COMPSS 222: Applied Statistics II

Spring 2025

Course Description

This course continues with the development of applied statistical methods appropriate for program evaluation and social science research initiated in COMPSS 212: Applied Statistics I. Topics may include the design of randomized control trials (RCTs), logistic regression, the Bayesian bootstrap, average treatment effect (ATE) estimation under unconfoundedness, local average treatment effects (LATE) and the method of instrumental variables, regression discontinuity design (RDD), quantile regression and discrete hazard analysis. Instruction will involve a mix of theory, illustration and application.

The central limit theorem at the beach!

https://en.wikipedia.org/wiki/Bean_machine

Course Logistics

Instructor: Bryan Graham, Department of Economics, University of California – Berkeley

Email: bgraham@econ.berkeley.edu

Time & Location: Monday 1PM to 4PM, 2121 Allston Way, 2nd Floor

Office Hours: To be determined.

Reader: Jinglin Yang, e-mail: jinglin.yang@berkeley.edu

Prerequisites: COMPSS 212 and a grade of C+ or better in Fall classes.

Units: 3 Credit Hours

Modality: Lecture will be recorded and made available on bCourses (with a lag). These recordings may include student questions and discussion. By enrolling and attending class you implicitly agree to this recording. While lectures will be recorded and made available for review on bCourses, the class modality is "in person". Regular attendance is essential to success in the class and forms a component of your grade. Lecture presents an opportunity for you to ask questions. Your questions also help me adjust lecture material to improve learning outcomes. A portion of lecture time will be set aside for computational demonstrations, exercises and small group work. Please prioritize regular class attendance. By working together we can make class more engaging and productive.

Big Picture: I hope you will find this class interesting and challenging (i.e., difficult). At the end I hope you will feel a sense of accomplishment, as well as ownership over some new and valuable skills. I do *not* want you to find the class stressful. I am mindful that difficulty and stress often go hand-in-hand, but with some thoughtfulness on our part we can avoid this. While I will set and maintain high academic standards, I will also do my very best be supportive, encouraging and helpful. I also strongly urge you – the students – to try to be supportive, encouraging and helpful *to one another*. You'll have more fun (and learn more) if you work together. If a classmate reaches out for help, be generous and offer it. You will not regret it. I do not grade on a pre-set "curve". Class is not a tournament and it is possible for all of you to excel (or not). By helping a classmate you will improve both your own, as well as their, understanding. You will both learn and, also, both do "better" in terms of grades. A related point is that I do not expect you to understand and master every idea introduced in lecture. The aim is to build fluency with a small number of statistical methods, familiarity with a somewhat larger set of tools and only "vague awareness" with a further set of methods. Confusion, appropriately dosed, can be productive. Lean on your classmates, don't be hesitant to ask questions, and please make use of office hours.

Learning outcomes:

1. To be a conversant consumer of quantitive social science research; particularly in the area of program evaluation;
2. To be able to design a randomized control trial (RCT);
3. Conversant in a multiple approaches to covariate adjustment as used to estimate average treatment effects (ATEs) under unconfoundness;
4. Understand both frequentist and Bayesian approaches to measuring statistical uncertainty;
5. Understand the method of instrumental variables (IV);
6. Competency in basic statistical programming in Python and/or R.

Grading: Grades for the class will equal a weighted average of those on homework (75%), class participation (15%) and the scribing assignment (10%). There will be 5 homework assignments. Homeworks are due at 5PM on the assigned due date (the Reader may elect to make small modifications to all things homework related). Homeworks are graded on a ten point scale with one-half point off per day late. In the interest of providing timely feedback, homework will not be accepted after five days from the assigned due date. You are free, indeed encouraged, to work in groups but each student must submit an individual write-up and/or accompanying computer code (please note on your write-up who you collaborated with). Your lowest homework grade will be dropped, with the average of the remaining scores counting toward your final grade. I will add 5 points to homework aggregates for students who make serious efforts to complete all five problem sets (concretely this means that students may amass up to 45 homework points). The due dates for the five problem sets are (exact topics/dates subject to possible change):

Problem Set	**Due Date**	**Topics**	**Dataset**
1	February 10th	Power calculations and minimum detectable effects	NLSY97, Honduras La Moskitia Buzos survey
2	February 24th	Logistic regression, regularization & Baye's	North Carolina recidivism dataset
3	March 17th	Covariate adjustment & Average Treatment Effects	Honduras SAT survey*
4	April 7th	Model Selection & Double Machine Learning	Baseball hitters dataset & SAT data*
5	April 28th	Quantile regression	Brazil PNAD dataset

**Non-publicly available data which will be made available in anonymized form on bCourses. In all cases actual datasets will be posted to bCourses.*

Homeworks are challenging and meant to mimic a "real world" exercise in data analysis. Please *start well before the due date*; doing so allows you to seek out assistance on portions of the assignment where you may require help.

