PPD 558: Multivariate Statistical Analysis

Nicolas Duquette

Spring 2021

Thursday – 2PM to 5:20 PM Pacific Time
Location: Online-only

Instructor: Nicolas Duquette
Office Hours: Online only
- Mondays 2-3 pm
- Thursdays 10-11 am
- Or by appointment

Email: nduquett@usc.edu

Teaching Assistant: Alison Holt
Office Hours: Wednesdays 12-2pm or by appointment
Email: alisonjh@usc.edu

Course Description

This course will provide you with the analytical and quantitative skills required to conduct applied statistical research and to think critically about methodology and proper interpretation of results when reading and analyzing empirical research such as that found in academic journals and other policy papers.

The foundation of this course is multivariate regression analysis. We will begin with the Ordinary Least Squares (OLS) model and expand our coverage to topics including logistic models, panel data, and experimental methods to evaluate the impacts of public policies. We will discuss common problems faced by these methods, techniques for diagnosing and addressing these problems, and selection of the appropriate econometric tools to answer any given question.

Prerequisite: PPD 502x, PPD 525, or equivalent

Learning Objectives

This course will focus on training students to be capable practitioners and sophisticated consumers of multivariate regression analysis for public policy. While we will be making use of econometric theory, it is viewed as a means to an end: this course has a strong applied (rather than theoretical) orientation, so our coverage of econometric theory will be limited to those elements that directly serve the primary goal of enabling students to be successful users of econometric analysis. A major goal of this course is to train students to effectively use econometric methods to inform the solution of complex policy, management, and planning problems.

Pandemic Adaptation

This course has been designated as online-only because of the Sars-CoV-2 pandemic. This syllabus is *subject to change as the public health environment changes*. Please be patient and understanding with any on-the-fly adjustments.

Similarly, I promise to be *accommodating and understanding* if the pandemic impinges on your ability to meet class obligations. Just talk to me.

Lectures will be posted as recordings to the Google Drive, along with PDFs of lecture slides. Students are expected to watch all the lectures *before* each week's class meeting.

We will meet synchronously via Zoom at the scheduled time (2pm Pacific time on Thursdays) to work through in-class exercises and have open discussion of any questions you have from the lecture material. Because the lecture content will have already been viewed asynchronously, these Zoom meetings will likely be substantially shorter than the full three hours and twenty minutes allocated.

Office Hours

Office hours are an opportunity to "stop by" and talk about course materials or any other questions you may have. This year, because of the pandemic, we will not literally be in offices. I will be on Zoom, and you can open the office hours link[1] to video chat during the designated times (2-3pm on Mondays and 10-11am on Thursdays). You can also make an appointment by email for a different meeting time.

The TA, Alison Holt, will also hold office hours, on Wednesdays from 12-2. Again, open the link to her Zoom room to stop by. Alison can also schedule meetings at other times by appointment.[2]

We will not have office hours on USC holidays or Wellness Days. All office hour times (and other times) are for Pacific Time.

[1] `https://usc.zoom.us/j/95821671253?pwd=Zm53bHZpeGFoK09wU0NzdU1sZG9pUT09`

[2] `https://usc.zoom.us/j/92385645793?pwd=MUg1T0U5VCtMZk1abzJydmphN1VFZz09`

Course Assignments and Grading Policies

Grades will be based on a mixture of problem sets and an analytical project. All of these assignments will involve working with data to answer questions appropriate for multivariate regression analysis of social and policy questions.

Problem Set Due Dates

Assignment	Due Date	Weight
Problem Sets		*50**
1 - Linear Models #1	February 4	10
2 - Linear Models #2	February 11	10
3 - Standard Errors	February 25	10
4 - Binary Outcomes	March 11	10
5 - Panels	March 25	10
6 - Cumulative	April 29	10
Analysis Project		*50*
Project Proposal	February 18	5
Data and Methods Report	March 4	5
Preliminary Results	March 18	5
Presentations	April 8, 15 or 29	10
Final Paper	May 12	20
Replication File	May 12	5
Total Weight		*100*

* – Problem sets sum to 50, not 60, because the lowest problem-set score is dropped.

