Academic Writing 1
for Accounting, Economics & Psychology

Session 6: Main Body

Prof. Dr. Dominika Langenmayr

Summer 2021

Agenda

1. Empirical Papers
2. Tables
3. Visualization
4. Theoretical Papers

Structure

Most theory papers have a structure very similar to the following:

1. Introduction
2. Possibly: Theoretical Framework or Background
3. Data and Descriptive Statistics
4. Empirical Framework
5. Results and Discussion
6. Conclusion

The structure is not set in stone. Quite often you see 4 before 3.

Identification Section

- The core of an empirical paper. It can be quite long!
- Structure
  1. Ideal experiment
  2. Deviations
  3. Potential threats to identification (and how you solve them)
  4. Equations (numbered!)
  5. Explain all variables, indices, and error term assumptions
- Use all supporting evidence, also qualitative or anecdotal.
- Tip: Read identification sections using similar methods to yours. Craft your text after it.

Further Reading

- Bellamare, M. (2020): How to Write Applied Papers in Economics.

Agenda

Importance

- Tables are the most common way to present empirical research results.
- Many readers will only read the introduction and then directly go to the tables.
- Each table (or figure) needs self-explanatory notes, often including data sources.
- These need to include all relevant information; the reader must be able to understand the table without referring to the main body of the paper.

Table Notes

Table 1: Local Labor Market Effects of Bonus Depreciation

	(1)	(2)	(3)	(4)	(5)
Employment					
Exposure × Post	0.021**	0.019***	0.021***	0.018***	0.020**
	(0.009)	(0.007)	(0.006)	(0.006)	(0.008)
Earnings					
Exposure × Post	0.022**	0.023***	0.019***	0.017**	0.021**
	(0.011)	(0.008)	(0.007)	(0.007)	(0.010)
Earnings-per-Worker					
Exposure × Post	-0.002	0.000	-0.005**	-0.004**	-0.003
	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
3-digit Industry-by-Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
State-by-Year Fixed Effects		Yes	Yes	Yes	Yes
County Characteristics			Yes	Yes	Yes
Winsorized Weights				Yes	
Dropping Small County-Industries					Yes

Notes: This table shows difference-in-differences estimates from Equation 2 where β is not allowed to vary by year. The outcomes are employment in the first row, earnings in the second, earnings-per-worker in the third. Column (1) shows estimates with only 3-digit NAICS industry-by-year fixed effects while column (2) adds state-by-year fixed effects. Column (3), the main specification, adds county level economic and demographic characteristics as control variables. The last two columns show robustness of the results to winsorizing the weights at the 5% level and to dropping county-3-digit industries with less than 1,000 workers in 2001. The sample for this table excludes 2002 as a transition year. Standard errors are clustered at the county level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Source: Garrett et al. (2020)

Some Tips

- Font size must not cause eye strain.
- Must convey takeaway within 10 seconds, without main text.
- Avoid acronyms. Really!
- Also avoid equation symbols or variables names without words to accompany them.
- Label every column header.
- Use as few decimal places as possible. Keep overprecision issues in mind.
- If using LaTeX, use the booktabs package.
- Do you really want table? Perhaps a figure is better?

Agenda

General Thoughts

Two reasons for graphing your data:

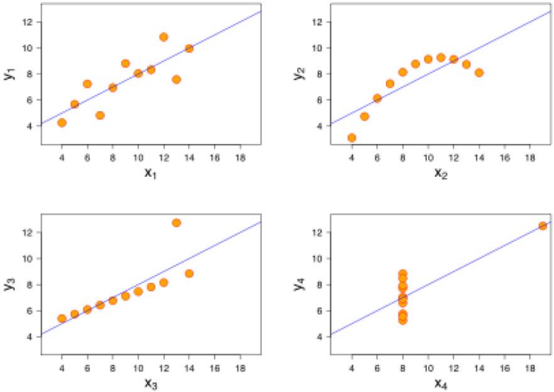
1. Graphical methods are important for understanding your data,
2. Good visualizations are key for getting your story across and making it memorable

   Your paper should have...

   - Fewer tables and more figures
   - At least one really good figure that summarizes the paper (for tweets etc.)

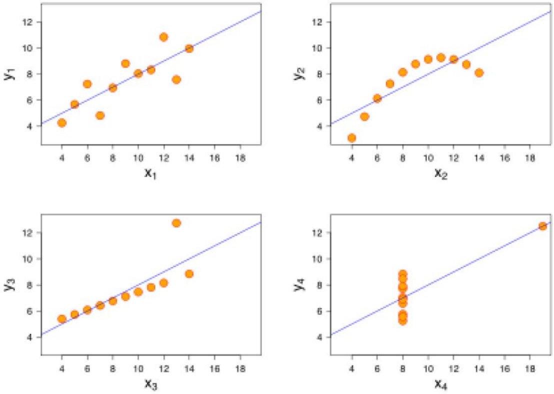
- The following slides are largely based on Paul Goldsmith-Pinkham's slides available here.

