Sharp Econometrics

Chris Sims

June 3, 2023

What are we talking about?

- "Sharp" is meant to contrast with "Mostly Harmless", from the title of a well known econometrics text book Angrist and steffen Pischke (2009).
- The new approaches, including the "Mostly Harmless" book, are sometimes called the "credibility revolution" in econometrics.
- These new approaches, have on the whole been a positive influence.
- David Card's Nobel lecture 2022 is a good summary of the positive aspects of these developments.

What's good about the new approaches

- The emphasis on the value of finding direct measures of policy action, especially where the actions are arguably not entangled with complicated reverse causality.

- The recognition that when this is possible, it can lead to reliable estimates of policy effects with fewer strong assumptions about economic behavior or probability distributions.

Characteristics of Mostly Harmless econometrics

- It starts from analysis of the "ideal" inference problem: A randomized controlled trial (RCT), which is an experiment in which some subjects are "treated" with a potential policy and others are left "untreated", with the treatments randomly assigned, and a single outcome, in the simplest cases itself binary, is measured.

- It extends then to applications where assignment to treatment is assumed to be "as good as random", treatments can take on more than two values, selection bias is present, etc.

- Inference is frequentist, based on asymptotic approximations, and avoids making distributional assumptions explicit. I.e., no specific probability model of the data is put forward.

Characteristics of Mostly Harmless econometrics

- It encourages use of simple estimators, even when they might be inefficient or might fail to be informative about the distribution of the data.

- For example, use of OLS with "robust" standard errors is encouraged, on the grounds that OLS regression of Y on X under weak assumptions consistently estimates the best linear predictor of Y from X, even when $E[Y \mid X]$ might be nonlinear and $\mathrm{Var}(Y \mid X)$ might not be scalar.

"Sharp" econometrics: Uses models

- Understanding what model would make your estimates and procedures exactly correct in small samples is generally worthwhile, even when you've found a "natural experiment" that allows the model to be fairly simple.

- Of course manageable models are often parametric; we usually believe they are at best approximately correct. Using them involves making an assumption about how far they are from being correct.

- The alternative, applying asymptotic distribution theory while making (almost) no *explicit* assumptions about distributions in fact makes implicit prior assumptions about the model;

- We apply the asymptotic theory in a particular sample, for which the asymptotic theory may or may not be a reasonable approximation.
- Proceeding as if it is amounts to making assumptions about how far the actual probability model is from the one that would justify our procedures as exactly correct.

Sharp econometrics: Is comfortable with making probability statements about parameters

- Of course this means that it is at least informally Bayesian.
- This is particularly important when not just the sample size, but also the number of parameters, is large.
- For example, consider a simple treatment-effect setup with randomly assigned treatment, plus some control variables X. If X has just a few columns, a multivariate regression will give more accurate estimates. But if the number of columns in X is large, including them all will make matters worse. Pre-sample-probability inference makes this a puzzle. For Bayesian inference it's all the same model. Priors generate "regularization" in a principled way.

Sharp econometrics: Is unembarrassed about specification searches

- Specification searches are examinations and comparisons of multiple models, with new models possibly considered in response to first results. This is what we all do, and should not be embarrassed about doing despite the havoc such searches wreak on frequentist inference. (See Leamer (1978).)

- From a Bayesian perspective, all inference is about tracing out likelihood functions, across models as well as across parameter spaces within models.

Sharp econometrics: Is unembarrassed about specification searches

- So long as the full search is reported, it doesn't matter to inference what order the models were introduced, whether the model space was fully specified in advance, or whether models were introduced in response to seeing data. (This is the same reasoning that leads to the "stopping rule paradox".)

- We should be teaching students how to do, and to interpret, specification searches. Not teaching them to be uneasy about considering any hypothesis or model that was not announced in advance of seeing the data.