The individual graded items will be converted into percentage points of the final course grade and then a letter grade for the course assigned as follows:

Letter grade assignment

Raw Grade Percentage	Letter Grade
$\geq$ 93 out of 100	A
90–92.99	A-
87–89.99	B+
83–86.99	B
80–82.99	B-
<80	$\leq$ C+

Problem Sets

There will be six problem sets in this class. Five of the six problem sets are worth 10 percent of the final course grade. The lowest of your six problem set grades will be dropped. So your five best problem sets will collectively account for half of your course grade.

Individual problem sets assign points to questions. These are weights on questions *within* the problem set. For example, problem set 1 will have three questions, worth 6, 16, and 8 points for a total of 30. If you received 23 of the possible points, you would receive $23/30 \times 10 \approx 7.667$ percentage points toward your course grade.

I will randomly assign you to a partner (group of two) for the first five problem sets. You are permitted but not required to work with your assigned partner on that problem set. However, if you do not wish to work with your assigned partner, **you are not allowed to work with any other classmates**.

I will assign a new partner for each of the first five problem sets. This policy is designed to allow students to learn from each other and meet classmates they do not yet know despite the barriers of remote learning. You may choose to work alone if you prefer, but to repeat: **you may not choose your own partner for these assignments**.

For the sixth problem set, you may work together with your a group of up to three classmates of your choosing.

Answers to each problem set must be submitted in PDF format with the along with Stata do-files used to generate all answers. Each group needs to submit only one copy of its problem set by email to `alisonjh@usc.edu`. You are encouraged to CC your email to all group members and to `nduquett@usc.edu`.

Problem sets are due at 1:59PM on the date assigned. Ordinarily, PPD 558 has strict penalties for late assignments. Because of the challenges we are all facing during the pandemic, I would like to be more flexible; however, it is also important for problem sets to be submitted close to on time so I can post answer keys to further student learning. I will therefore allow you to turn in your problem sets up to one week late without penalty, *as long as you notify me that you need the extra time as soon as possible*. Once I have posted answer keys, it is too late to submit and your problem set will receive a score of zero. This late policy does *not* apply to analysis project components, and is subject to change if I think it is not serving the class well.

Analysis project

The other major component of this course will be an original analysis project. Detailed instructions and advice on the individual components of the project are in a guide at the end of this

document. Here, briefly, is how the project will work and be factored into your final course grade.

In our first class meeting, we will introduce ourselves and everybody will note their policy interests (such as labor and employment, tax policy, environment, education, and so on). With my help and the TA's, you will then organize yourselves into groups of 2–4 students with shared interests. This group will conduct an original analysis of real data to answer a question of interest, and report on their findings in a written document and class presentation.

To maintain a successful trajectory for the group project, groups will be required to submit three items—a project proposal, a data and methods plan, and a set of preliminary results—during the semester. These are worth five percent of your course grade each, and will be scored on the basis of compliance with the detailed directions provided in the guide. Please send your materials both to the instructor (`nduquett@usc.edu`) and the TA (`alisonjh@usc.edu`).

A presentation of your project in the final weeks of the class will count for ten percent of your grade. These will be spread out over the three course meetings late in the semester. Each of those class meetings will have 3–4 groups present for 15 minutes with another 15 minutes for class discussion. Presentation materials are due at noon on the date of the presentation, and late submissions will receive a grade of zero.

Lastly, the projects will culminate in a final written paper and replication file. The paper will be a ten-page report explaining the research project, data used, and methods applied, then reporting and interpreting the results obtained. It is worth twenty percent of the course grade. Along with the final report, groups must submit a full replication file, that is, a collected archive of all do-files and data which will make it easy for an outside researcher to recreate the entire analysis from start to finish exactly as reported; this is worth five percent of your final grade.

The final report and replication file are both due at 11:59pm (23:59) Pacific Time on the final day of the exam period. However, because students are working in different time zones and managing many different obligations at the end of the semester, I strongly encourage you to *submit these materials earlier if you can*! Late submissions will receive a grade of zero.

What about attendance and participation?

In normal times, PPD 558 has strict attendance and participation policies. These are not normal times. I recognize that some of you are taking this class at odd hours in your local time zone, or with unreliable Internet connectivity. There will therefore not be formal grade consequences for attendance or participation. (I'm not scheduling any timed exams for the same reason.)