Why use plots



Source: Denny (2021)

Why use plots



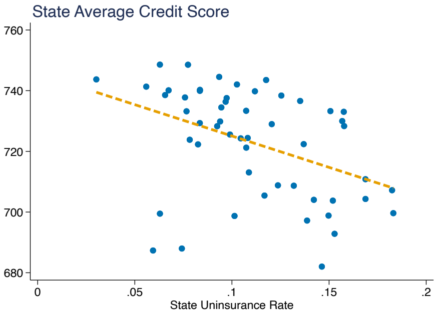
Source: Denny (2021)

- All x variables have the same mean and variance.
- Same is true for all y variables and their covariance.

$\rightarrow$ Same regression line, but relationships between x and y are very different!

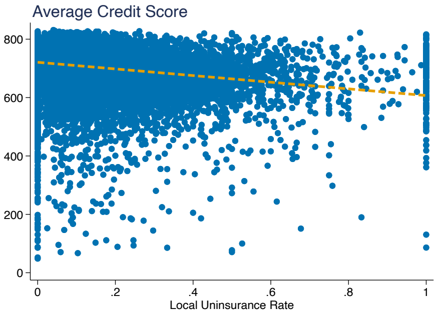
Visualizing a relationship

- Intuitively, for many papers, we plot an outcome Y_i and want to describe the effect of D_i
- The line is a useful summary description of it, but the data already does a pretty good job. Why do we need the line?



Visualizing a relationship

- Intuitively, for many papers, we plot an outcome Y_i and want to describe the effect of D_i
- The line is a useful summary description of it, but the data already does a pretty good job. Why do we need the line?
- Well, sometimes we have a LOT more data and it's harder to see the relationship
- The line is an excellent summary



Binscatter approach

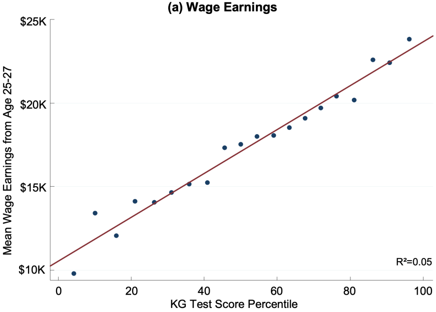
$$Y_i = f(D_i, \theta) + \epsilon_i$$

- There are a number of ways to approximate this function in the econometrics literature
  - One common approach is called *binscatter*, which uses spaced bins to construct means
- Why is this useful? Well, much of the time in our plots it is hard to see the underlying conditional expectation function.
- The dots reflects averages within 20 equally spaced quantiles

Source: Chetty et al. (2011)

Binscatter approach

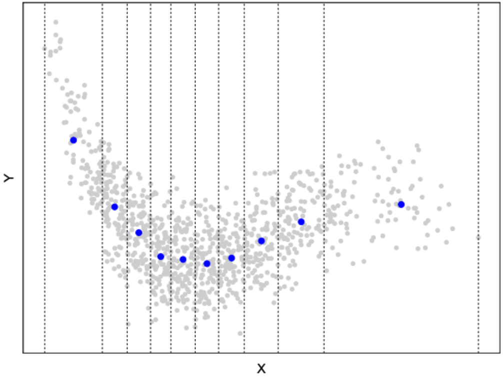
- Two things worth noting from this (very nice) graph
  - The R^2 is not enormous, which suggests lots of unexplained variation
  - We don't have a good reason for the bin choice
- In a discrete case, the bin choice is obvious
- So what's going on under the hood?



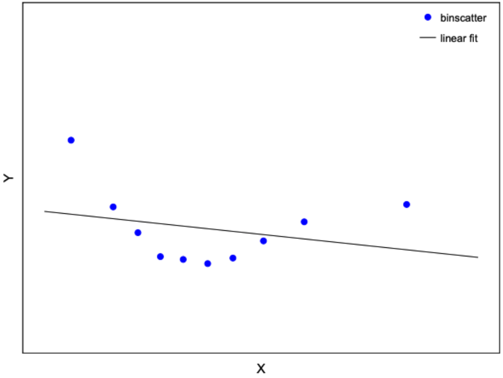
Source: Chetty et al. (2011)

How a binscatter graph is made (Cattaneo et al. (2019)

Figure 1: The basic construction of a binned scatter plot.



