Linear Regression II: Semiparametrics + Visualization

Paul Goldsmith-Pinkham

February 2, 2023

Linear Regression: Why so Popular?

- Linear regression is incredibly popular as a tool. Why?

- Many reasons:
  - Fast (easy analytic solution and matrix inversion has gotten better)
  - Efficient (under some settings, OLS is BLUE)

- My view: linear regressions is
  1. an intuitive summary of data relationships
  2. A good default – many “better” options are only good in some settings, and linear regression is not bad in many
  3. Does a good job with many of the things we throw at our models (high dimensional fixed effects, lots of data)

- Today: how to stay in the world of linear regression as much as possible, improving our presentation
  - As a side goal, we will do a discussion on good visualization practice

General framework of causal relationships

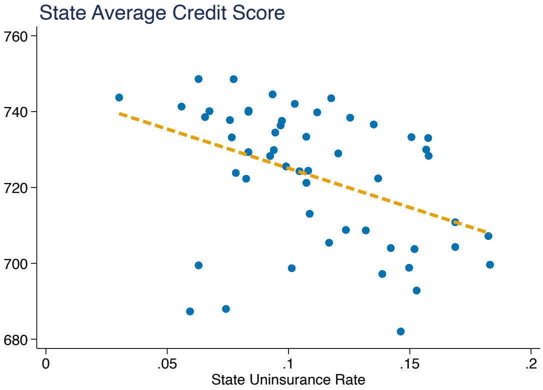
- Without any structure, we can describe our usual relationships as $Y_i = F(D_i, W_i, \epsilon_i)$
  - D_i is some causal variable we care about
  - W_i is controls / heterogeneity
  - ϵ_i is unobservable noise
  - Very unrestricted!

- This function is very challenging to estimate with non-seperable ϵ_i and if the dimension of D_i or W_i is high
  - Simpler: $Y_i = F(D_i, W_i) + \epsilon_i$
  - What do we report from this? $E\left(\frac{\partial F}{\partial D_i} \middle| W_i = w\right)$? $E\left(\frac{\partial F}{\partial D_i}\right)$?

- What does a simple linear model get us to? $Y_i = D_i\tau + W_i\beta + \epsilon_i$
  - Can be more complex! E.g. $Y_i = D_i\tau + W_i\beta_1 + D_i \times W_i\beta_2 + \epsilon_i$, etc.
  - However, in this setting there is not a “single” number either

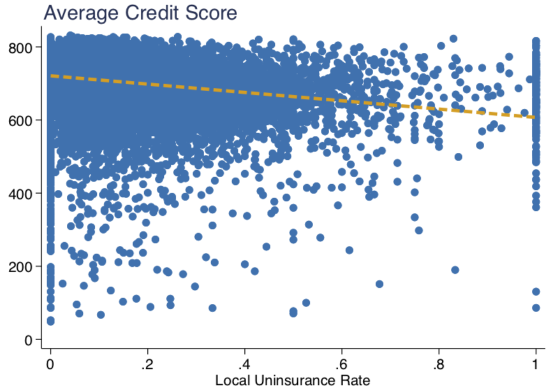
Visualizing a relationship

- Intuitively, for many papers, we plot an outcome Y_i and want to describe/assert a relationship/effect from 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/assert a relationship/effect from 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



Visualizing a *multivariate* relationship

- What about controls? E.g. we have a causal estimand conditional on a set of covariates W
- First, an aside. Let W be discrete – e.g., we think the effect of D is causal, but only conditional on fixed effects.
  - How can we think about the OLS regression?
- In the pscore setting, we would estimate $\tau(w) = E(Y|D_i = 1, W = w) - E(Y|D_i = 0, W = w)$, and then aggregate this using the distribution of the w (using IPW)
  - With OLS, this is done for us automatically. How?
- Recall in a regression, our setup is

$$Y_i = \tau D_i + \beta W_i + \epsilon_i$$

Residual Regression

$$Y_i = \tau D_i + \beta W_i + \epsilon_i$$

- Consider the projection of D_i and Y_i onto W_i
    - Note that if W and D are uncorrelated, we don't have to worry about controlling for it.
- We define a projection matrix as $\mathbf{P}_W = \mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n$
    - Note that $\mathbf{P}_W\mathbf{W}_n = \mathbf{W}_n, \mathbf{P}_W\mathbf{P}_W = \mathbf{P}_W$
    - Also note that $\mathbf{P}_W\mathbf{D}_n$ gives you the predicted values from a linear regression:

$$D_i = \gamma W_i + u_i$$

- Finally, denote $\mathbf{M}_W = \mathbf{I}_n - \mathbf{P}_W$ as the annhilator matrix
    - This gives us the residual from the regression on W_i! (e.g. u_i above).