Each of you must "scribe" one lecture. It is expected that you will partner with classmates on this assignment. A LaTex style file will be made available on bCourses for your scribing notes. A well-scribed lecture should summarize key points, include theoretical and computational details where needed and – ideally – move beyond the lecture material to include information from readings and/or discussions on computation in R and/or Python. I will work with scribing groups to ensure that their assignment is correct and maximally useful.

Overall numerical course grades will be calculated as follows:

$$\text{Grade} = 75 \times \frac{\text{Homework Points}}{45} + 15 \times \frac{\text{Class Participation}}{20} + 10 \times \frac{\text{Scribing}}{20}$$

Numerical grades will be mapped into letter grades. A default mapping is 100 - 97 A+, 93 to 96 A, 90 to 92 A-, 87 to 89 B+, 83 to 86 B, 80 to 82 B- and so on. In practice grades are sometimes "curved". I do not curve to enforce a certain grade distribution. In past years I have found that 30 to 40 percent of students earn a grade of A- and above, 40 to 50 percent a grade of B- to B+, with the balance scoring lower. If student performance merits it, I am delighted to award more As, likewise if student learning is less that expected, I will (reluctantly) award fewer As. One thing I want to emphasize is that it is optimal for you to help one another. If you understand the material you will earn a higher grade; helping a classmate will strengthen your understanding and also help them. Class is not a tournament: the *Hunger Games* makes for engaging fiction, but it is a poor model for a learning community.

Textbook: Course materials will be made available on bCourses. There is no official textbook for the class.Virtually all of the assigned readings are available in electronic form via the University Library. I will make an effort to locate and share any unusually difficult to find resources (e..g., chapters from undigitized books). The books by Wickham et al. (2025) and McKinney (2022) provide useful information on computation using R and Python respectively.

Computation: Computational work will be done in either R or Python. Python is a widely used general purpose programming language with good functionality for scientific computing. There are lots of ways of accessing Python. We will use https://datahub.berkeley.edu for computation. More information will be provided in section on how to access and use this platform. For those wishing to manage a Python environment on their personal computer, the Anaconda distribution, which is available for download at

https://www.anaconda.com/products/individual

is a convenient way to get started. Some basic tutorials on installing and using Python, with a focus on economic applications, can be found online at `https://quantecon.org/`.

While issues of computation will arise regularly during lecture, I will not formally teach R and Python programming. *This is something you will largely master outside of class* (although help will be provided in office hours and lecture). I do not expect this to be easy. I ask that those

students with strong backgrounds in technical computing to assist classmates with less experience. Problem sets will be more fun if you all work together and assist each other.

Extensions: Routine extensions for assignments will not be granted (i.e., extensions are for exceptional circumstances only). The penalty for lateness is relatively minor and I also drop the lowest homework grade. These features are designed to allow you some flexibility in workload management during busy times of the semester. Late work, in addition to being undesirable for the individual student, can delay your classmates getting feedback. Please do your best to start work well before the due date.

Accommodations: Any students requiring academic accommodations should request a 'Letter of Accommodation' from the Disabled Students Program at http://dsp.berkeley.edu/ *immediately*. I will make a good faith effort to accommodate any special needs conditional on certification.

Illness, Religious Observances and other excused absences: From time-to-time illness, a religious observance or another important obligation (e.g., participating in intercollegiate athletics) may prevent you from attending class. Please communicate planned absences in advance so that we can partner on any appropriate accommodations. Briefly check in after any illnesses. Lecture recordings and scribed lecture notes will be the primary means of "making-up" missed material.