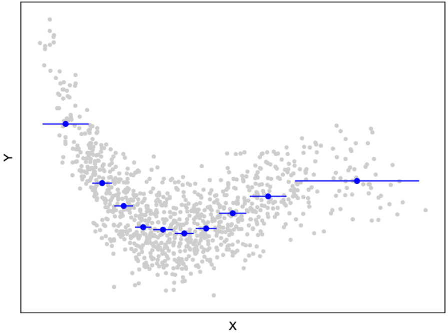
(a) Scatter and Binscatter Plots



(b) Binscatter and Linear Fit

Cattaneo et al. “On Binscatter”

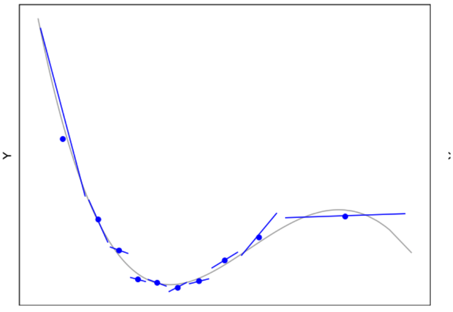
- Paper provides several generalizations to binscatter approach
- First contribution: highlight that the “traditional” binscatter approach is presenting a particular non-parametric estimation
- Initially assumes that constant within bin
  - Not crazy! But could do more.
- Piece-wise functions can be made very flexible



(a) Binned Scatter Plot with Piecewise Constant Fit

Cattaneo et al. “On Binscatter”

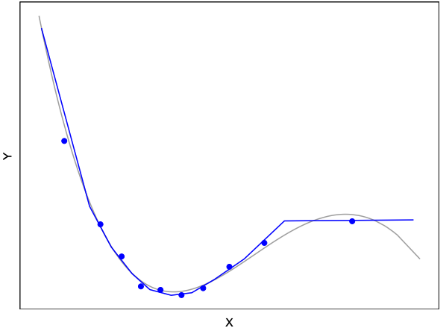
- Paper provides several generalizations to binscatter approach
- First contribution: highlight that the “traditional” binscatter approach is presenting a particular non-parametric estimation
- Initially assumes that constant within bin
  - Not crazy! But could do more.
- Piece-wise functions can be made very flexible



(a) $p = 1$ and $s = 0$

Cattaneo et al. "On Binscatter"

- Paper provides several generalizations to binscatter approach
- First contribution: highlight that the "traditional" binscatter approach is presenting a particular non-parametric estimation
- Initially assumes that constant within bin
  - Not crazy! But could do more.
- Piece-wise functions can be made very flexible



(b) $p = 1$ and $s = 1$

Cattaneo et al. "On Binscatter"

- Second contribution: Choosing bins!
- Reframe as non-parametric problem. Estimation problem is tradeoff:
  - bias (picking too few bins makes your function off)
  - and noise (pick too many bins and they're very noisy)
- In canonical binscatter, $\approx n^{1/3}$
  - This is data driven tuning, so you tie your hands a bit

Why was binscatter so successful?

- As an intellectual history, binscatter approach is a very recent innovation in applied work
  - Became a staple of much of Raj Chetty and coauthors' work
- Extremely successful as an example of improving our data visualization to communicate results
  - The status quo of big regression tables is *bad*
- Some more general points to improve visual presentation of results

Table 8

Opting Out and the Value of Cash

Panel A

Notes: The dependent variable is the annualized excess return of the firm relative to the Fama and French (1993) 25 size and book-to-market portfolios. Cash includes cash and marketable securities. Many variables are scaled by the market value of equity (ME). Number of Opt Outs is the number of governance categories that firms opt out of. Earnings is earnings before extraordinary items plus interest, deferred taxes, and investment tax credits. Net assets is the value of assets net of cash, and R&D is the value of R&D expenses. Interest expenses include total interest and related expenses. Dividends include common dividends paid, and lagged cash is the lagged value of cash. Debt/Market Value is the ratio of the sum of long-term and short-term debt to the sum of the long-term debt, short-term debt, and the market value of equity. New Finance is the sum of net equity issues and net debt issues. All specifications include fixed effects for the exchange the firm is listed on and the country where the firm is headquartered. Heteroskedasticity-consistent standard errors that correct for clustering at the country level appear in parentheses. ***, **, and * denote significance at the 1, 5, and 10 percent levels, respectively.