Frisch-Waugh-Lovell? More like Frisch-Wow-Lovell!

$$Y_i = \tau D_i + \beta W_i + \epsilon_i$$

- Now if we transform $\mathbf{Y}_n^* = M_W \mathbf{Y}_n$ and $\mathbf{D}_n^* = M_W \mathbf{D}_n$, we can run

  $$Y_i^* = \tau D_i^* + \tilde{\epsilon}_i$$

  and get the right coefficient τ! (This is the Frisch-Waugh-Lovell theorem)

- Consider W as a discrete set of covariates. This will demean D and Y within each group. It is not too difficult to show that this regression estimate will get you

$$\tau = \frac{E(\sigma_D^2(W_i)\tau(W_i))}{E(\sigma_D^2(W_i))}, \qquad \sigma_D^2(W_i) = E((D_i - E(D_i|W_i))^2|W_i) \tag{1}$$

  Let's derive this, and show how it can fail more generally.

- To build intuition, consider both W_i and D_i binary. Then add another treatment arm.
- Consider regression
$$Y_i = \alpha + D_i\beta + W_i\gamma + U_i,$$
with $D_i, W_i \in \{0, 1\}$. By definition, U_i mean-zero regression residual uncorrelated with (D_i, W_i)
- Stylized Project STAR example: D_i is small classroom dummy, Y_i is avg test score of student i
    - Randomization stratified: probability of assignment to small vs large classroom depends on school. W_i denotes school FE
    - Binary W_i: only 2 schools for simplicity

Potential outcomes and key assumption

- To characterize β, use potential outcomes notation $Y_i(d)$
  - Individual treatment effect $\tau_{i1} = Y_i(1) - Y_i(0)$, conditional treatment effect $\tau_1(w) = E[\tau_{i1} \mid W_i = w]$
  - Observed outcome $Y_i = Y_i(0) + \tau_{i1} D_i$
  - Propensity score: $p_1(W_i) = \Pr(D_i = 1 \mid W_i) = E[D_i \mid W_i]$
- Treatment (as good as) randomly assigned conditional on W_i: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid W_i$
- Random assignment assumption delivers key result from Angrist (1998):

$$\beta = \phi\tau_1(0) + (1-\phi)\tau_1(1), \quad \phi = \frac{\operatorname{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1} \operatorname{var}(D_i \mid W_i = w)\Pr(W_i = w)},$$

Derivation

$$\beta \overset{(1)}{=} \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{EE[\tilde{D}_i Y_i(0) \mid W_i]}{E[\tilde{D}_i^2]} + \frac{EE[\tilde{D}_i D_i \tau_{i1} \mid W_i]}{E[\tilde{D}_i^2]}$$

$$\overset{(2)}{=} \frac{E[\operatorname{var}(D_i \mid W_i)\tau(W_i)]}{E[\operatorname{var}(D_i \mid W_i)]}$$

$$= \phi\tau(0) + (1-\phi)\tau(1) \quad \phi = \frac{\operatorname{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1} \operatorname{var}(D_i \mid W_i = w)\Pr(W_i = w)},$$

- (1) follows from FWL theorem; $\tilde{D}_i$ residual from regressing D_i on W_i.
- (2) follows by random assignment, and the fact that $E[\tilde{D}_i \mid W_i] = 0$ (not just $\operatorname{corr}(\tilde{D}_i, W_i) = 0$).

Key features of this estimator

$$\beta = \phi\tau(0) + (1-\phi)\tau(1), \quad \phi = \frac{\text{var}(D_i \mid W_i = 0)\Pr(W_i = 0)}{\sum_{w=0}^{1} \text{var}(D_i \mid W_i = w)\Pr(W_i = w)},$$

- $\phi \in (0, 1)$
- No need to estimate propensity score
- Puts larger weight on strata with higher variation in D_i
  - $\neq$ ATE! (unless $\tau(w)$ constant or $p_1(w)$ constant across strata)
  - May lead to unusual or "unrepresentative" estimand (Aronow and Samii (2016)
  - But this sort of weighting necessary to avoid loss of identification under overlap failure (e.g. $p_1(0) = 0$), or lack of precision under weak overlap ($p_1(0)$ close to 0)

Multiple treatments

- Project STAR in fact had additional treatment arm in addition to small class ($D_i = 1$): full-time teaching aide ($D_i = 2$).