Sharp econometrics helps meet the challenge from machine learning

- Big data sets allow use of complex models with many parameters.
- Machine learning models implicitly use large numbers of parameters and allow for complex non-linearity, though they often abandon probability-based inference.
- Economists following the "mostly harmless" paradigm often end up presenting estimated linear equations in which every displayed estimated coefficient has three stars (significant at .001 level) after it.
- We know our models are very seldom complete or linear; when we see such results we should assume we are not done with the analysis, not accept the linear model's estimated "effect" as the end product.

Sharp econometrics tries to model heterogeneity, not just average over it

- In small-T dynamic panel models we often estimate fixed effects. These estimates are inconsistent, and the distribution of the estimates is not a good estimate of the distribution of the fixed effects themselves. Bayesian approaches have no difficulty with estimating the distribution of the true fixed effects, which is often central to the analysis. (See e.g. Liu (2017))

- When we are looking at the effect of a "treatment" and have a rich array of controls, there is a choice between settling for estimating the average effect, avoiding the need for elaborate modeling of the controls, and getting a much more useful mapping between the controls and the treatment effects.

Looking again at the Card-Krueger minimum wage paper

Stop for questions

- Their classic paper looked at the introduction of a minimum wage for fast-food workers in New Jersey, but not Pennsylvania, as a natural experiment.

- They estimated a regression of the change in employment between a period before the minimum wage took effect, to several months after it took effect, on a state indicator variable or a "gap" variable that measured how far below the minimum the store was initially.

Their result

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.490      1.008  -2.471  0.01390 *
STATENJ        2.943      1.123   2.621  0.00911 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.785 on 389 degrees of freedom
(19 observations deleted due to missingness)
Multiple R-squared:  0.01735,Adjusted R-squared:  0.01483
F-statistic: 6.869 on 1 and 389 DF,  p-value: 0.009112
```

Queries

- They combined separate data on full time and part time workers, with the part timers given weight .5.
- Why? Fast food establishments have tremendous labor turnover, many 150% per year. Employment and hours fluctuate week to week, and full time and part time employment are likely to have different, and related, dynamics.
- A system of two equations in the two employment series seems to make more sense, i.e. a small VAR.

Measurement error vs. non-trivial dynamics

- C&K had evidence, from a few cases where restaurants were contacted twice by accident, that there was error in the reported employment numbers. They allowed for this by treating it as mean-zero, serially uncorrelated, "classical" measurement error — itself a strong assumption.

- But to justify their regression specification, they then had to rule out significant dynamics — inertia in reaction to the minimum wage shock, e.g.

- Two models of the underlying dynamics that justify what they did are:

$$L_{it} = c_i + \varepsilon_{it}$$

$$L_{it} = L_{i,t-1} + \varepsilon_{it}$$

- If the true model involves both inertia and measurement error, the NJ dummy's effect is not identified.

- We're going to go ahead assuming dynamics, but no measurement error. This is not a better model — it's just different and interesting.

The two equations

```
lm(formula = EMPFT2 ~ EMPFT + EMPPT + STATE)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.52444    1.27978   1.973 0.049255 *
EMPFT        0.23734    0.04543   5.224 2.86e-07 ***
EMPPT        0.13404    0.03848   3.483 0.000552 ***
STATENJ      1.54859    0.98922   1.565 0.118290

lm(formula = EMPPT2 ~ EMPFT + EMPPT + STATE)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.85565    1.47963   3.958 9.00e-05 ***
EMPFT        0.29201    0.05257   5.555 5.15e-08 ***
EMPPT        0.56888    0.04445  12.798  < 2e-16 ***
STATENJ     -0.43847    1.14313  -0.384    0.702
```

Model fit

- Note that the two lagged employment variables' coefficients are all precisely estimated, while the state indicator makes only a marginally significant contribution to fit, in only one of the equations.

- The part-time variable shows a *decline* in employment in response to the minimum wage variable, though this effect is not statistically significant.

- Similar equations with the `gap` variable instead of the state indicators show even less contribution to fit from the policy variable.

Non-random assigment

- In the treatment-effects jargon, what is showing up here is a pair of "controls", lagged full- and part-time employment, that are strongly correlated with the "treatment", here the state indicator.

- Furthermore, introducing these controls changes the size and significance of the treatment effect estimates. So this is not a pure natural experiment, with "treatment" assigned "as good as randomly".

- To justify what C&K did, they needed the assumption of "common trends" — that the distribution of true *changes* in employment was the same in PA and NJ, except for the effects of the minimum wage legislation, and despite the fact that the distribution of *levels* of employment differed.

Aside: C&K's consideration of lagged employment

- Card and Krueger did consider a range of control variables, a few of which did reduce the significance of the state indicator.

- They realized one might consider including the lag of their aggregated employment variable, but noted that this variable's coefficient might not be one, if their employment variable had "measurement error", so they did not present results like this:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.77439    1.15071   6.756  5.2e-11 ***
STATENJ      1.42111    0.94398   1.505    0.133
emp1         0.48711    0.03925  12.409  < 2e-16 ***
```