Dependent Variable	Annualized Excess Returns				
Countries in Sample:	All	Common Law	Civil Law	High Anti-Self-Dealing	Low Anti-Self-Dealing
	(1)	(2)	(3)	(4)	(5)
Change in Cash / ME	1.6878***	1.8587***	1.6502***	1.8252***	1.7478***
	(0.2122)	(0.3166)	(0.2708)	(0.2989)	(0.3254)
Number of Optouts * Change in Cash / ME	-0.0945	0.0515	-0.1997***	-0.0276	-0.1532*
	(0.0573)	(0.0670)	(0.0412)	(0.0446)	(0.0760)
Number of Optouts	0.0056	0.0145*	-0.0074	0.0062	-0.0034
	(0.0050)	(0.0072)	(0.0086)	(0.0048)	(0.0148)
Change in Earnings / ME	0.4976***	0.4601**	0.4660***	0.5373***	0.3824***
	(0.1175)	(0.1706)	(0.1157)	(0.1525)	(0.1311)
Change in Net Assets / ME	0.1776***	0.2135	0.1651***	0.2546	0.1334***
	(0.0584)	(0.1364)	(0.0510)	(0.1462)	(0.0337)
Change in R\&D / ME	0.2540	0.6163	0.3232	0.4092	0.2201
	(1.0234)	(1.3368)	(0.9671)	(1.2351)	(1.5280)
Change in Interest Expense/ME	-2.2656***	-4.1335***	-1.3077*	-2.4315**	-1.1878
	(0.6087)	(0.5239)	(0.6808)	(1.0161)	(0.8087)
Change in Dividends / ME	2.2446***	2.8959***	1.6995*	3.7704***	0.8538
	(0.6355)	(0.7491)	(0.9179)	(0.5778)	(0.6349)
Lagged Cash / ME	0.1691**	0.2995**	0.0506	0.1828	0.0836
	(0.0776)	(0.1167)	(0.0490)	(0.1091)	(0.0862)
Debt / Market Value	-0.1579***	-0.0722	-0.2092***	-0.1332	-0.2089**
	(0.0418)	(0.0809)	(0.0481)	(0.0776)	(0.0938)
New Finance / ME	-0.1545	-0.1415	-0.1745	-0.3153**	-0.0767
	(0.1058)	(0.1776)	(0.1443)	(0.1343)	(0.1635)
Lagged Cash/ME * Change in Cash Holdings/ME	-0.6712***	-1.1451***	-0.4458**	-1.0258***	-0.3146***
	(0.1415)	(0.2654)	(0.1639)	(0.1659)	(0.1137)
Leverage * Change in Cash Holdings/ME	-0.0021	0.0662	0.0119	0.2040	-0.6240
	(0.2675)	(0.3769)	(0.2809)	(0.3315)	(0.5702)
Country Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Exchange Fixed Effects?	Yes	Yes	Yes	Yes	Yes
No. of Obs.	2370	1180	1190	1203	1072

Some goals

1. Minimize tables
2. Have describable goals for every exhibit
3. Focus the reader and craft not-ugly figures
    - Ideally beautiful, but at minimum not ugly
4. Do not *mislead* your readers

Some goals

1. Minimize tables
2. Have describable goals for every exhibit
3. Focus the reader and craft not-ugly figures
    - Ideally beautiful, but at minimum not ugly
4. Do not *mislead* your readers

Within figures, Schwabish's guidelines are excellent:

1. Show the data
2. Reduce clutter
3. Integrate graphics and text
4. Avoid providing extraneous information
5. Start with grey

1. Minimize Tables

- Tables suck but are important storage units of information.
  - They should be stored in an online appendix
- Tables make it very hard to actually compare results and contrast things
- Tables also tend to report things that are unnecessary
  - The coefficient on the controls are usually not interpretable in a causal way (Hünermund and Louw (2020))
  - Why bother reporting them?
- Even when not doing regressions!

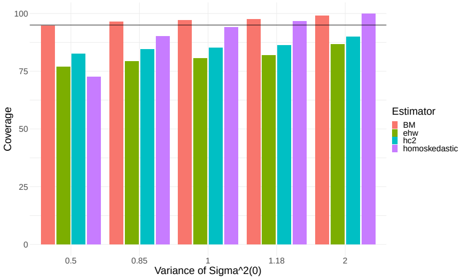
1. Minimize Tables

TABLE 1.—COVERAGE RATES AND NORMALIZED STANDARD ERRORS (IN PARENTHESES) FOR DIFFERENT CONFIDENCE INTERVALS IN THE BEHRENS-FISHER PROBLEM