$$Y_i = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2 + W_i\gamma + U_i,$$

- General notation:
    - $X_i = [X_{i1}, X_{i2}]'$, $X_{ij} = \mathbb{1}\{D_i = j\}$
    - $Y_i = Y_i(0) + X_i'\tau_i$, where $\tau_{ik} = Y_k(k) - Y_i(0)$.
    - Let $\tau_k(W_i) = E[\tau_{ik} \mid W_i]$ and $p_{ok}(w) = E[X_{ik} \mid W_i = w]$.
- Assignment still conditionally random, $(Y_i(0), Y_i(1), Y_i(2)) \perp X_i \mid W_i$

Causal interpretation of β_1

Again, due to FWL,

$$\beta_1 = \frac{E[\tilde{\tilde{X}}_{i1} Y_i]}{E[\tilde{\tilde{X}}_{i1}^2]} = \frac{E[\tilde{\tilde{X}}_{i1} Y_i(0)]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i1} \tau_{i1}]}{E[\tilde{\tilde{X}}_{i1}^2]} + \frac{E[\tilde{\tilde{X}}_{i1} X_{i2} \tau_{i2}]}{E[\tilde{\tilde{X}}_{i1}^2]}$$
$$= E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)],$$

where $\lambda_{11}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1} X_{i1} | W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \geq 0$, and $\lambda_{12}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1} X_{i2} | W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \neq 0$ in general.

Key point $\tilde{\tilde{X}}_{i1}$ is residual from regressing X_{i1} on W_i, constant, and X_{i2}

- $\tilde{\tilde{X}}_{i1} \neq X_{i1} - E[X_{i1} \mid W_i, X_{i2}]$, since X_{i2} depends non-linearly on X_{i1}
- As a result, β_1 contaminated by τ_{i2}.

Stylized Example: No overlap

- Suppose only units in stratum $W_i = 0$ receive treatment 2. Let $n_k(w) = \sum_{i=1}^{N} \mathbb{1}\{W_i = w, X_i = k\}$.
- Then

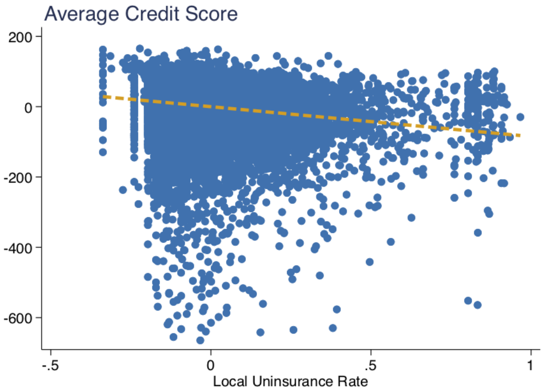
$$\hat{\beta} = \begin{pmatrix} \phi\hat{\tau}_1(0) + (1-\phi)\hat{\tau}_1(1) \\ \frac{n_1(0)(1-\phi)}{n_1(0)+n_0(0)}\left[\hat{\tau}_1(1) - \hat{\tau}_1(0)\right] + \hat{\tau}_2(0) \end{pmatrix},$$

where $\phi = \frac{(1/n_1(0)+1/n_0(0))^{-1}}{\sum_{w=0}^{1}(1/n_1(w)+1/n_0(w))^{-1}}$.
- E.g., with equal-sized strata, $n_0(0) = n_1(0) = n_2(0)$, and $n_0(1) = n_1(1)$,

$$\hat{\beta} = \begin{pmatrix} \frac{2}{5}\hat{\tau}_1(0) + \frac{3}{5}\hat{\tau}_1(1) \\ \frac{3}{10}\left[\hat{\tau}_1(1) - \hat{\tau}_1(0)\right] + \hat{\tau}_2(0) \end{pmatrix}.$$

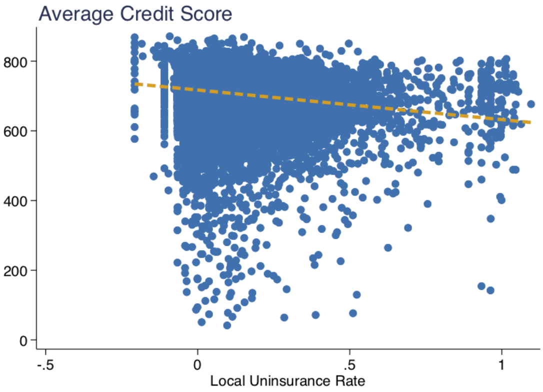
Exploiting FWL for visualization

- Key point: we can still plot our line, but it would be nice to lay the line over data
- Why don't we exploit FWL and plot Y^* and D^*?
  - Add in state fixed effects
- Kind of hard to intuit b/c demeaned



Exploiting FWL for visualization

- Key point: we can still plot our line, but it would be nice to lay the line over data
- Why don't we exploit FWL and plot Y^* and D^*?
  - Add in state fixed effects
- Kind of hard to intuit b/c demeaned
- Easy solution – add back the overall means
  - Can you see an issue here?