However, this class does depend on some level of participation to be successful. Therefore, I reserve the right to adjust your final grade if your class citizenship is especially good and not reflected in your raw score. It is particularly important that you attend the class meetings where your peers present their semester projects and ask constructive questions.

Technology

The Stata software package is required for in-class data analysis, take-home assignments, and the final project. You must have access to Stata to complete assignments and participate in class.

The Price school recommends you use the virtual machines equipped with Stata/SE at `https://cloudapps.usc.edu`.[3] If you prefer to purchase Stata for installation directly on your personal computer, student licenses are available at StataCorp's web site.[4] A six-month license of Stata/IC (the cheapest but quite limited option) should be sufficient for everything we do in this course.

USC technology rental program

We realize that attending classes online and completing coursework remotely requires access to technology that not all students possess. If you need resources to successfully participate in your classes, such as a laptop or internet hotspot, you may be eligible for the university's equipment rental program. To apply, please submit an application at `https://studentbasicneeds.usc.edu/resources/technology-assistance/`. The Student Basic Needs team will contact all applicants in early August and distribute equipment to eligible applicants prior to the start of the fall semester.

Course Web Sites

Lecture videos, lecture slides, assignments, data sets, and other useful resources will be posted to the class Google Drive.[5] You must be logged into your USC account to access the materials.

Recorded Zoom meetings and grades will be posted to the class Blackboard site.

USC **prohibits sharing of any synchronous and asynchronous course content** outside of the learning environment. Sharing of course materials will lead to academic integrity sanctions. Specifically, SCampus Section 11.12(B) states

Distribution or use of notes or recordings based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study is a violation of the USC Student Conduct Code. This includes, but is not limited to, providing materials for distribution by services publishing class notes. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the Internet or via any other media. (See Section C.1 Class Notes Policy).

[3] See also the list of software available to USC students at `https://software.usc.edu/`.

[4] `https://www.stata.com/order/new/edu/gradplans/student-pricing/`

[5] `https://drive.google.com/drive/folders/1FM4ITq8C5pWwTfsY13a8xPF37IHw1vbk?usp=sharing`

Synchronous session recording notice

Zoom class meetings will be recorded and provided to all students asynchronously via Blackboard. SCampus Section 11.12(C) prohibits student recording of class meetings. Do not redistribute class recordings.

Office hours will not be recorded.

Readings Overview

Each class meeting has an associated list of suggested readings, drawn primarily from this list of texts. Additional readings will be posted as PDFs to the course Google Drive. Bibliographies for all examples will be included in the reference section of the slide decks. Students should consider their own priorities before purchasing any of the non-required, non-free texts; I recommend them because they are useful, but they are not necessary.

- Studenmund, A. (2017). *A Practical Guide to Using Econometrics*. Pearson Education Limited, Harlow England, seventh edition. **Our primary textbook**. The 6th edition of the Studenmund text can be used instead (successful completion of the course does not require the newest edition), but the 7th edition has incorporated some substantial revisions. Stata companion and Textbook resources website. An e-version can also be purchased directly from Pearson for about $60.

- Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press, New Haven. Approachable overview of causal inference with workhorse estimators in econometrics. Copious Stata examples. Web version is free of charge at `https://mixtape.scunning.com/`.

- Long, J. S. and Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press, College Station, TX, third edition. Excellent coverage of binary and multinomial dependent variable models. Includes authors' own code for straightforward calculation of marginal effects. Does not cover new features in Stata v16, such as `cmset` commands.

- Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. Recommended for PhD students. Works through the mathematics for all major econometric techniques, with examples.

- Kennedy, P. (2008). *A Guide to Econometrics*. Blackwell Publishing, sixth edition. An excellent complement to Studenmund if you would like a second resource for understanding regression.

- Acock, A. (2018). *A Gentle Introduction to Stata*. Stata Press, College Station, Texas, sixth edition. Recommended for students who would like more resources on Stata itself.

Course Schedule: A Weekly Breakdown

Below is a list of topics by weeks and associated methodological readings.