Angrist-Pischke Unbalanced Design, $N_0 = 27, N_1 = 3$, Normal Errors

		I		II		III		IV		V	
	$\sigma(0)$	0.5		0.85		1		1.18		2	
A. Coverage Rates and Median Standard Errors											
Variance Estimator	Dist/dof	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE	Cov. Rate	Med. SE
$\hat{\mathbb{V}}_{\text{homo}}$	∞	72.5	(0.33)	90.2	(0.52)	94.0	(0.60)	96.7	(0.70)	99.8	(1.17)
	$N-2$	74.5	(0.34)	91.5	(0.54)	95.0	(0.63)	97.4	(0.73)	99.8	(1.22)
$\hat{\mathbb{V}}_{\text{EHW}}$	∞	76.8	(0.40)	79.3	(0.42)	80.5	(0.44)	81.8	(0.45)	86.6	(0.55)
	$N-2$	78.3	(0.42)	80.9	(0.44)	82.0	(0.46)	83.3	(0.47)	88.1	(0.57)
	wild	89.6	(0.73)	89.4	(0.70)	89.6	(0.69)	89.9	(0.68)	91.8	(0.69)
	wild_0	89.7	(0.55)	97.5	(0.75)	98.7	(0.85)	99.5	(0.99)	99.9	(1.64)
$\hat{\mathbb{V}}_{\text{HC2}}$	∞	82.5	(0.49)	84.4	(0.51)	85.2	(0.52)	86.2	(0.53)	89.8	(0.62)
	$N-2$	83.8	(0.51)	85.6	(0.53)	86.5	(0.54)	87.4	(0.56)	91.0	(0.65)
	wild	90.3	(0.76)	90.3	(0.74)	90.5	(0.73)	90.8	(0.72)	92.4	(0.73)
	wild_0	89.8	(0.55)	97.5	(0.75)	98.7	(0.85)	99.4	(0.99)	99.9	(1.64)
	K^*_{Welch}	96.1	(1.02)	96.8	(0.98)	97.0	(0.95)	97.1	(0.93)	96.7	(0.87)
	K_{Welch}	93.1	(1.00)	92.5	(0.93)	92.4	(0.90)	92.5	(0.87)	93.5	(0.80)
	K_{BM}	94.7	(0.90)	96.4	(0.94)	97.0	(0.95)	97.6	(0.98)	99.1	(1.14)
$\hat{\mathbb{V}}_{\text{HC3}}$	∞	87.2	(0.60)	88.6	(0.61)	89.2	(0.62)	89.9	(0.63)	92.4	(0.71)
	$N-2$	88.2	(0.62)	89.5	(0.64)	90.1	(0.65)	90.8	(0.66)	93.4	(0.74)
$\max_{\text{EHW}}$	∞	82.2	(0.41)	91.8	(0.54)	94.7	(0.62)	97.0	(0.71)	99.8	(1.17)
$\max_{\text{HC2}}$	∞	86.1	(0.49)	93.2	(0.57)	95.4	(0.64)	97.3	(0.73)	99.8	(1.17)
B. Mean Effective dof											
	K^*_{Welch}	2.1		2.3		2.5		2.7		4.1	
	K_{Welch}	2.8		3.8		4.4		5.1		8.6	
	K_{BM}	2.5		2.5		2.5		2.5		2.5	

Cov. Rate refers to coverage of nominal 95% confidence intervals (in percentages), and "Med. SE" refers to standard errors normalized by $t^K_{0.975}/t^{\infty}_{0.975}$. Variance estimators and dof adjustments are described in the text, and wild bootstrap confidence intervals ("wild" and "wild_0") are described in section 2 in the appendix; $\max_{\text{EHW}} = \max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{EHW}})$, and $\max_{\text{HC2}} = \max(\hat{\mathbb{V}}_{\text{homo}}, \hat{\mathbb{V}}_{\text{HC2}})$. Results are based on 1 million replications, except for wild bootstrap-based confidence intervals, which use 100,000 replications and 1,000 bootstrap draws in each replication.



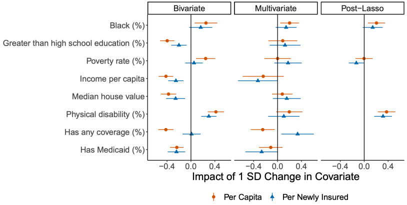
1. Minimize Tables

Appendix Table A4: Correlates with reduction in collections debt at age 65

		Bivariate		Multivariate		Post-Lasso	
Covariate	Estimate Type	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Black (%)	Per Capita	-7.17	(2.77)	-5.74	(2.28)	-6.23	(2.08)
Greater than high school education (%)	Per Capita	11.30	(1.74)	-2.47	(3.52)	4.86	(2.46)
Has any coverage (%)	Per Capita	12.00	(1.86)	7.09	(2.94)		
Has Medicaid (%)	Per Capita	6.75	(1.65)	3.24	(2.87)		
Hospital beds per capita	Per Capita	-1.09	(1.4)	1.86	(1.48)		
Income per capita	Per Capita	11.90	(1.79)	6.86	(5.01)		
Median house value	Per Capita	10.70	(1.88)	-2.25	(2.49)		
Hospital occupancy rate (%)	Per Capita	6.56	(1.68)	-0.90	(3.12)		
Physical disability (%)	Per Capita	-11.90	(2)	-5.60	(3.21)	-7.41	(2.56)
Poverty rate (%)	Per Capita	-7.01	(2.34)	-0.01	(3.24)	1.02	(2.16)
Payment by charity care patients ($)	Per Capita	-1.52	(1.65)	-1.78	(1.46)	-2.70	(1.53)
Medicare spending per enrollee ($)	Per Capita	-6.48	(2.08)	-0.63	(2.98)		
For-profit hospitals (%)	Per Capita	-10.20	(1.96)	-4.96	(2.17)	-8.29	(1.97)
Teaching hospitals (%)	Per Capita	9.69	(1.51)	6.14	(3.32)		
Cost of charity care per patient day ($)	Per Capita	0.07	(3.1)	-0.96	(2)	-1.26	(2.21)