Can we do more?

- Residual regression is powerful
- Maybe we could use it to do something more flexible? When I plot my data, it's not totally obvious that a straight line is the best fit. But it's hard to see because there's so much data.
- Recall that we're acutally interested in conditional expectation functions – e.g. $E(Y|D)$
  - What's a way to approximate this?

An aside on non-parametric vs. semiparametric vs. parametric

- What I view as the formal definition:
  - Parametric: model where data generating process is specified as finite dimensional. Hence,
    $$Y_i = D_i\beta + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
    is a fully parametric model (conditional on D)
  - Non-parametric: model where the data generating process is specified as infinite dimensional. E.g.
    $$Y_i = F(D_i, \theta_i)$$
    where θ_i is infinite-dimensional parameter
  - Semi-parametric: a combination. E.g. even OLS with robust standard errors:
    $$Y_i = D_i\beta + \epsilon_i, \qquad \epsilon_i \sim F(\theta_i),$$
    where θ_i is infinite dimensional and β is finite dimensional

- Important to distinguish between *nuisance* parameters (e.g. we don't care about actually estimating θ_i in the robust standard error example) and parameters of interest.

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
  - Idea: points reflect $f(D_i)$

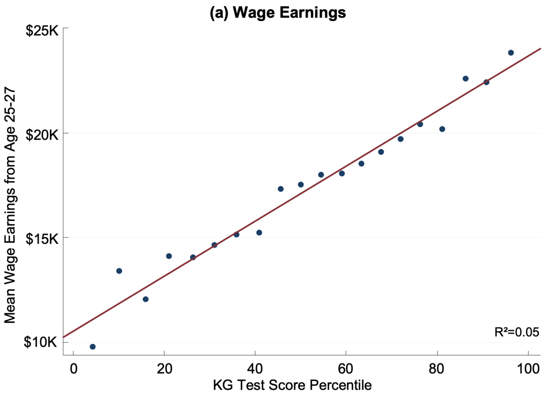
Chetty et al. (2011) - Kindergarten scores on adult earnings

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
  - Non-parametrics is (easier) when discrete!

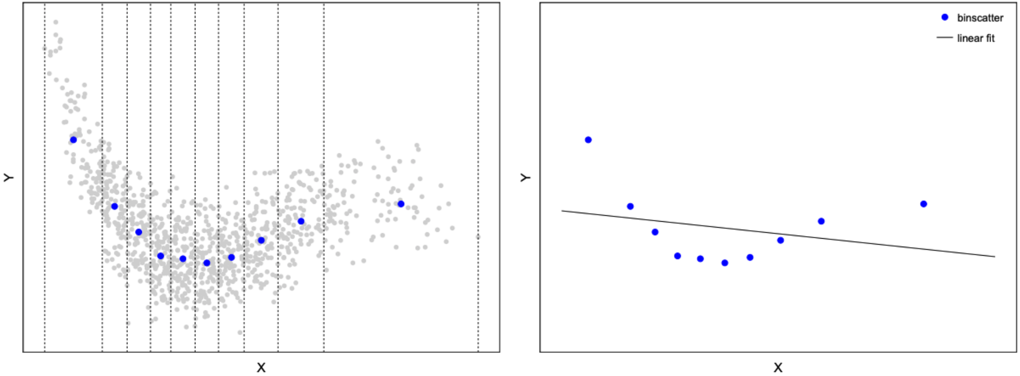
- So what's going on under the hood?



Chetty et al. (2011) - Kindergarten scores on adult earnings

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

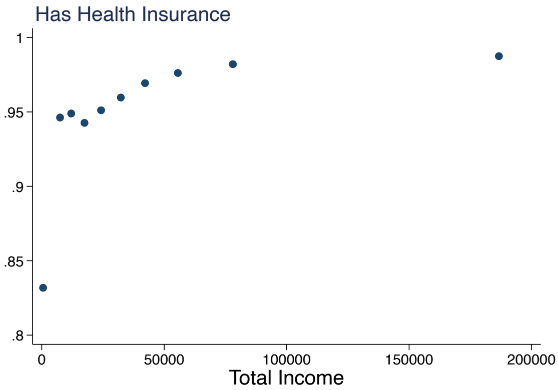
Figure 1: The basic construction of a binned scatter plot.