(`emp2`, the dependent variable, and `emp1` are the first and second period values of the Card/Krueger employment aggregate.)

Aside: C&K's consideration of lagged employment

- They tried using the number of cash registers in the establishment, or the number of registers active at 11AM, as instruments for lagged employment.

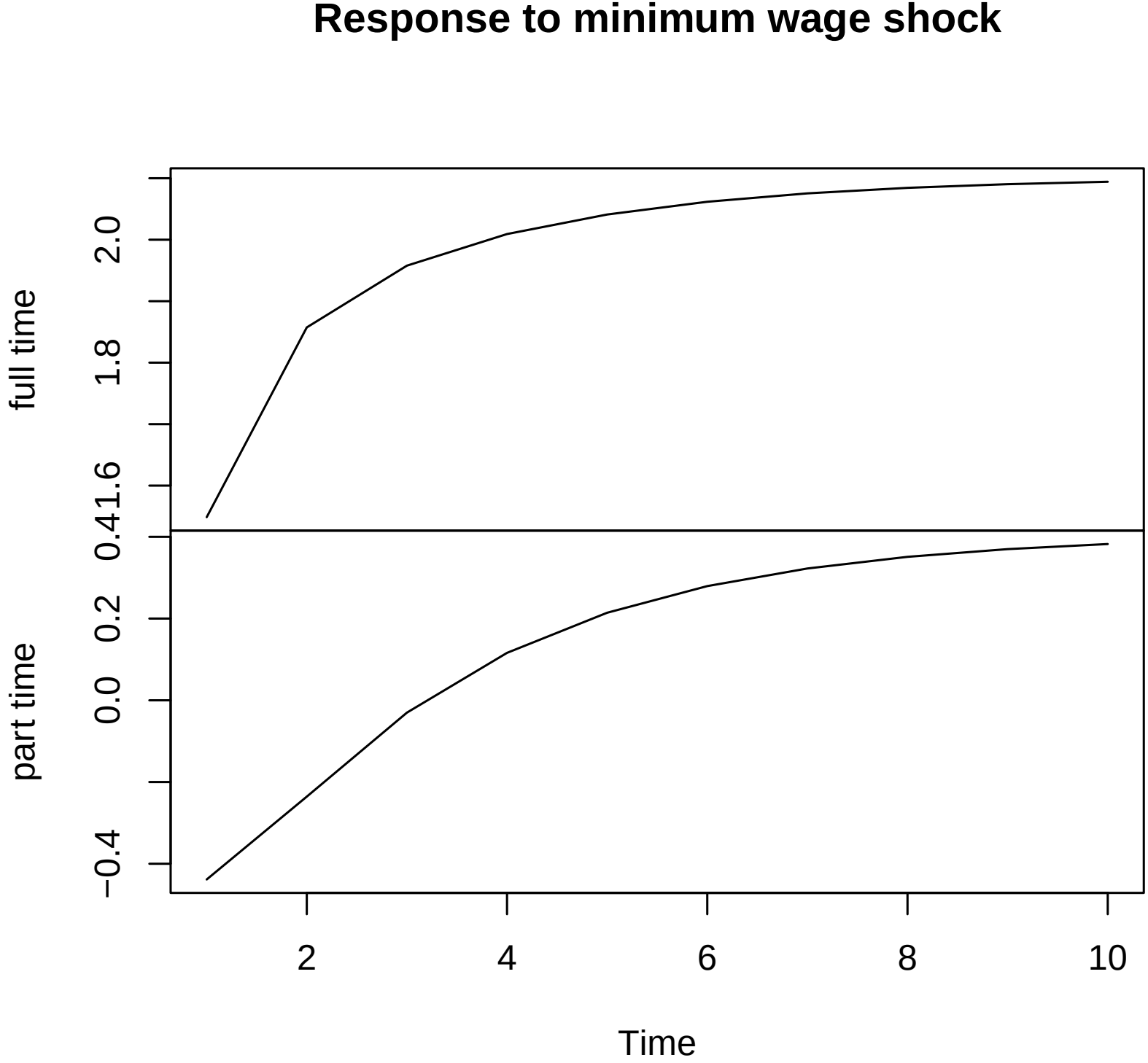
- If the C&K model were correct, including its assumption of i.i.d. measurement error in employment, and if in addition the distribution of first period employment were the same for both states as in a pure natural experiment, estimates with lagged employment on the right would give consistent, and more accurate, estimates of the NJ coefficient.

