Applied Econometrics for Fun and Profit

Sahil Chinoy

July 26, 2017

Section 1

Why bother?

- Economists and academics are bad at many things. One of them is graphics.

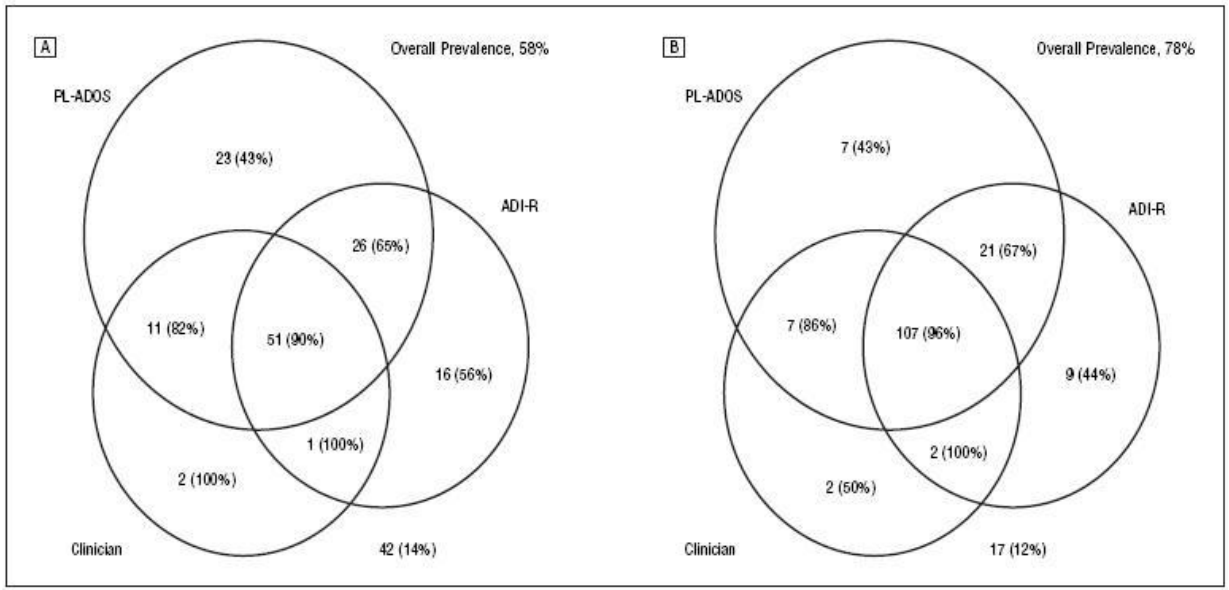
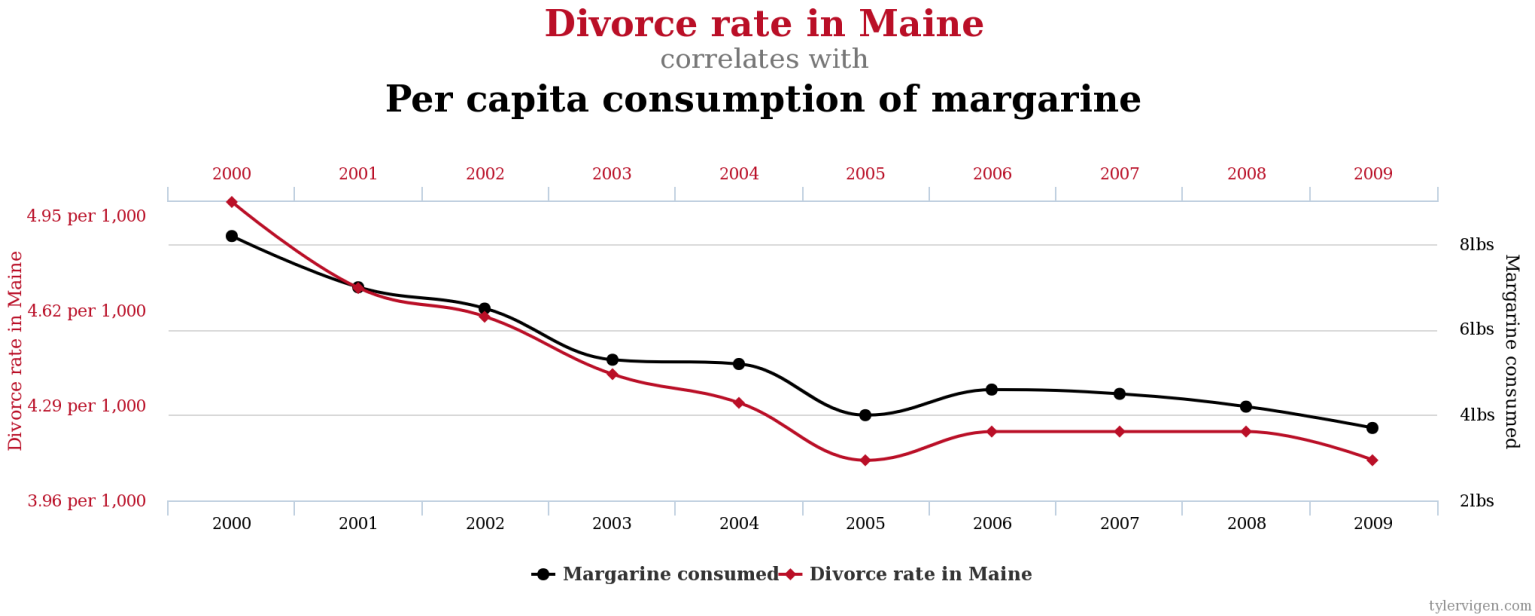


Figure 2. Frequency of diagnostic combinations at age 2 years and prevalence of best-estimate diagnosis (in parentheses) at age 9 years. A, Autism. B, Autism spectrum. PL-ADOS indicates Pre-Linguistic Autism Diagnostic Observation Schedule; ADI-R, Autism Diagnostic Interview–Revised.