(a) Scatter and Binscatter Plots (b) Binscatter and Linear Fit

Start with binscatter

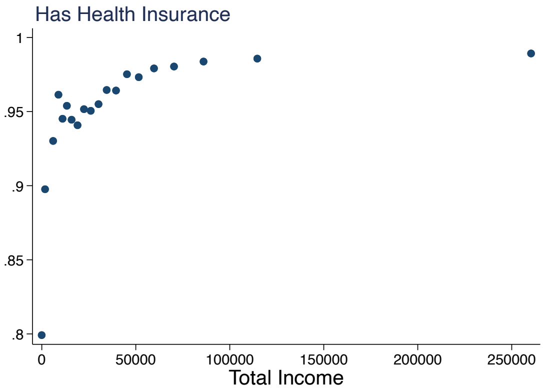
- Choice of bin is not obvious
- How you pick bins can influence interpretation



income on health insurance, 10 bins

Start with binscatter

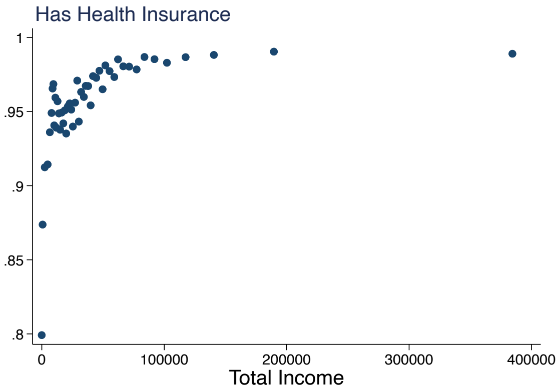
- Choice of bin is not obvious
- How you pick bins can influence interpretation



income on health insurance, 20 bins

Start with binscatter

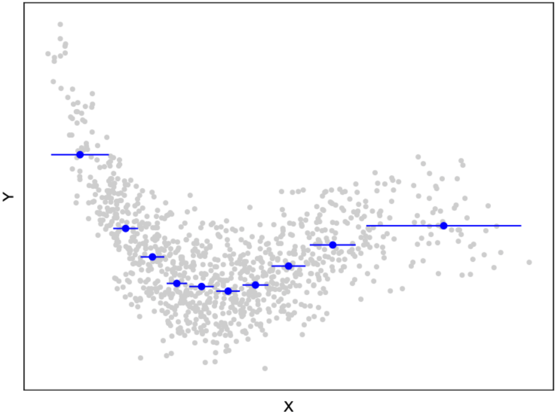
- Choice of bin is not obvious
- How you pick bins can influence interpretation
- This is a statistical problem!



income on health insurance, 50 bins

Cattaneo et al. “On Binscatter”

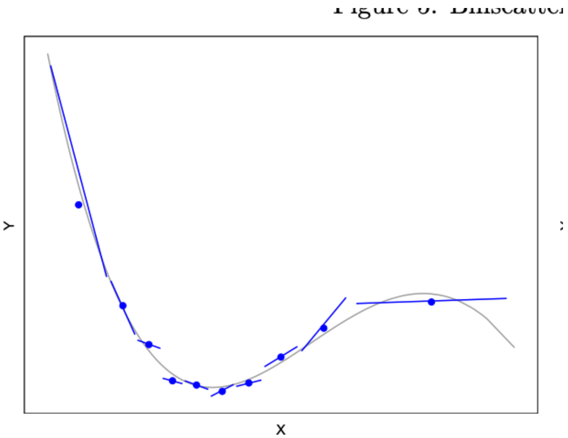
- Paper provides several generalizations to binscatter approach
- First contribution: highlight that the “traditional” binscatter approach is presenting a particular non-parametric estimation
- Initially assumes that constant within bin
  - Not crazy! But could do more.
- Piece-wise functions can be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



(a) Binned Scatter Plot with Piecewise Constant Fit

Cattaneo et al. “On Binscatter”