January 21, 2021

Course Introductions

- Studenmund, 6E Ch. 17 (available online at this link).[6]
- Kennedy, chapter 1

January 28, 2021

Regression Analysis: Estimation and Evaluation

- Studenmund, chapters 1-5
- Kennedy, chapter 2 up to 2.7

February 4, 2021

Regression Analysis: Assumptions, Properties, and Model Specification

- Studenmund, chapters 6-7
- Cunningham, "Properties of Regression"
- Cameron and Trivedi, chapter 4
- Kennedy, chapters 3 and 5

February 11, 2021

Heteroskedasticity and Serial Correlation

[6] https://media.pearsoncmg.com/ph/bp/bp_studenmund_econometrics_7/Studenmund6e_ch17.pdf

- Studenmund, chapters 9-10
- Cameron and Trivedi, chapters 11 and 24.
- Kennedy, chapters 8

February 18, 2021

Practical Challenges and Diagnostic Tools

- Studenmund, chapters 8 and 11

February 25, 2021

In-class analysis #1

March 4, 2021

Categorical Dependent Variable Models

- Studenmund, Ch. 13
- Long and Freese, chapters 5–6.
- Cameron and Trivedi, Ch. 14.

March 11, 2021

Experimental Methods

- Studenmund, Ch. 16 (through p. 472)
- Cunningham, "Potential outcomes causal model"

March 18, 2021

Panel Data and Fixed Effects

- Studenmund, Ch. 16 (starting p. 473)
- Cunningham, "Panel data" and "Differences in differences"

- Cameron and Trivedi, chapters 21–22.
- Kennedy, chapter 18

March 25, 2021

In-class analysis #2

April 1, 2021

Multiple-category outcome models

- Long and Freese, chapters 7–8.
- StataCorp (2019). *Stata Choice Models Reference Manual*. Stata Press, release 16 edition. `https://www.stata.com/manuals/cm.pdf`
- Cameron and Trivedi, chapter 15.

April 8, 2021

Project presentations

April 15, 2021

Project presentations

April 22, 2021

USC Wellness Day—No Class

April 29, 2021

Project presentations

Academic Accommodations

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to the instructor (or to a TA) as early in the semester as possible. DSP is located in STU 301 and is open 8.30 AM to 5.00 pm Monday through Friday. Website and contact information for DSP: `http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html` (213) 740–0776 (Phone), (213) 740–6948 (TDD only), (213) 740–8216 (FAX), ability@usc.edu

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" http://policy.usc.edu/scampus--part--b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, http://policy.usc.edu/scientific--misconduct.

Support Systems

Student Health Counseling Services – (213) 740-7711 – 24/7 on call
http://engemannshc.usc.edu/counseling
Free and confidential mental health treatment for students, including short–term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline – 1 (800) 273-8255 – 24/7 on call
http://suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention Services (RSVP) – (213) 740–4900 – 24/7 on call
http://engemannshc.usc.edu/rsvp
Free and confidential therapy services, workshops, and training for situations related to gender–based harm.

Office of Equity and Diversity (OED) | Title IX – (213) 740-5086
http://equity.usc.edu, http://titleix.usc.edu
Information about how to get help or help a survivor of harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following protected characteristics: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations.

Bias Assessment Response and Support – (213) 740-2421
http://studentaffairs.usc.edu/bias--assessment--response--support

Avenue to report incidents of bias, hate crimes, and microaggressions for appropriate investigation and response.

The Office of Disability Services and Programs – (213) 740-0776
`http://dsp.usc.edu`
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy – (213) 821-4710
`http://studentaffairs.usc.edu/ssa`
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC – (213) 740-2101
`http://diversity.usc.edu`
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency – UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
`http://dps.usc.edu,emergency.usc.edu`
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety – UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
`http://dps.usc.edu`
Non–emergency assistance or information.

Guide to the Analysis Project

Half your grade in PPD 558 will be a project applying the tools of multivariate regression analysis to real data. This will give you and your groupmates to put theory into practice by carrying out an analysis and explaining your findings in a presentation and written research report. Your grade will be based on a combination of these final projects as well as intermediate steps designed to ensure your group is making steady, adequate progress.

Developing Your Project

In the initial stages of your project, you will from a group with my help and develop a research topic you are interested in and which will enable you to fulfill the objectives of the assignments.