- The big change in the coefficient of NJ when the coefficient on lagged employment is left free implies, under the C&K model, a big difference in firm sizes across states, and thus that the "common trends" relaxation of the pure natural experiment assumption is required.

Dynamic response

- We can ask what the estimated system implies as the cumulative response over time to the minimum wage, assuming that the minimum wage persists in NJ and is not introduced in PA.

- This is shown on the next plot.


Response to minimum wage shock
full time
part time
1.6
1.8
2.0
0.4
0.2
0.0
−0.4
2
4
6
8
10
Time

Conclusion about minimum wage effects?

- The point estimate of the effects of the minimum wage, if we think that's what the NJ state indicator measures, imply convergence to an effect almost as large as the initial Card-Krueger estimates.

- However these effects are borderline significant at best. They might be the effects of the minimum wage, but most of the explanatory power of the minimum wage variable is absorbed by the lagged employment variables.

- Someone convinced the minimum wage had no effect on employment could interpret these results as showing that the state indicator variable is just picking up the effects of different initial distributions of employment in the two states, and that this becomes clear when we allow lagged employment variables in the system rather than just a single differenced aggregated employment.

An example specification search

Stop for questions

- I consider the data set used in the classic Angrist and Krueger (1991) study of the returns to education.
- What I'll show is based on a teaching exercise and is not itself an example of a complete specification search — just of initial steps on the path.

Data and initial model

- The data include log of wage, years of education (0 through 20), state of birth, year of birth and quarter of birth for over 300,000 men aged 40-49 in the 1970 census..

- The original paper focused on the possibility that OLS estimates of returns to schooling might be subject to selection bias — people who have higher ability might be more likely to stay in school, for example.

- The contribution of the paper was to note that quarter of birth, interacted with place of birth, could serve as an instrument for years of education and eliminate this bias.

- They showed that in fact there was little indication of bias, and in what I describe below I assume the selection bias is not important.