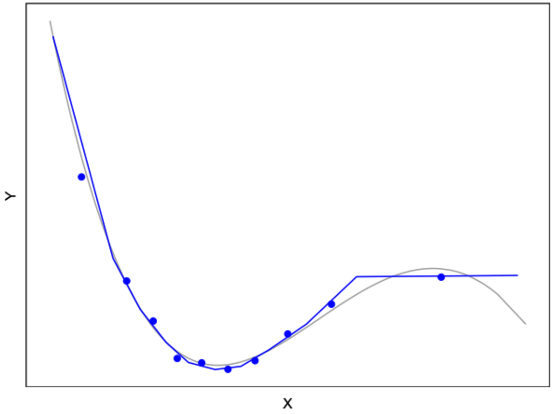
- Paper provides several generalizations to binscatter approach
- First contribution: highlight that the “traditional” binscatter approach is presenting a particular non-parametric estimation
- Initially assumes that constant within bin
    - Not crazy! But could do more.
- Piece-wise functions can be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

Cattaneo et al. "On Binscatter"

- Third contribution: back to residual regression

- Recall our approach was to residualize D_i by our controls to do residual regression
  - Exploiting Frisch-Waugh-Lovell theorem

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

- In this setting, you can't residual D_i and get back the function f if f is non-linear
  - Unfortunately, this is what historically has been the default in Stata package

- Correct way to view this – imagine binning D_i and running the regression. You want to plot the coefficients

Cattaneo et al. “On Binscatter”

- Third contribution: back to residual regression

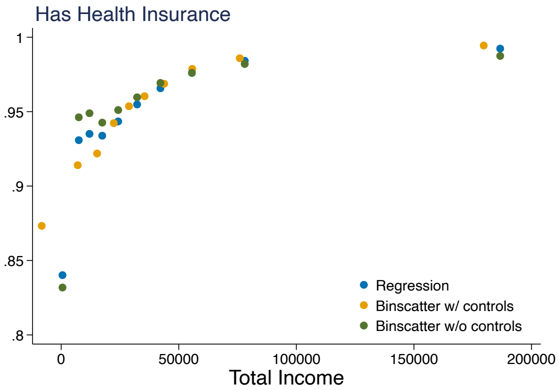
- Recall our approach was to residualize D_i by our controls to do residual regression
  - Exploiting Frisch-Waugh-Lovell theorem

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

- In this setting, you can’t residual D_i and get back the function f if f is non-linear
  - Unfortunately, this is what historically has been the default in Stata package

- Correct way to view this – imagine binning D_i and running the regression. You want to plot the coefficients

Comparison of methods:

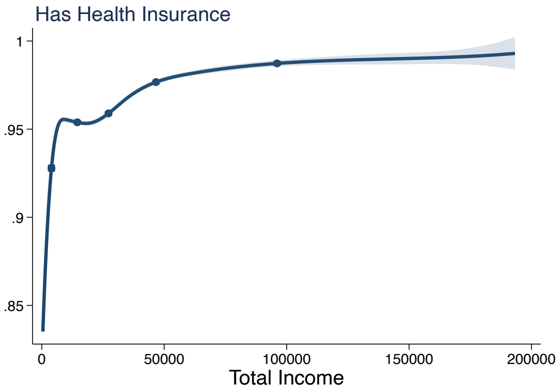


Controls: age, sex, and state of residence

Cattaneo et al. “On Binscatter”

- Final contribiution: testing the CEF
- By defining the estimand, we can actually test properties of it
  - Confidence intervals
  - Test monotonicity
- We actually see a noticeable dip across income – maybe driven by Medicaid eligibility thresholds?
- Code for this is all available here: `https://nppackages.github.io/binsreg/`
- If you just want to fix the FWL issue: `https://github.com/mdroste/stata-binscatter2`

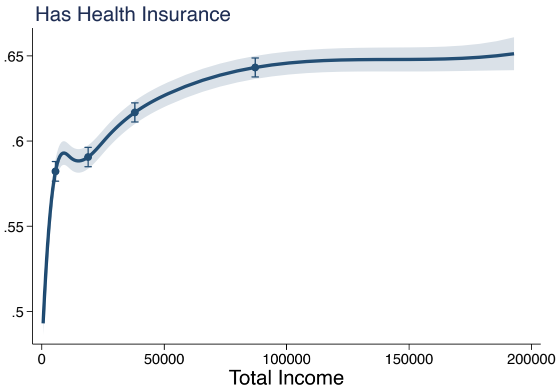
Comparison of methods:



Cattaneo et al. “On Binscatter”