Considerations for paper topic:

- Value added: your analysis contributes something new and doesn't just replicate the work of others (and does not just use different data with essentially the same model that someone else already developed)
- Non-trivial question: your question needs to be sufficiently challenging to allow you to demonstrate the skills that are the focus of this course (ask yourself: did you really need econometrics to answer your question?)
- Make sure you have a clearly defined research question: if someone asks about your topic and you can only give a very general response like "I'm doing my paper on crime," it's likely that you haven't clearly defined your question. You should be able to say something less vague, like "I'm studying the effects of community policing on city crime rates in the United States." Can your research question be framed as a testable hypothesis?

Everyone needs to be involved the in analysis. It's not acceptable to divide the work in a way that has, e.g., one person's only contribution be writing the introduction and conclusion—some division of labor is reasonable, and some individuals may take a lead role in any given part, but everybody needs to be involved in all of the analytical portions of the project (working on the model, conducting analysis, interpreting results).

Unless you obtain instructor approval to do otherwise, you should use cross-sectional data from either IPUMS or a similarly clean, well-documented, public data source. IPUMS (`https://ipums.org/`) is an outstanding resource for access to large, rich datasets with excellent documentation from the US Census and international equivalents.

If for some reason you want to use data from another source (not recommended!), please discuss this with me in advance—approval is not guaranteed, as searching for an alternative data source, cleaning the data, etc., can be prohibitively time-consuming, and data collection / processing is not the focus of this course. Some other major sources of data you might explore include:

- ICPSR (`https://www.icpsr.umich.edu/icpsrweb/`). This is a huge repository of data files from published papers across the social sciences. Search by keyword.
- Harvard DataVerse (`https://dataverse.harvard.edu/`). Another huge, searchable repository of data from published papers.
- Data Is Plural Archive (Hyperlink) This is a great resource that tracks a wide diversity of interesting data sets—consider joining the mailing list at `https://tinyletter.com/data-is-plural`.
- PSID (`https://psidonline.isr.umich.edu/`). Panel survey of families across a wide range of variables going back decades.
- VoteView (`https://voteview.com/data`). Detailed data on individual legislators and their votes in the US Congress.
- FRED (`https://fred.stlouisfed.org`). Mostly macroeconomic US time series data, but also some sub-federal and international data.

Project Proposal

Submit your proposal in PDF format by emailing it to the instructor and TA. Send one email and CC the other group members. The top of the first page should have a header indicating that this is a project proposal and listing the names of the group members. Please include the following information (maximum 2 pages):

Introduction

- Briefly describe the policy issue, i.e. provide appropriate context to understand the research question you will be working on and its importance.
- Clearly state your specific research question; it should be about the causal relationship between a particular explanatory variable and a particular outcome.
- Identify the intended audience/client; what decision-makers will care about the results of your research, and why is it relevant to them?

Data

Describe the dataset that you will use in the analysis. You do not need to have obtained / cleaned / analyzed the data yet, just identified a dataset that meets your needs (for example, it should contain your dependent variable and explanatory variable of interest).

Empirical Methodology

Unless you obtain instructor approval to use one of the alternative methods covered in the second half of the course, your method of analysis should be OLS regression. List and briefly describe your dependent variable, explanatory variable of interest, and other potentially important control variables.

Timeline

Provide a weekly timeline for the work that you and your team members will conduct through the end of the semester.

The introduction will be graded on a 0–5 scale, with one point for completion of each of the above subparts and a fifth point for overall compliance with these directions. Late submissions will be penalized as described in the grading section of the syllabus.

Data and Methods Report

The data and methods report is designed to ensure that groups are on track to carry out their analysis. You are still free to modify your dataset and model after this; I just want to make sure that you have successfully obtained data that will allow you to perform your analysis and have completed the initial design of your regression model.

Again, submit your report in PDF format by emailing it to the instructor and TA. Send one email and CC the other group members. The top of the first page should have a header indicating that this is a project proposal and listing the names of the group members.

Please include the following information (maximum 2 pages):

Data

- Describe your data source
- List and briefly describe all of your dependent and independent variables
- Provide a table of descriptive statistics (of the sort produced by the summarize command in Stata) that includes all of the variables in your model

Empirical Methodology

- Write out the primary regression model that you plan to estimate
- Provide a justification for each the variables included in your model (as well as the functional form that you have used for each variable – logs, polynomials, interactions, etc.)
- Describe any supplementary analyses (e.g., diagnostic tests) that you plan to carry out and the purpose of each (what will you learn from these analyses?)

