firm rotation (e.g., audit fees, audit quality); \( TREAT = 1 \) if the company is a public interest entity, 0 otherwise; \( POST = 1 \) in the years following mandatory audit firm rotation, 0 in the years before mandatory rotation.

Record the expected values of \( Y \) in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Before the introduction of mandatory audit firm rotation</th>
<th>After the introduction of mandatory audit firm rotation</th>
<th>Difference (after vs. before)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-public interest entities</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Public interest entities</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Difference (PIE vs. non-PIE)</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

e. What assumptions do researchers generally make when they estimate: i) difference-in-differences designs, ii) regression discontinuity designs. How can researchers test whether the assumptions are valid?

f. Another student identifies a different country where mandatory audit firm rotation was recently introduced for companies with a market cap exceeding $50 million. The student is unsure whether to test the consequences of mandatory audit firm rotation using a difference-in-differences design or a regression discontinuity design. What would you suggest?