- Final contribiution: testing the CEF
- By defining the estimand, we can actually test properties of it
  - Confidence intervals
  - Test monotonicity
- We actually see a noticeable dip across income – maybe driven by Medicaid eligibility thresholds?
- Code for this is all available here: `https://nppackages.github.io/binsreg/`
- If you just want to fix the FWL issue: `https://github.com/mdroste/stata-binscatter2`

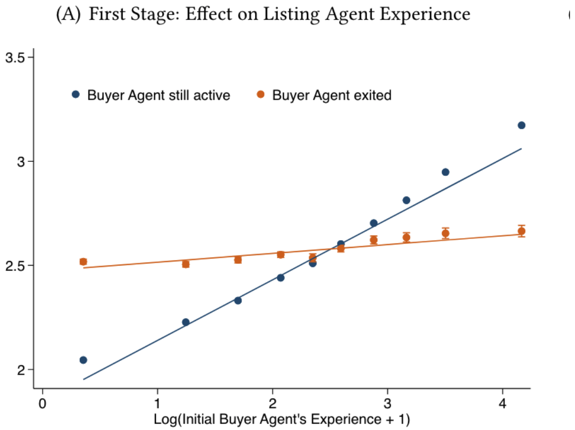
Comparison of methods:



Controls: age, sex, and state of residence (Note, level is off b/c program currently does not recenter correctly with covariates)

Binscatter

- Key point: Binscatter is super useful, but needs to be done correctly
  - Do not mess up the Frisch-Waugh-Lovell point
- Taking serious the estimand adds a lot of tools into your toolset!
- But, a lot of times these approaches are buttressing a simple reported linear number
  - Nuance is important, but a paper has many pieces – useful to have summary numbers



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 coauthor's work
- Extremely successful as an example of improving our data visualization to communicate results
    - The status quo of big regression tables is *bad*
- Will finish by discussing ways to improve visual design and improving communication in papers

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

My design 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

My design 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 necessary to generate strong ignorability are not interpretable in a causal way (Hunermund and Louw (2020))
  - Why bother reporting them?
- Even when not doing regressions!

1. Minimize Tables

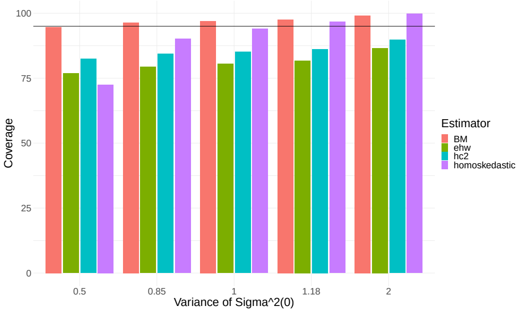
- Several examples of tables vs. regression improvements
- Imbens and Kolesar siulations

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

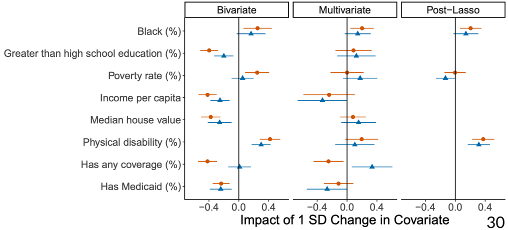
- Several examples of tables vs. regression improvements
- Imbens and Kolesar siulations
- Regression output!

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

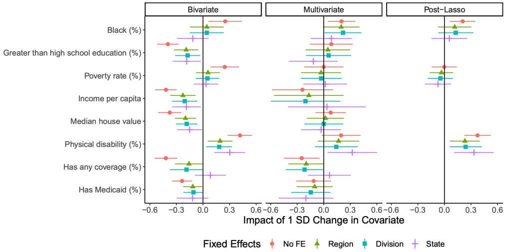


1. Minimize Tables

- Several examples of tables vs. regression improvements
- Imbens and Kolesar siulations
- Regression output!
- Can compress a lot of information

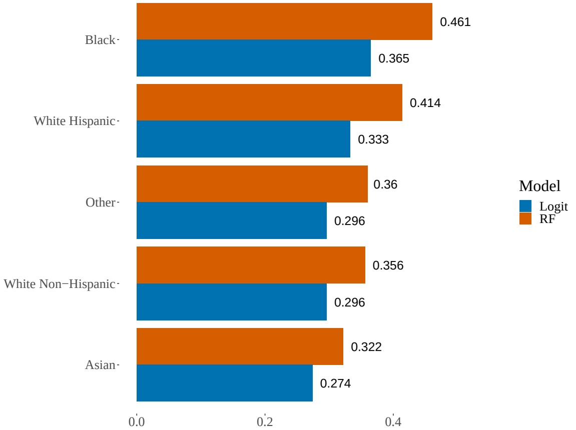
Appendix Figure A12: Correlates with reduction in collections debt at age 65, with Fixed Effects