Grading will be on a 0-5 scale, with two points for each of these sub-parts and one point for overall compliance with these directions. Late submissions will be penalized as described in the grading section of the syllabus.

Preliminary Results

Please include the following information:

Empirical Methodology

- Write out your primary regression model in the form of an equation
- Describe and justify any changes that have been made to the model that you previously provided in the Data and Methods Report

Results

- Present your OLS regression results (in the form of a standard table of regression output)
- Provide a clear interpretation of your coefficient estimates, statistical significance of these estimates, and any other pertinent information (e.g., R2)
- Present the results of any supplementary analyses (e.g., diagnostic tests) that you have conducted, and interpret the results
- Describe your initial findings (i.e., what implications do your preliminary results have for your research question?)

Timeline

- Briefly explain any challenges that you are dealing with (if any) and how you plan to resolve them, along with any other details relevant to your plans for successful completion of the project.
- Provide an updated timeline for completion of the remainder of the project. Please include your original timeline submitted in your Project Description along with the activities that you have performed or expect to perform each week through final submission. Don't forget to allocate time for "non-analysis" tasks such as preparing your in-class presentation (both creating your slides and practicing your presentation), creating and formatting any tables and graphs for inclusion in the final paper (raw Stata output is fine for rough drafts, but will not suffice for the finished product!), writing / editing / proofreading the final draft, etc.

This assignment will be graded on a 0-5 scale, with 1 point for methodology, 2 points for results, and 1 points for timeline and 1 for overall compliance. Late submissions will be penalized as described in the grading section of the syllabus.

Presentations

Must submit slides (pdf format, not Powerpoint) via Blackboard by 11:59am Pacific Time on day of presentation. Each group will give a 12-15 minute planned presentation, followed by Q&A with me and with class members. Groups who fail to submit their materials on time, and/or present in their assigned slot, will necessarily receive a zero so *please* plan ahead and work to make sure you are able to submit this assignment on time!

Since presentations are remote, you may if you wish submit your presentation as a movie instead of a live presentation over Zoom; in this case, you must submit both a PDF of any slide deck and the movie file for grading.

All group members must present a non-trivial portion of the project.

Final Paper

The final paper is due, in PDF format, at the end of the semester. Each group should submit a single assignment to the instructor and TA with one member CCing other group members on the emailed submission; there is no need for each group member to submit a copy. I strongly urge you to submit your paper before the deadline.

Your final paper should be modeled after the research papers written and circulated by policy and academic institutions worldwide. The paper must not exceed 10 pages, excluding tables, figures,

and references. Use standard formatting: double-spaced text, 12 point Times New Roman or similar font, 1-inch margins, and so on—don't play silly games with this.

Your paper must include the following sections (in this order):

1. A brief introduction (typically around 1 page, no more than 2) that describes the motivation for the analysis and basic background on the problem you are addressing. If a brief literature review is important for context, this is the place for it (and the introduction section, literature review included, should still be at most 2 pages), but a literature review is not required.

2. A detailed description of your data, their sources, and their reliability. Any limitations of the data should be described here.

3. A description of your methods, justification for your model (included variables and other model specification choices) and assumptions required for the results to be valid.

4. A discussion of the results of your analysis (including your regression analyses and any supplementary results such as diagnostic tests) and any limitations or concerns about its validity.

5. A brief conclusion section that clearly lays out your assessment of the implications of your analysis for the research question that you addressed; if there are relevant policy implications, specific issues relating to generalizability of your findings, etc., this is the place to summarize them. Remember that "conclusions" should be understood to mean "conclusions that can be drawn from the analysis presented in this paper" – this is not the place for speculative claims that are not based on the evidence you have provided!

6. A references section that identifies the sources of any materials cited in your paper.

7. An appendix that provides a table of descriptive statistics for all of the variables used in the analysis, all of your regression results presented in standard tabular form (including relevant information such as point estimates, standard errors, etc.), and any other output (tables, figures, etc.) from additional analyses conducted (e.g., diagnostic tests).