Starting point

The starting point is the model

$$logwage = c + \alpha\, educ + \beta\, yob + \varepsilon\ .$$

Estimates of it, using OLS, are

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1620767  0.0137642  375.04   <2e-16 ***
educ         0.0710812  0.0003390  209.68   <2e-16 ***
yob         -0.0049081  0.0003829  -12.82   <2e-16 ***
---
```

Nonlinearity

- It's quite plausible that a year of education has different effects at different stages. From the significance levels of t-statistics on coefficients, which are zero to machine precision, it's clear that we have room to estimate more coefficients.

- So let's let each level of education have a different coefficient. (Accomplished in R by just applying `as.factor()` to `educ`.)

Nonlinear education effects

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.1861454	0.0290997	178.220	< 2e-16	***
as.factor(educ)1	0.0408756	0.0505731	0.808	0.418949	
as.factor(educ)2	0.1417293	0.0366588	3.866	0.000111	***
as.factor(educ)3	0.1485511	0.0323701	4.589	4.45e-06	***
as.factor(educ)4	0.2401967	0.0312590	7.684	1.55e-14	***
as.factor(educ)5	0.3072038	0.0296365	10.366	< 2e-16	***
as.factor(educ)6	0.3811710	0.0279425	13.641	< 2e-16	***
as.factor(educ)7	0.4483723	0.0272370	16.462	< 2e-16	***
as.factor(educ)8	0.5500115	0.0264852	20.767	< 2e-16	***
as.factor(educ)9	0.6259080	0.0265909	23.538	< 2e-16	***
as.factor(educ)10	0.6650902	0.0264738	25.123	< 2e-16	***
as.factor(educ)11	0.7113150	0.0265701	26.771	< 2e-16	***
as.factor(educ)12	0.8249697	0.0260712	31.643	< 2e-16	***
as.factor(educ)13	0.8985823	0.0264311	33.997	< 2e-16	***