Panel A: Area-level demographic characteristics



1. Minimize Tables

- Several examples of tables vs. regression improvements
- Imbens and Kolesar siulations
- Regression output!
- Can compress a lot of information
- Also can use it for model output (this is really effective in presentations)



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

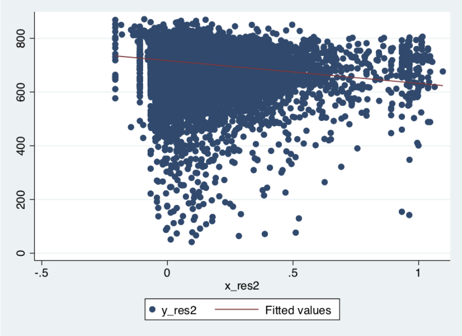
- Jon Schwabish describes this as "preattentive processing" – how do we emphasize certain pieces of a figure for the reader?

3. Craft not-ugly figures

- There is huge variation in how much researchers value figures
    - I'm quite aware I fall on an extreme of that distribution

- 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

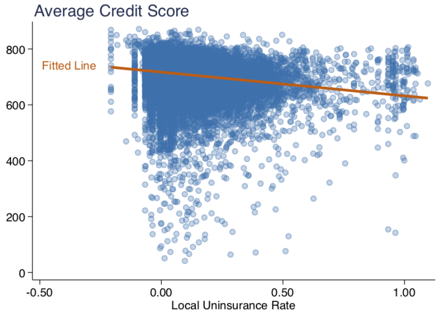


3. Craft not-ugly figures

- There is huge variation in how much researchers value figures
    - I'm quite aware I fall on an extreme of that distribution

- 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

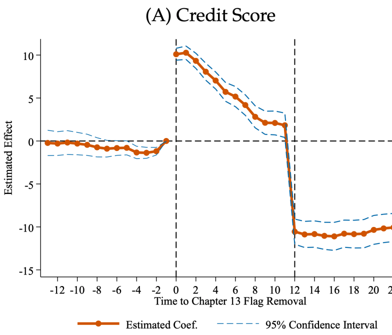


4. Do not mislead your readers

- Readers will percieve things in certain ways, and you can exploit that
  - For good or for evil! Pick good.

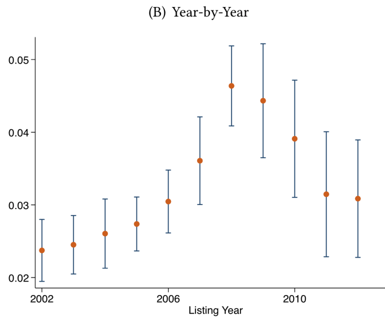
- Consider the following example (from my own work which I have since changed)
  - In many event study settings, we plot the dynamic coefficients
  - We typically have period by period data – don't want to imply smoothness that isn't there
  - My (updated) view: better to use pointwise caps, as the smooth lines imply something that is not true

- Also important – keep improving your graphs! All graphs can be improved, but you don't have to improve every graph.



4. Do not mislead your readers

- Readers will percieve things in certain ways, and you can exploit that
  - For good or for evil! Pick good.
- Consider the following example (from my own work which I have since changed)
  - In many event study settings, we plot the dynamic coefficients
  - We typically have period by period data – don't want to imply smoothness that isn't there
  - My (updated) view: better to use pointwise caps, as the smooth lines imply something that is not true
- Also important – keep improving your graphs! All graphs can be improved, but you don't have to improve every graph.



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.
- Don't put confidence intervals on bar graphs. Use a point range plot instead
- 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, put zero padding
- Label your axes, but label your y-axis at the top of your graph rather than turned 90 degrees on the side
- Use gestalt principles to highlight things in your graphs:
  - Shapes, thickness, saturation, color, size, markings, position, sharpness

Making good figures is hard

- We are not the NYTimes – we do not need to make insanely polished visualizations
- Most of our results will be relatively simple, but we will have a lot of versions of it that we need to convey
  - Key: provide a polished way to provide a bite-sized piece of information
  - Then, once the reader understands that, a large host of other information is also easily processed
  - E.g., consider these figures from my paper
- A lot going on, but in given panel, can break down into bite sized pieces
  - Each subsequent result is then easily understood

Figure 1: Changes in health insurance, financial health, and covariates at age 65