Academic Integrity: Please read the Center for Student Conduct's statement on Academic Integrity at http://sa.berkeley.edu/conduct/integrity. We should all take issues of intellectual honesty *very* seriously. Cheating, of any type, will not be tolerated. I also encourage you to familiarize and embrace UC Berkeley's Principles of Community.

Additional notes: I prefer to avoid having substantive communications by e-mail. Please limit e-mail use to short (ideally yes/no) queries. I am unlikely to be able to respond to a long/complex e-mail. However, don't be shy about approaching me with questions immediately before/after class (I will make an effort to arrive early and am generally able to linger for a few minutes after). For longer questions please make use of my office hours. This is time specifically allocated for your use; please come by. I look forward to getting to know all of you!

Table 1: **Course Outline** ([r] = "required", [b] = "background")

Date	**Topic**	**Readings**	**Case Studies**
Mon 1/27	Randomized Control Trials (RCTs)	[r] Holland (1986), [r] Duflo et al. (2007) [b] Bloom (1995) , [b] Evans (2023)	(i) Problem Set 1; design of a hypothetical RCT in La Mosquitia (Background #1, #2)
Mon 2/3	Science & Induction	[b] Clayton (2022) [b] Godfrey-Smith (2021, Chs. 1-4)	(i) Castaneda v. Partida, [b] Kaye (1986)
Mon 2/10	Logistic and Discrete Choice Regression	[r] Wooldridge (2010, Chs. 15 & 16) [r] Murphy (2022, Ch. 10), [b] Larson et al. (2016)	(i) Schmidt & Witte (1988) (See Problem Set #2); (ii) Evans et al. (1999) (data extract)
Mon 2/17	**Presidents' Day**		
Mon 2/24	Bayesian Inference & Bootstrap	[r] Murphy (2022, Ch. 10), [r] Xu et al. (2020) [b] Chamberlain & Imbens (2003)	(i) Propublica Compas audit study; (ii) Neal & Johnson (1996) "replication" w/ NLSY97.
Mon 3/3	Covariate Adjustment	[r] Imbens (2004), [b] Robins et al. (1992) [b] Graham et al. (2016, 2012); Graham & Pinto (2022)	(i) McEwan et al. (2015) (See Problem Set 3)
Mon 3/10	Model Selection	[r] Efron (2004), [r] Ye (1998) [b] Efron & Hastie (2016, Chs. 8 & 12)	(i) James et al. (2021, Ch. 8) (See Problem Set 4); (ii) Graham & Powell (2012) calorie demand dataset.
Fri 3/14 *9AM to 12PM*	Double Machine Learning (DML)	[r] Belloni et al. (2014); Chernozhukov et al. (2018) [b] James et al. (2021, Ch. 5)	Knaus (2022) (DML applied to SAT data analyzed in Problem Set 3)
Note there is no class on Monday 3/17, but there is class on Friday 3/14 9AM to 12PM!			
Mon 3/24	**Spring Recess**		
Mon 3/31	LATE and Instrumental Variables (IV)	[r] Angrist et al. (1996), [b] Bloom (1984) [r] Angrist et al. (2000), [b] Borusyak & Hull (2023)	Card (1995), Neal (1997) Evans et al. (1999) Autor et al. (2013) (IV exercise using David Card data)
Mon 4/7	Model-Based IV & EM Algorithm	[r] Imbens & Rubin (1997), [r] Page et al. (2015) [b] Gupta & Chen (2010)	
Mon 4/14	Regression Discontinuity Design	[r] Cattaneo et al. (2024)	RDD datasets
Mon 4/21	Quantile Regression	[r] Koenker & Hallock (2001), [r] Chamberlain (1994) [b] Mood et al. (1974)	(i) Angrist et al. (2006) (See Problem Set 5, illustration with PNAD data from Brazil)
Mon 4/28	Discrete Hazard Analysis	[r] Efron (1988), [b] Jenkins (1995), [b] Singer & Willett (2003, Chs. 9 - 12)	(i) Schmidt & Witte (1988), [b] Larson et al. (2016) (See Problem Set 2, Recidivism and Discrete Hazard Analysis)
Mon 5/5	**RRR Week** (Make-Up Day)		

References

Angrist, J. D., Chernozhukov, V., & Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica*, 72(2), 539 – 563.