Figure 3: Commuting zone characteristics correlated with the reduction in collections debt at age 65

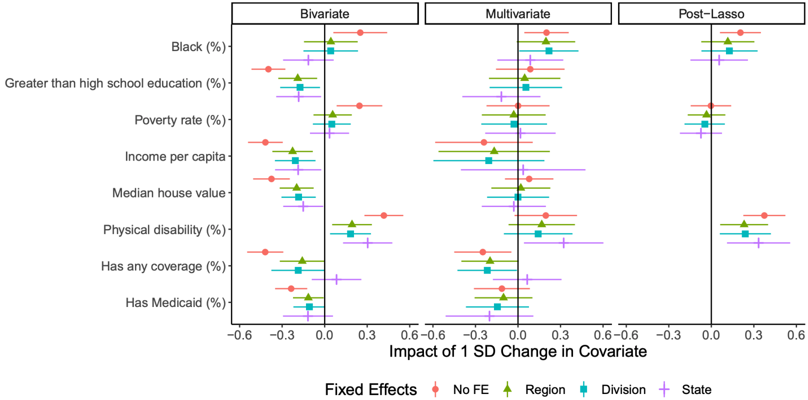
Panel A: Demographic characteristics



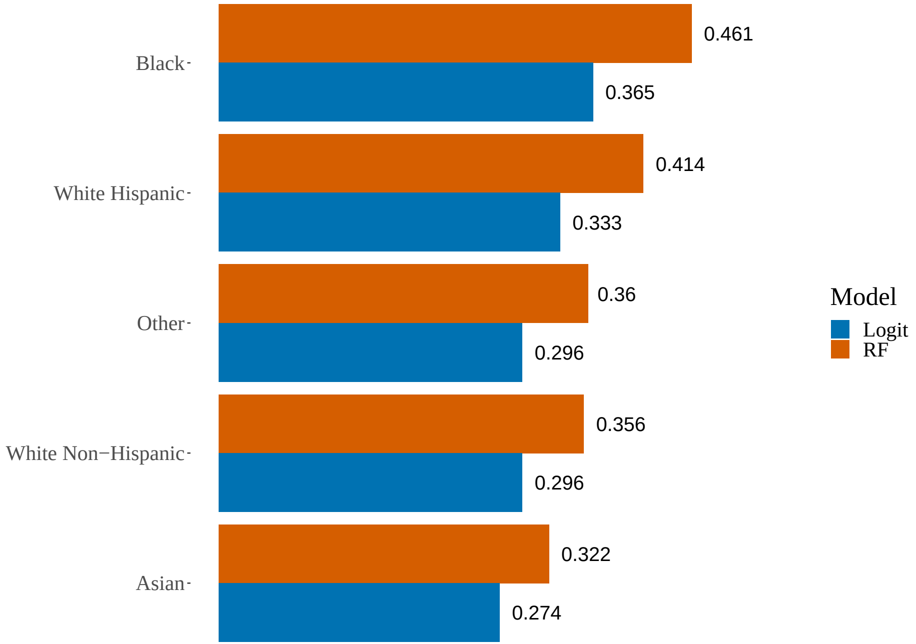
1. Minimize Tables

Appendix Figure A12: Correlates with reduction in collections debt at age 65, with Fixed Effects

Panel A: Area-level demographic characteristics



1. Minimize Tables

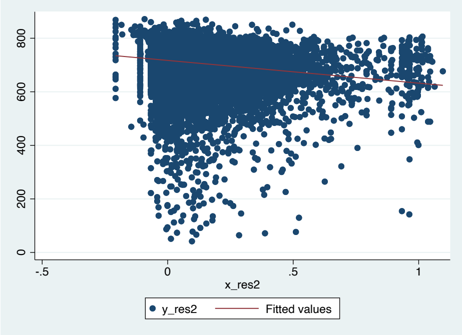


2. Describable Goals

- When considering a figure, for most papers you want the result to be obvious
  - Research papers' exhibits typically are not "exploratory"
- If it is not immediately obvious what the goal of an exhibit is, one of two things are likely occuring
  - You have too much information, and the story you are telling is lost
  - You have too little information or highlighting of the relevant piece that you're interested in
- Jon Schwabish describes this as "preattentive processing" – how do we emphasize certain pieces of a figure for the reader?