- But they are good at two things that journalists can learn.
  - Causal inference.
  - Quantifying uncertainty.
- Why should you care?
  - Dissect studies.
  - Talk to academics in their language.
  - Use econometric tools in your own reporting.

- The standard tool of econometrics is the ordinary least-squares (linear) regression. It's useful and well-understood.
  - But it's easy to abuse to find spurious correlations.[1]
  - Also, it's a little boring.



[1] http://www.tylervigen.com/spurious-correlations

- Today, we'll run through three other techniques in applied econometrics.
  - This will be focused on examples, not theory or coding.
  - It's intended to help you get familiar with the terminology, understand what's possible and know how to get started.

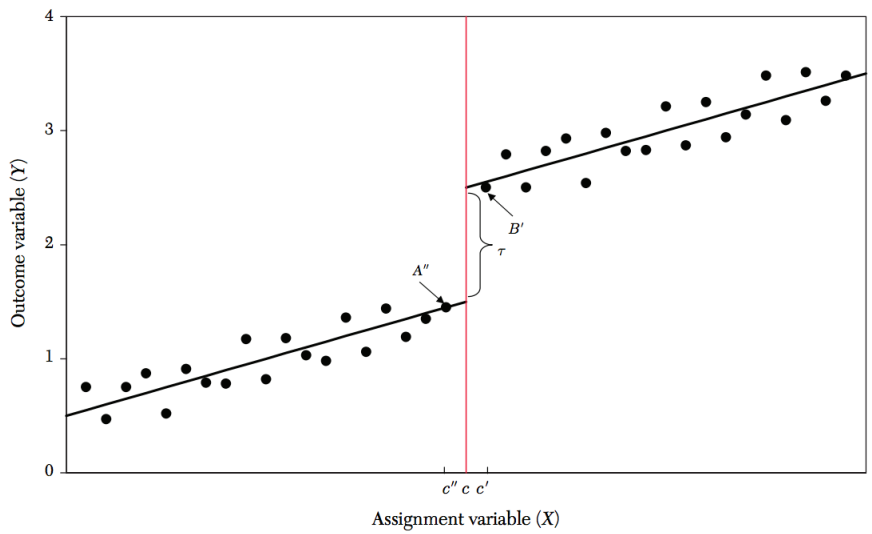
Section 2

Regression discontinuity

- Like researchers, we're often after causality. It's a better story to say that X causes Y, not just that X and Y are related.
- The gold standard: randomized trials.
  - Think military drafts. Because whether you're drafted is essentially random, we can use the difference in outcomes between the two groups to analyze the causal effect of joining the military.
- Often in real life, we don't have that kind of randomness.[2]
- So, we take some sort of cutoff, and assume that if you're close to the cutoff, which side you're on is *basically* random.

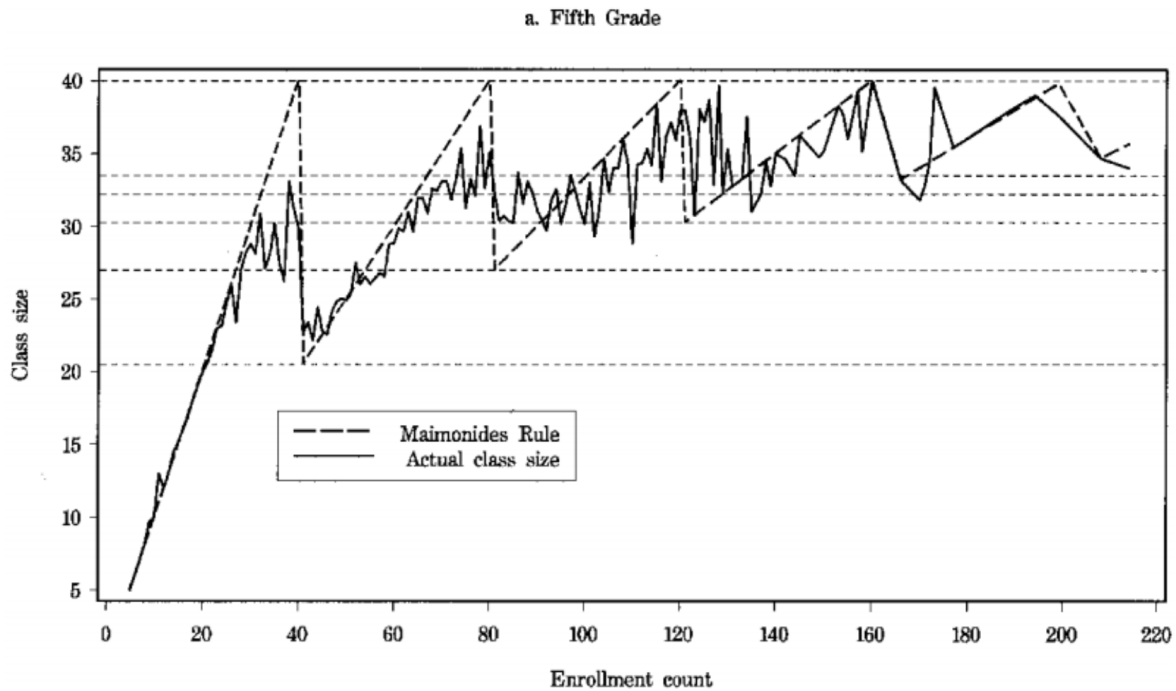
[2] For example, we can't randomly give aid to some poor nations but not others.

- Here's the classic example.[3]
  - Say that if you get a 2000 or above on your SAT, you're guaranteed to get into college, but if you get below 2000, you don't.
  - The difference in ability between someone who scores 1980 and 2020 is basically nothing.
  - So we can use the difference in the earnings of people who scored below 2000 and above 2000 to get the causal effect of college admission on earnings.