Angrist, J. D., Graddy, K., & Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3), 499 – 527.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444 – 455.

Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The china syndrome: local labor market effects of import competition in the united states. *American Economic Review*, 103(6), 2121 – 2168.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608 – 650.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, 8(2), 225 – 246.

Bloom, H. S. (1995). Minimum detectable effects: a simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5), 547 – 556.

Borusyak, K. & Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6), 2155 – 2185.

Card, D. (1995). Earnings, schooling, and ability revisted. *Research in Labor Economics*, 14(23 - 48).

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2024). *A Practical Introduction to Regression Discontinuity Designs*. Cambridge: Cambridge University Press.

Chamberlain, G. (1994). *Advances in Econometrics: Sixth World Congress*, volume 2, chapter Quantile regression, censoring, and the structure of wages, (pp. 171 – 209). Cambridge University Press: Cambridge.

Chamberlain, G. & Imbens, G. W. (2003). Nonparametric applications of bayesian inference. *Journal of Business and Economic Statistics*, 21(1), 12 – 18.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1 – C68.

Clayton, A. (2022). *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. Columbia University Press.

Duflo, E., Glennerster, R., & Kremer, M. (2007). *Handbook of Development Economics*, volume 4, chapter Using randomization in development economics research: a toolkit, (pp. 3895 – 3962). North-Holland: Amsterdam.

Efron, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American Statistical Association*, 83(402), 414 – 425.

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, 99(467), 619 – 632.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.

Evans, D. K. (2023). Towards improved and more transparent ethics in randomised controlled trials in development social science. *Journal of Development Effectiveness*, (pp. 1 – 11).

Evans, W. N., Farrelly, M. C., & Montgomery, E. (1999). Do workplace smoking bans reduce smoking? *American Economic Review*, 89(4), 728 – 747.

Godfrey-Smith, P. (2021). *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press, 2nd edition.

Graham, B. S. & Pinto, C. (2022). Semiparametrically efficient estimation of the average linear regression function. *Journal of Econometrics*, 226(1), 115 – 138.

Graham, B. S., Pinto, C., & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3), 1053 – 1079.

Graham, B. S., Pinto, C., & Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, 31(2), 288 – 301.

Graham, B. S. & Powell, J. L. (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, 80(5), 2105 – 2152.

Gupta, M. R. & Chen, Y. (2010). Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3), 223 – 296.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945 – 960.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86(1), 4 – 29.

Imbens, G. W. & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variable models. *Review of Economic Studies*, 64(4), 555 – 574.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R.* New York: Springer-Verlag, 2nd edition.

Jenkins, S. P. (1995). Easy estimaton methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics*, 57(1), 129 – 137.

Kaye, D. H. (1986). Is proof of statistical significance relevant? *Washington Law Review*, 61(4), 1333 – 1365.

Knaus, M. C. (2022). Double machine learning based program evaluation under unconfoundedness. *Econometrics Journal*, 25(3), 602 – 627.

Koenker, R. & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4), 143 – 156.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the compas recidivism algorithm.

McEwan, P. J., Murphy-Graham, E., Irribarra, D. T., Aguilar, C., & Rápalo, R. (2015). Improving middle school quality in poor countries: evidence from the honduras sistema de aprendizaje tutorial. *Educational Evaluation and Policy Analysis*, 37(1), 113 – 137.

McKinney, W. (2022). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. Cambridge: O'Reilly, 3rd edition.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics*. New York: McGraw-Hill Book Company, 3rd edition.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: The MIT Press.

Neal, D. (1997). The effects of catholic secondary schooling on educational achievement. *Journal of Labor Economics*, 15(1), 98 – 123.

Neal, D. A. & Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104(5), 869 – 95.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, 36(4), 514 – 531.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2), 479 – 495.

Schmidt, P. & Witte, A. D. (1988). *Predicting Recidivism Using Survival Models*. New York: Springer-Verlag.

Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis*. Oxford: Oxford University Press.

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2025). *R for Data Science*. O'Reilly Media Inc., 2nd edition.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2nd edition.

Xu, L., Gotwalt, C., Hong, Y., King, C. B., & Meeker, W. Q. (2020). Applications of the fractional-random-weight nootstrap. *The American Statistician*, 74(4), 345 – 358.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441), 120 – 131.