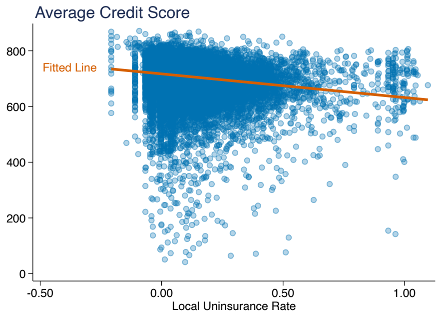
3. Craft not-ugly figures

- There is huge variation in how much researchers value figures
- Nonetheless, there's almost no good reason to have *bad* figures
- Avoiding this entails a small amount of work for big returns. For this example, we could:
  1. Fix the scheme (e.g. blue on white is ugly)
  2. Label our axes
  3. Make our color scheme clearer
  4. Thicken the line fit, and lighten the points
- Keep in mind that
  1. 1 in 12 men is color blind
  2. Many print in black and white



3. Craft not-ugly figures

- There is huge variation in how much researchers value figures
- Nonetheless, there's almost no good reason to have *bad* figures
- Avoiding this entails a small amount of work for big returns. For this example, we could:
  1. Fix the scheme (e.g. blue on white is ugly)
  2. Label our axes
  3. Make our color scheme clearer
  4. Thicken the line fit, and lighten the points
- Keep in mind that
  1. 1 in 12 men is color blind
  2. Many print in black and white

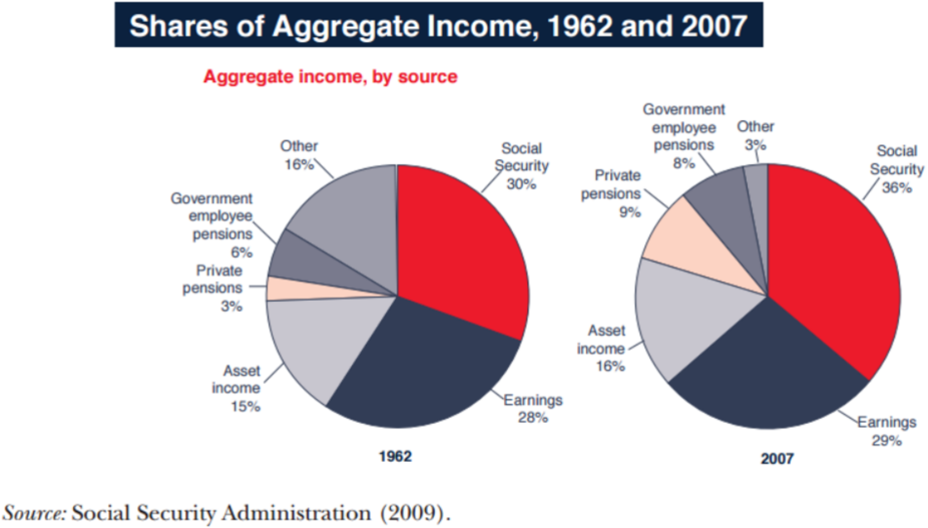


Making good figures is hard

Some suggestions:

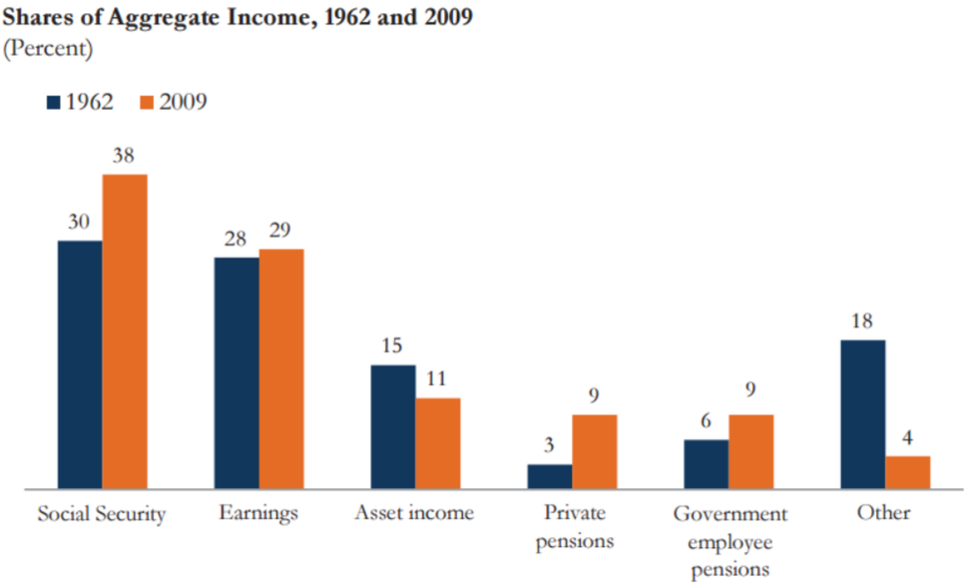
- Bar graphs are always good places to start. Make them horizontal (almost always) so that your labels are readable.
- Directly label on your figure as much as you can – it makes it much easier for the reader to pay attention to what is going on
- Fix your units
  - Round numbers, add commas, put dollar signs...
- Label your axes, but label your y-axis at the top of your graph rather than turned 90 degrees on the side
- Use shapes, thickness, saturation, color, size, markings, position, sharpness... to highlight things in your graphs