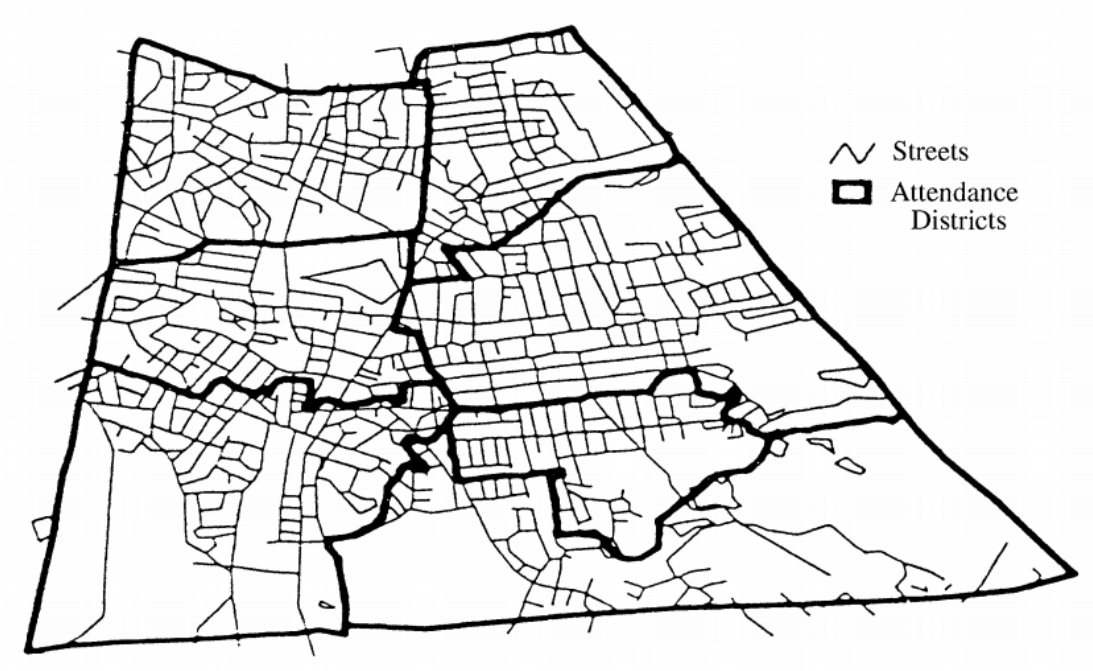
[3] More examples in [Lee and Lemieux, 2010], pp. 339-42.

- Discontinuities due to policy or custom.[4]
  - The maximum number of students per teacher in Israel is 40.
  - But a cohort with 40 students is basically the same as a cohort of 41 students except for how many teachers they have.
  - This can be used to get the causal effect of classroom size on test scores.



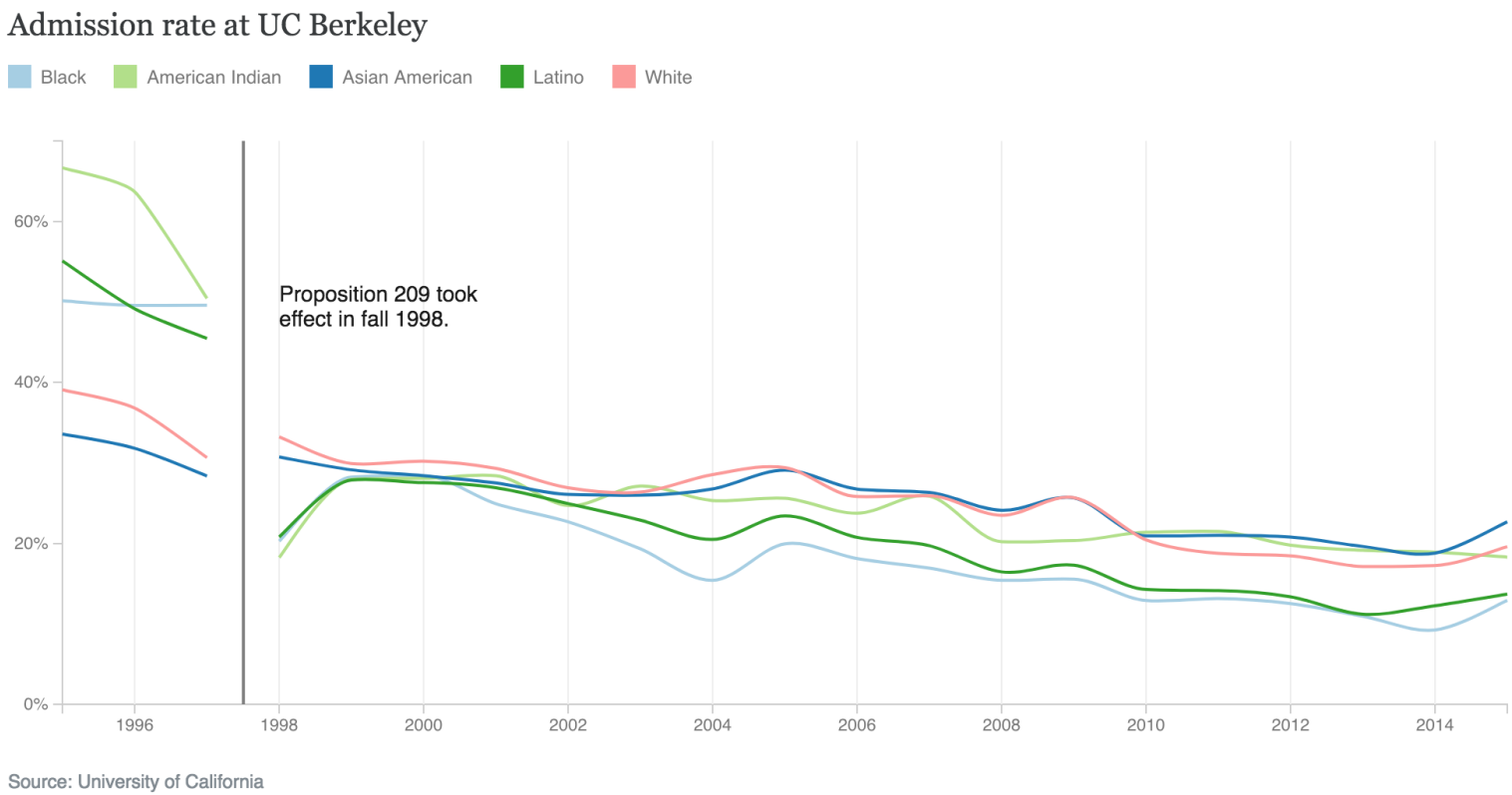
[4] See [Angrist and Lavy, 1999].