Make the case that the key assumptions required for successful causal estimation via OLS are satisfied by your model. If they're not, explain how you're addressing those issues (e.g., using heteroskedasticity-robust standard errors). If you cannot solve all of the potential problems (it's likely that you will not be able to solve them all), you should identify any unaddressed issues, explain why they could not be addressed, and discuss how they are likely to impact your results (e.g., analyze the expected sign of any bias in your estimates)

Justify your model specification – why each of your variables was included, why any variables that might seem like reasonable candidates were not included, why you chose the functional

forms that you did, etc. Conduct sensitivity analysis if alternative specifications seem sensible in order to see if the results are robust to reasonable modifications to your model (e.g., adding or dropping particular control variables, changing functional form, defining variables somewhat differently). If the key conclusions are sensitive to these sorts of changes, this can indicate problems.

To be more specific about that results section and "limitations or concerns about its validity": it is likely that some or many groups will find that their analysis has unanticipated flaws or weaker results than the group initially hoped, and students worry that they will receive bad grades. Fear not: the goal of this assignment is not to carry out a perfect analysis in a single semester; it is to apply and demonstrate mastery of the tools we learn in this course. If you and your group find yourselves in this situation, you have two options: either (1) take the appropriate steps to fix these problems, or (2) discuss how you could fix them in principle, why you can't do it in this particular case, and how your results are likely to be affected.

The first option (fixing the problem) is of course the ideal response to an identified problem, but the second option is perfectly acceptable in the context of this assignment. Keep in mind that the results produced by your analysis aren't the ultimate goal of this project as far as grading is concerned; the goal is for you to demonstrate your understanding of the course material. The fact that your analysis has limitations (and it will!) is not a problem in and of itself; this is only problematic if you fail to identify the problems that (given the material we covered in this course) you can be reasonably expected to detect (the sorts of problems you should be watching out for are not a secret: most of the semester is spent discussing what these potential problems are, how to detect them, how to solve them, and what their consequences are if not fixed!), don't attempt to address problems you have identified, or fail to discuss whatever problems you are ultimately unable to fix.

For example, if you have identified what you believe is likely to be a relevant omitted variable: in principle, how can you solve the problem created by the omission of a relevant variable? (Include it, include a proxy, etc.) Why weren't these solutions feasible in your situation? What are the consequences of this problem, given that you were unable to address it? Be as specific as possible; e.g., rather than simply stating that omitting a relevant variable can lead to bias, discuss the expected bias in this particular situation.

Replication File

A replication file allows any other analyst your researcher recreate your results in full from your source data. A good replication archive clearly documents what the researchers did, enabling others to confirm, verify, or extend your work. Planning to produce a replication archive is also good practice while you're working in a group with others, since it forces you to keep files organized and clearly documented.

Your replication file must be submitted as a single compressed directory in `.zip`, `.7z` or `.tgz` format. You may send it to me directly by email, or if the file size is too large, as a shared file via your USC Google Drive account. The directory should name the group members directly so that it will be distinguishable as I am grading multiple projects, that is, it should be named something like `PPD558 Ali Singh Wang.zip`. When I decompress that file, the folder should include the following subdirectories and files

- A *data* folder containing the raw, unchanged data files you obtained from IPUMS or elsewhere.
- A *code* folder containing you Stata-do files. These should include a main do-file that sets the parameters for your project and calls each of the other do-files in the correct order to process the raw data and carry out your analyses.
- A *temp* folder where you can store data after you have processed it. Do not save processed data in your data folder, and *never* overwrite your downloaded data with the processed data! You might need to change the way you process the data if you find a mistake.
- One or more *output* folders where you save tables and figures you generate in your analysis.
- You might find it helpful to go beyond this simple system and add more stuff. Do you want to have a "read me" file in your main directory where you keep track of your own decisions? A folder for log files? Go for it! It's up to you.

To read more about good practices, I strongly recommend reading Gentzkow, M. and Shapiro, J. M. (2014). *Code and Data for the Social Sciences: A Practitioner's Guide*. Available at `https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf`. I have also posted this to the Google Drive.

You can also download an example replication file I have created from the course Google Drive. It is posted to the "Resources" folder.

I will assign five percentage points of your final course grade on the basis of whether your do-files have clear and consistent explanatory comments (1 point), whether I am able to run the entire work flow from start to finish simply by changing a file path at the start of your code (1 point), and whether that workflow fully replicates your entire work from raw data to tables and figures consistent with those presented in your paper (3 points).