Some Examples



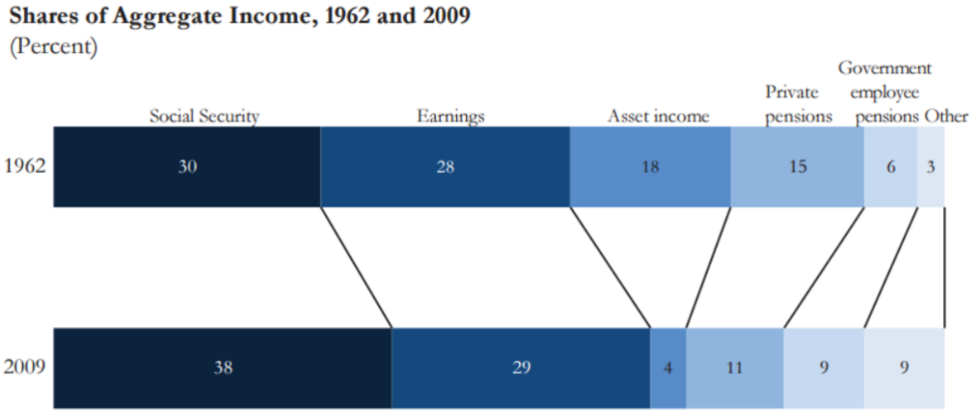
Source: Schwabish (2014)

Some Examples



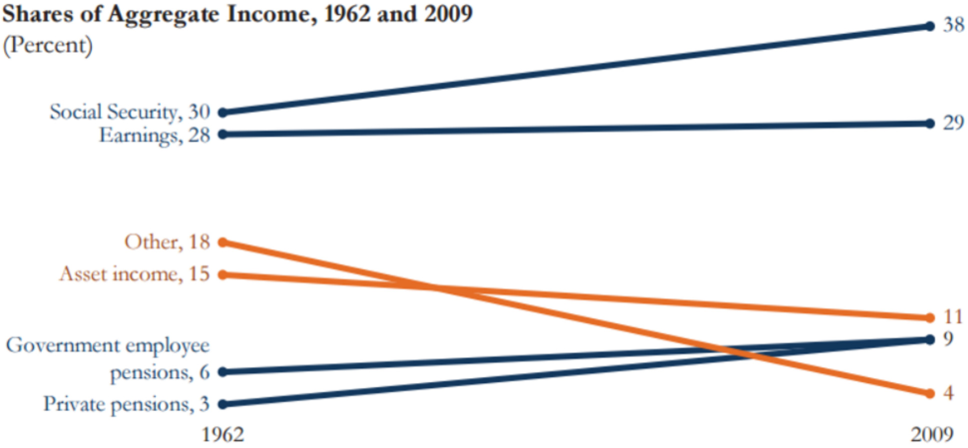
Source: Schwabish (2014)

Some Examples



Source: Schwabish (2014)

Some Examples



Source: Schwabish (2014)

Further Reading

- Denny, Kevin (2021): Basic Stata Graphics for Social Science Students. UCD Geary Insitute for Public Policy Discussion Paper Series, 2021/02.
- Schwabish, Jonathan (2021): Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks.
- Schwabish, Jonathan (2014): An Economist's Guide to Visualizing Data. Journal of Economic Perspectives, 28 (1): 209-34.
- Healy, Kieran (2019): Data Visualization: A Practical Introduction.

Agenda

1. Empirical Papers
2. Tables
3. Visualization
4. Theoretical Papers

Structure of Theory Papers

Most theory papers have a structure very similar to the following:

1. Introduction
2. Model Setup
    - First general setup (describe the world in the model)
    - Then individual parts (consumers, firms...)
3. Equilibrium/Optimal tax structure/...
4. Extension to xy
5. Discussion
6. Conclusion

How to Write Theory

- Motivate your assumptions.
- Answer the following early on (in this order)
  1. Who are the players?
  2. What can they do?
  3. What do they know?
  4. What are their objectives?
- Don't vary assumptions in the main text
  - Say in a footnote what happens without it.
  - Or in the discussion section.
- Provide intuition for each result (and most equations).

Formal Things

- Define all variables. Follow the literature in your choice of variables.
- Number the equations (maybe not all).
- Show your results
  - Lemmata: intermediate results, “stepping stones”
  - Propositions: Main results (2-4), need a proof, but also explain intuition in text
  - Theorems: Very important propositions (rare)
  - Corollaries: Side statement that follow from a proposition/theorem, need a proof
  - Conjectures: no proof
- The differences between them are often subjective.

Further Reading

- Glaeser, Ed (2003): How to write a theory paper,
  https://www.princeton.edu/~reddings/tradephd/Glaeser_Lecture_11.pdf
- Varian, Hal (1989): What Use is Economic Theory?,
  http://people.ischool.berkeley.edu/~hal/Papers/theory.pdf