- Geographic distance from a border where some meaningful policy change happens.[5]
  - Consider houses on two sides of a school district border. They are basically the same except for the school they're assigned to.
  - This can be used to figure out how much parents are willing to pay to send their children to better schools.



[5] See [Black, 1999].

- How to do it
  - Make a scatterplot, then draw two trend lines on either side of the cutoff.
  - The visual discontinuity can be powerful even without the underlying statistics, but make sure the effect you observe is significant.



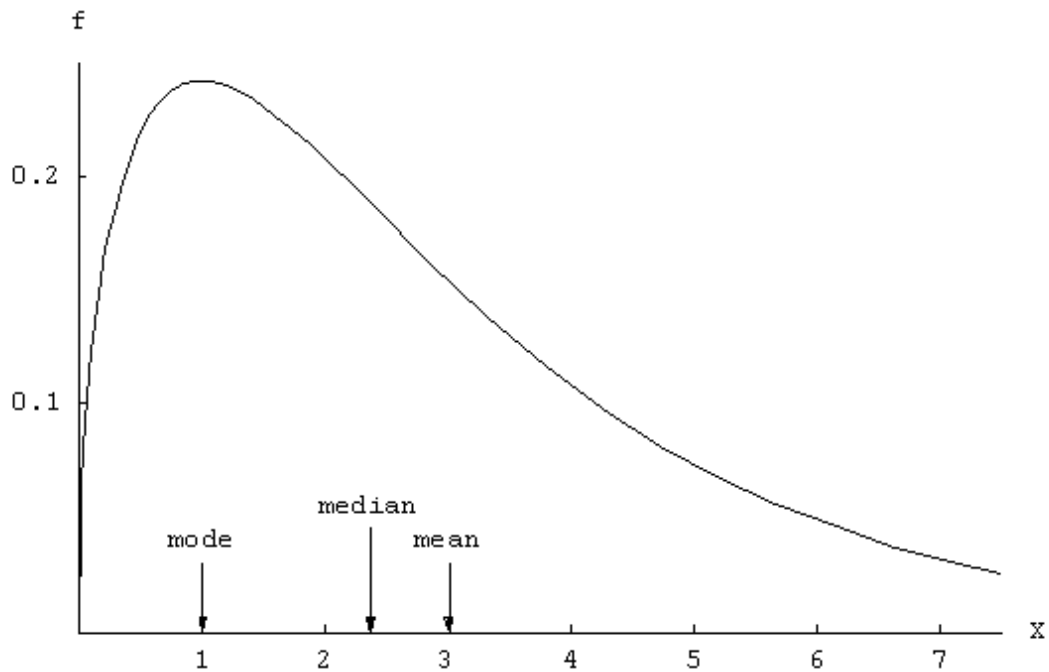
- Run a linear regression with a binary variable to indicate the side of the cutoff you're on. Then check the magnitude and significance of the coefficient.
  - Python: `statsmodels.OLS`.
  - See `github.com/natematias/research_in_python` for an example in a Jupyter notebook.

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	70.1188	1.151	60.908	0.000	67.856	72.382
csize	-0.1392	0.119	-1.168	0.243	-0.373	0.095
small	3.9531	1.800	2.197	0.029	0.416	7.491

Section 3

Quantile regression

- Ordinary least-squares estimates the conditional mean of a dependent variable given particular values of the independent variables.
- But for skewed distributions, the mean can be misleading. Sometimes the median is more informative.
- What if we could estimate the conditional median? Or the 25th percentile, or the 99th percentile...

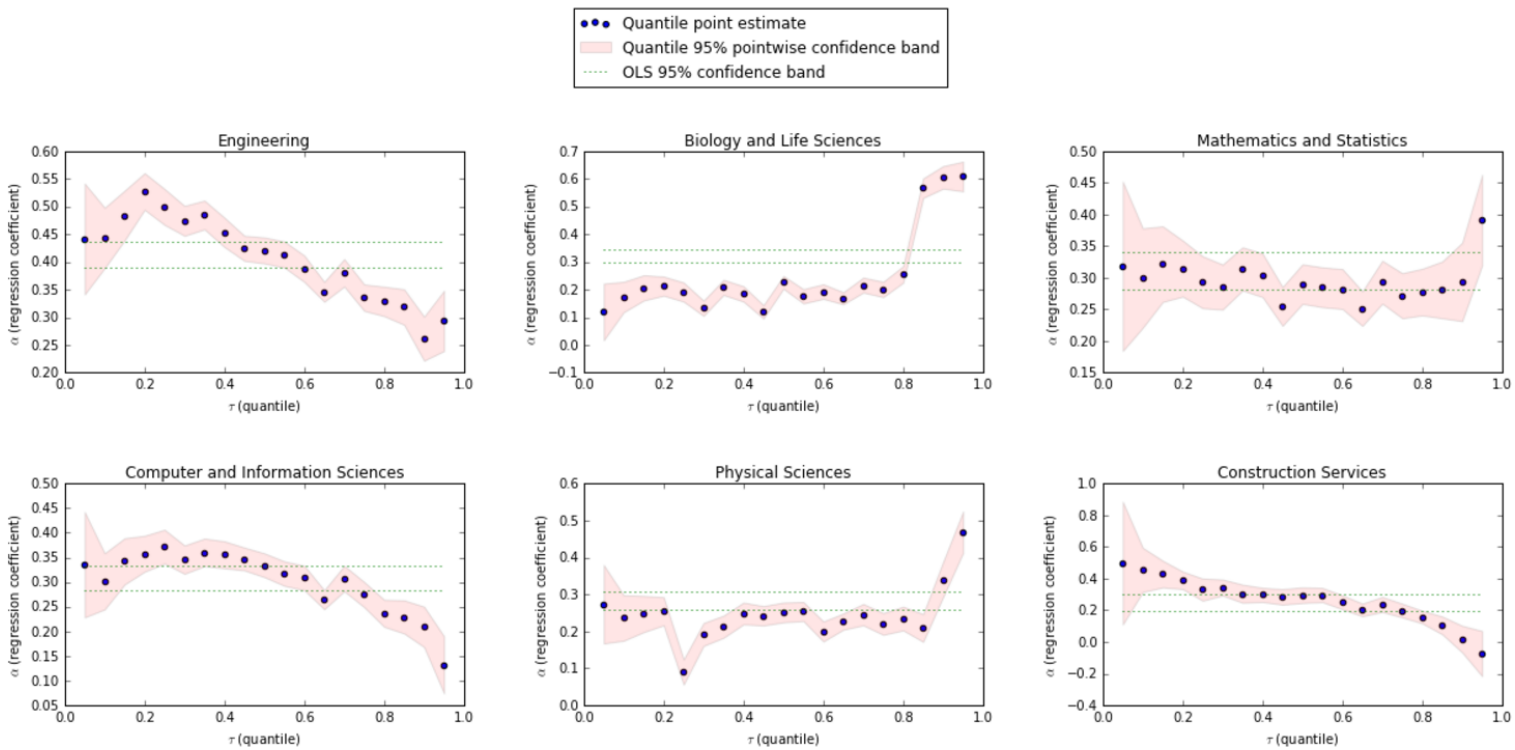


- Example: I want to know what to major in to boost my future earnings.
- First attempt: Average earnings by major.
  - But men earn more on average, and engineering majors are more likely to be male.

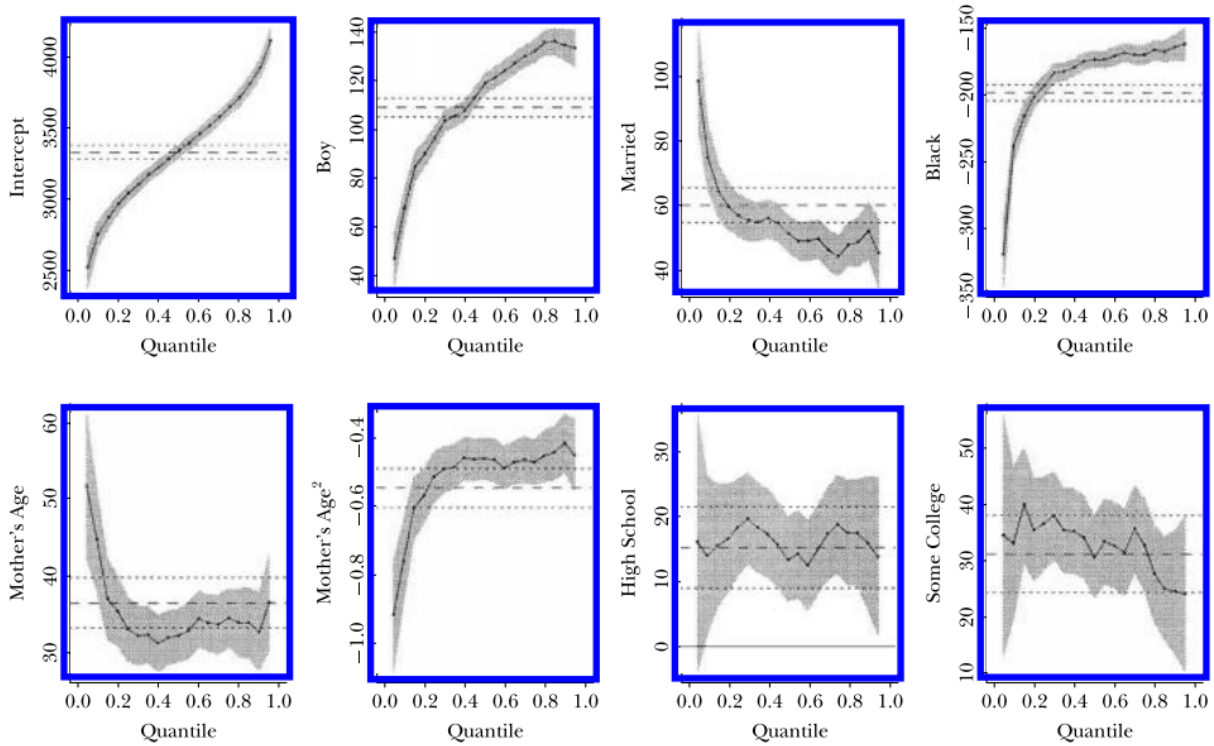
	INCWAGE
DEGFIELD	
Engineering	111888.19
Biology and Life Sciences	104841.85
Physical Sciences	102970.25
Mathematics and Statistics	100254.41
Computer and Information Sciences	95371.24
Social Sciences	95007.78
Transportation Sciences and Technologies	91478.05
Business	89170.16
Construction Services	88169.26
Engineering Technologies	87997.63

- Second attempt: Linear regression to account for gender, race and age.
  - Engineering Technologies drops out of the list because it is skewed male.
  - Biology still appears second on the list.
  - Doctors make a lot of money, but many biology majors end up as medical technicians.
  - Maybe we want to look at the median earnings. Or better yet, maybe we want a fuller picture of the distribution of earnings.

- Third attempt: Quantile regression. What is the effect on earnings *at some quantile* of majoring in X (relative to some reference major), controlling for race, gender and age?
  - Earnings at the 95th percentile for biology majors look very different than the median.



- Effect of mother's race on birthweight.[6]
  - At the lowest quantiles, babies born to black mothers weigh 300 g less than those born to white mothers. At the top quantiles, the difference is more like 150 g.



[6] See [Koenker and Hallock, 2001].

- Ethnicity and wealth across 19th-century American cities.[7]
  - At the lowest quantiles, German workers in Indianapolis are far wealthier than others. At the top quantiles, this effect disappears.

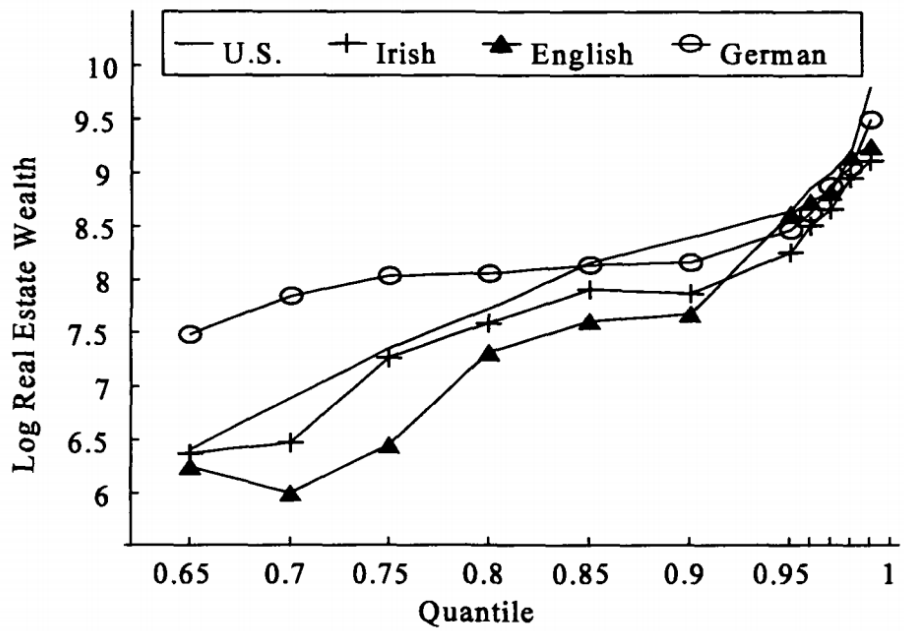


FIGURE 4
ESTIMATED REAL ESTATE WEALTH FOR SKILLED WORKERS IN INDIANAPOLIS

[7] See [Conley and Galenson, 1998].

- Effect of unions on wages.[8]
  - At the lowest quantiles, unions make a big difference in wages. At the top of the wage distribution, unions hardly matter.

	Quantile				
	0.10	0.25	0.50	0.75	0.90
A Manufacturing					
Union					
λ	2.293	1.720	2.169	2.760	−0.793
	(0.340)	(0.855)	(0.661)	(0.778)	(0.581)
Ratio	0.725	0.878	0.824	0.785	1.256
Non-union					
λ	0.827	0.770	−0.928	−0.440	−1.421
	(0.337)	(0.313)	(0.358)	(0.441)	(0.393)
Ratio	1.065	1.082	1.845	1.498	1.898
Union wage effect[a]	0.292	0.272	0.201	0.117	0.041
	(0.011)	(0.013)	(0.012)	(0.012)	(0.010)

[8] See [Chamberlain, 1994].

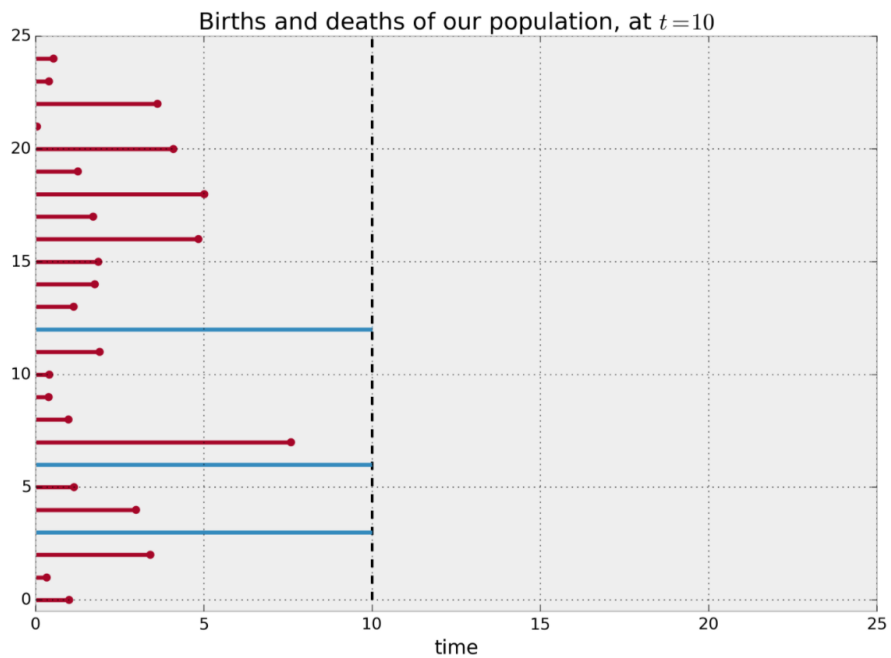
- How to do it
  - Pretty much like an OLS regression computed for a given quantile.
  - Works by dividing your sample into "cells" for each combination of the explanatory variables, so it requires a large sample size.[9]
  - Python: `statsmodels.quantReg`.
  - See me for the college majors example in a Jupyter notebook.

[9] Sort of.

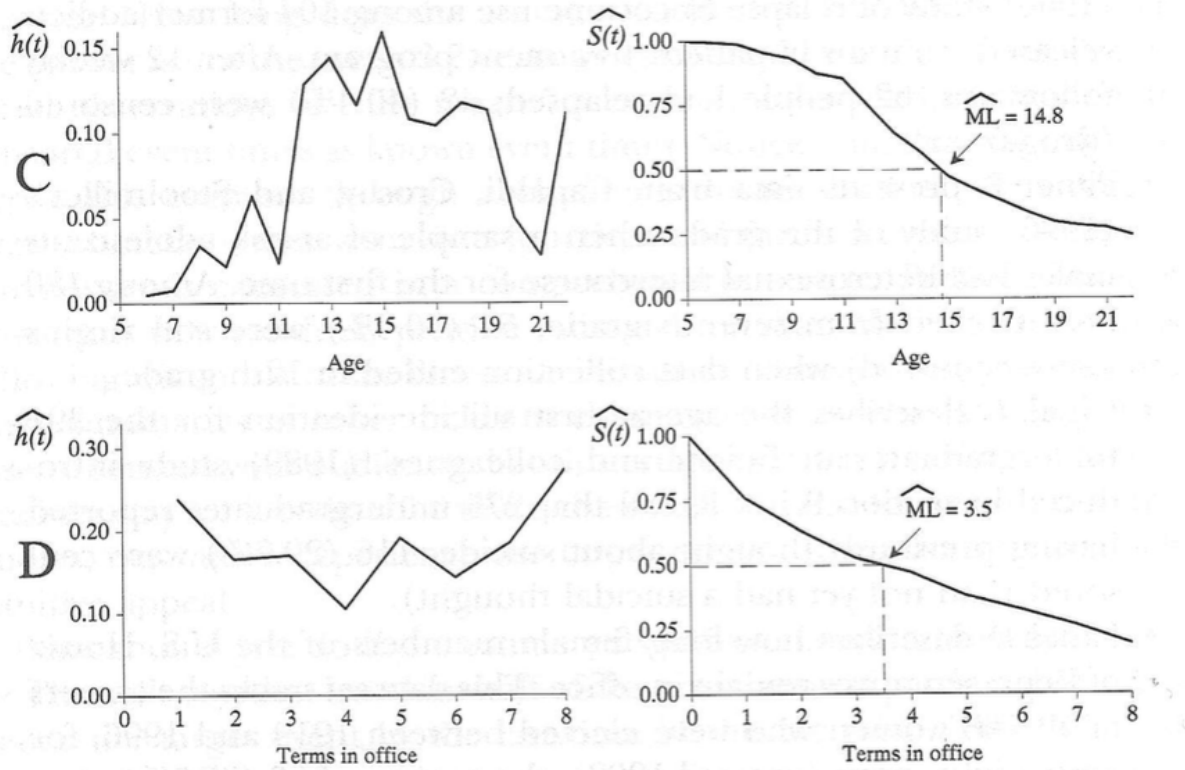
Section 4

Survival analysis

- We're interested in durations and how they differ across groups.
  - Length of a politician's time in office for different parties.
  - Patient lifetime after heart surgery for different doctors.
- The problem: We don't observe the "death" event for every individual.
  - This is known as censorship, and will lead us to underestimate the true lifetime.

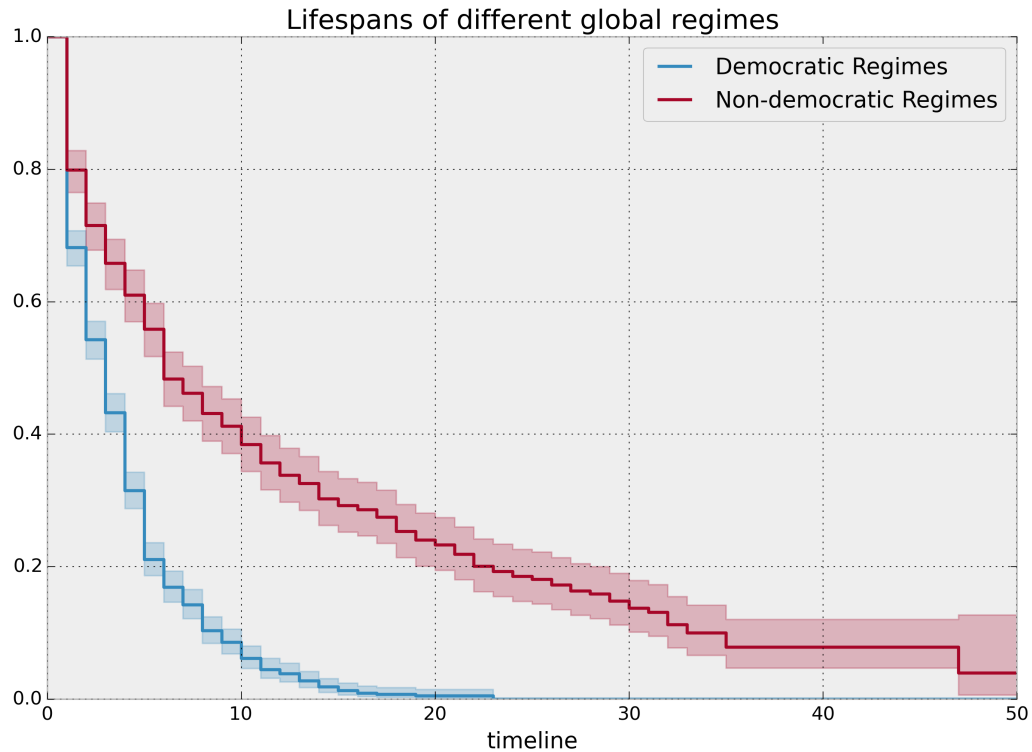


- Survival analysis is essentially a bag of tools to get around the censorship problem.
- It lets us estimate the *survival function*[10] – the fraction of the population that survives past a given time – and the *hazard rate* – the probability of "dying" at a given time.



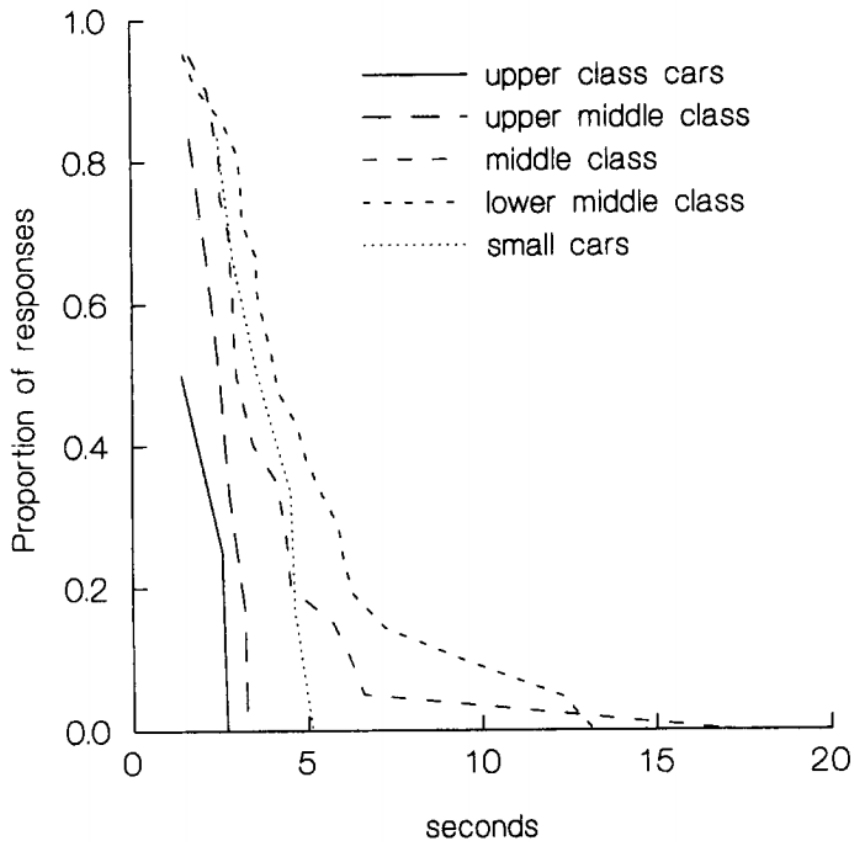
[10] Known as the Kaplan-Meier estimator.

- We can run a survival regression to analyze how the baseline hazard rate changes with some explanatory variable.
- But it's sometimes enough just to look at the survival functions for subgroups.
    - Consider the lifetime for democratic and nondemocratic regimes.[11]



[11] `lifelines.readthedocs.io`

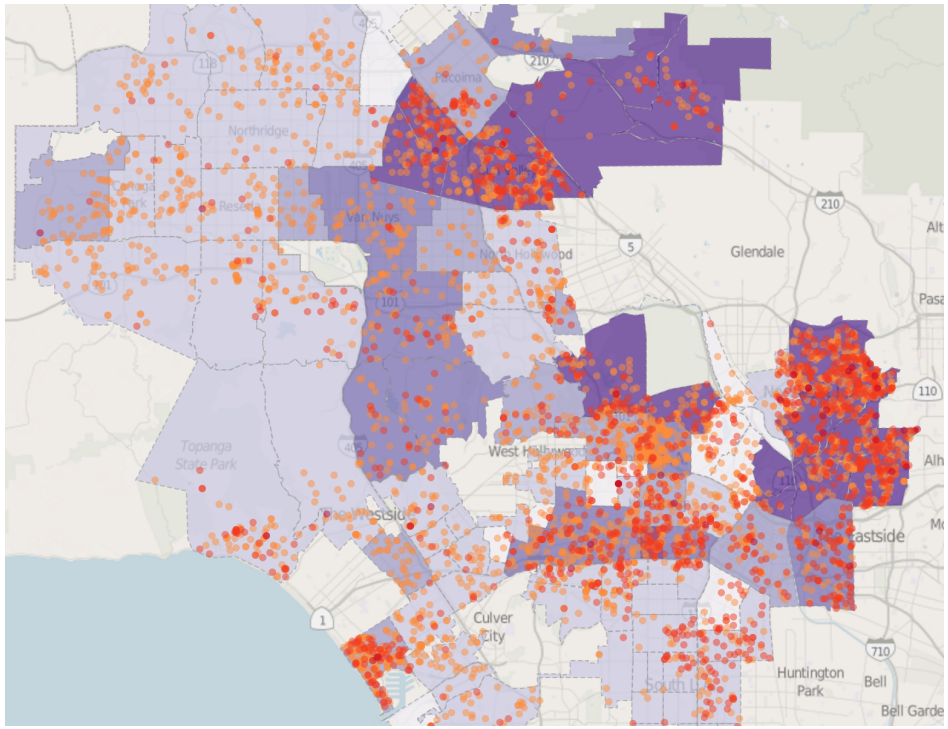
- Rich drivers are more likely to honk faster when they're blocked at an intersection.[12]



[12] See [Diekmann et al., 1996] (but note that $N = 57$).

- Armand (et al.) used this to look at building safety complaints in Los Angeles.[13]

  "Last year, the citywide median wait for a visit from an inspector or other initial action by the building department following a complaint was eight days – compared with 26 days in the eastern parts of the city."



[13] latimes.com/local/cityhall/la-me-building-safety-delay-20141219-story.html

- How to do it
  - The Python `lifelines` package is good and well-documented.
  - Use the `KaplanMeierFitter` to estimate the survival function and the `CoxPHFitter` to run a survival regression.
  - Median lifetime is an intuitive and powerful statistic.[14]

[14] But nothing is easy, see [Singer and Willett, 2003] p. 347 for caveats.

Section 5

References


Angrist, J. D. and Lavy, V. (1999).
Using Maimonides' rule to estimate the effect of class size on scholastic achievement.
The Quarterly Journal of Economics, 114(2):533–575.


Black, S. E. (1999).
Do better schools matter? Parental valuation of elementary education.
The Quarterly Journal of Economics, 114(2):577–599.


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


Conley, T. G. and Galenson, D. W. (1998).
Nativity and wealth in mid-nineteenth-century cities.
The Journal of Economic History, 58(2):468–493.

Diekmann, A., Jungbauer-Gans, M., Krassnig, H., and Lorenz, S. (1996).
Social status and aggression: A field study analyzed by survival analysis.
The Journal of social psychology, 136(6):761–768.

Koenker, R. and Hallock, K. (2001).
Quantile regression: An introduction.
Journal of Economic Perspectives, 15(4):43–56.

Lee, D. S. and Lemieux, T. (2010).
Regression discontinuity designs in economics.
Journal of Economic Literature, 48(2):281–355.

Singer, J. and Willett, J. (2003).
Describing discrete-time event occurrence data.
Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. New York: Oxford University Press, Inc, pages 325–56.