```
as.factor(educ)14  0.9437695  0.0263093  35.872  < 2e-16 ***
as.factor(educ)15  0.9883090  0.0267765  36.910  < 2e-16 ***
as.factor(educ)16  1.2120061  0.0262243  46.217  < 2e-16 ***
as.factor(educ)17  1.2105012  0.0266451  45.431  < 2e-16 ***
as.factor(educ)18  1.2491305  0.0266671  46.842  < 2e-16 ***
as.factor(educ)19  1.1923295  0.0272132  43.814  < 2e-16 ***
as.factor(educ)20  1.2918082  0.0266770  48.424  < 2e-16 ***
yob               -0.0048086  0.0003825 -12.571  < 2e-16 ***
```

Note that the coefficients on education are monotone increasing up through 16 years of education, but not beyond that. Also, the increments from year to year are not uniform, with college completion (16), high school graduation (12) and 8th grade graduation (8) standing out as larger. All coefficients are ridiculously "significant", except the first year of education (vs. none).

Are schooling effects monotone increasing?

The original article assumed uniform effects of education at all levels, but our point estimates are not completely monotone. The estimates are uncertain, though. Is there some substantial probability of monotone effects?

This is easily assessed in a Bayesian framework. With this sample size, the posterior on the residual variance is extremely precise, so we will treat its estimated value as known. The posterior joint distribution of the coefficients is then normal, and we can sample from it. Making a thousand draws from it, we can count how often the coefficient vector is monotone from 2 through 20. It turns out that not a single one of a thousand draws is monotone.

We can do the same thing for 2 through 16 years: 599 of 1000 draws are monotone over this range.

Next steps

- There's obviously room for furthe exploration of the non-linearity. Jumps at 12 and 16? Are returns to the first few years of graduate training negative? Or is this people who took extra time to get through undergraduate programs?

- But for our example we will turn now to looking at interactions. Suppose education technology has gotten better over time. Then interactions between schooling and year of birth might be important and might affect our estimates of returns to schooling.

Interactions

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(educ)	20	18506	925.31	2288.8783	< 2.2e-16	***
as.factor(yob)	9	67	7.41	18.3280	< 2.2e-16	***
as.factor(educ):as.factor(yob)	180	142	0.79	1.9477	4.577e-13	***
Residuals	329299	133123	0.40			

Once again everything is ridiculously "significant". But the interaction terms, despite delivering a highly significant p-value, generate an F statistic of only 1.95. This almost satisfies the Akaike criterion (which would look for an F of 2 or greater) and is well below the critical value for the BIC, which here, for the F statistic, is just log of sample size, which is over 12. Note that `yob` as a factor is accepted by both the AIC and BIC.

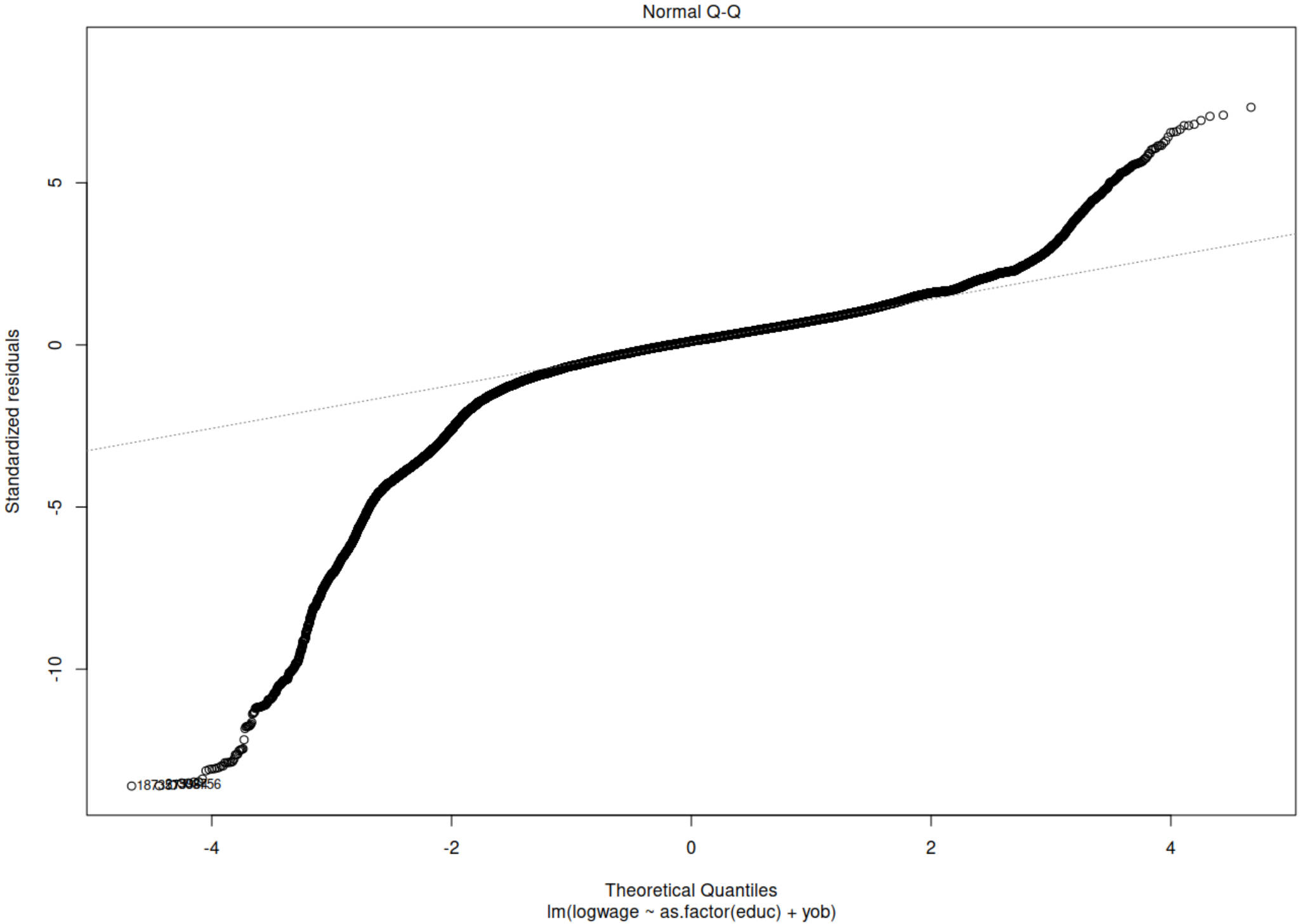
Limits of AIC and BIC

Though better than simple F tests, AIC and BIC are both arbitrary. There is no good way to assess posterior odds on models other than a fully Bayesian approach, using a prior, because with the large number of parameters we are considering here, priors are inevitably important. BIC is motivated by posterior odds, but it does not approximate them, even in large samples. So this is another loose end in our analysis, as we move to the next step.

Non-normality

- Because asymptotics guarantees nearly Gaussian posteriors in very large samples (assuming finite variance), non-normality of residuals may seem unimportant.

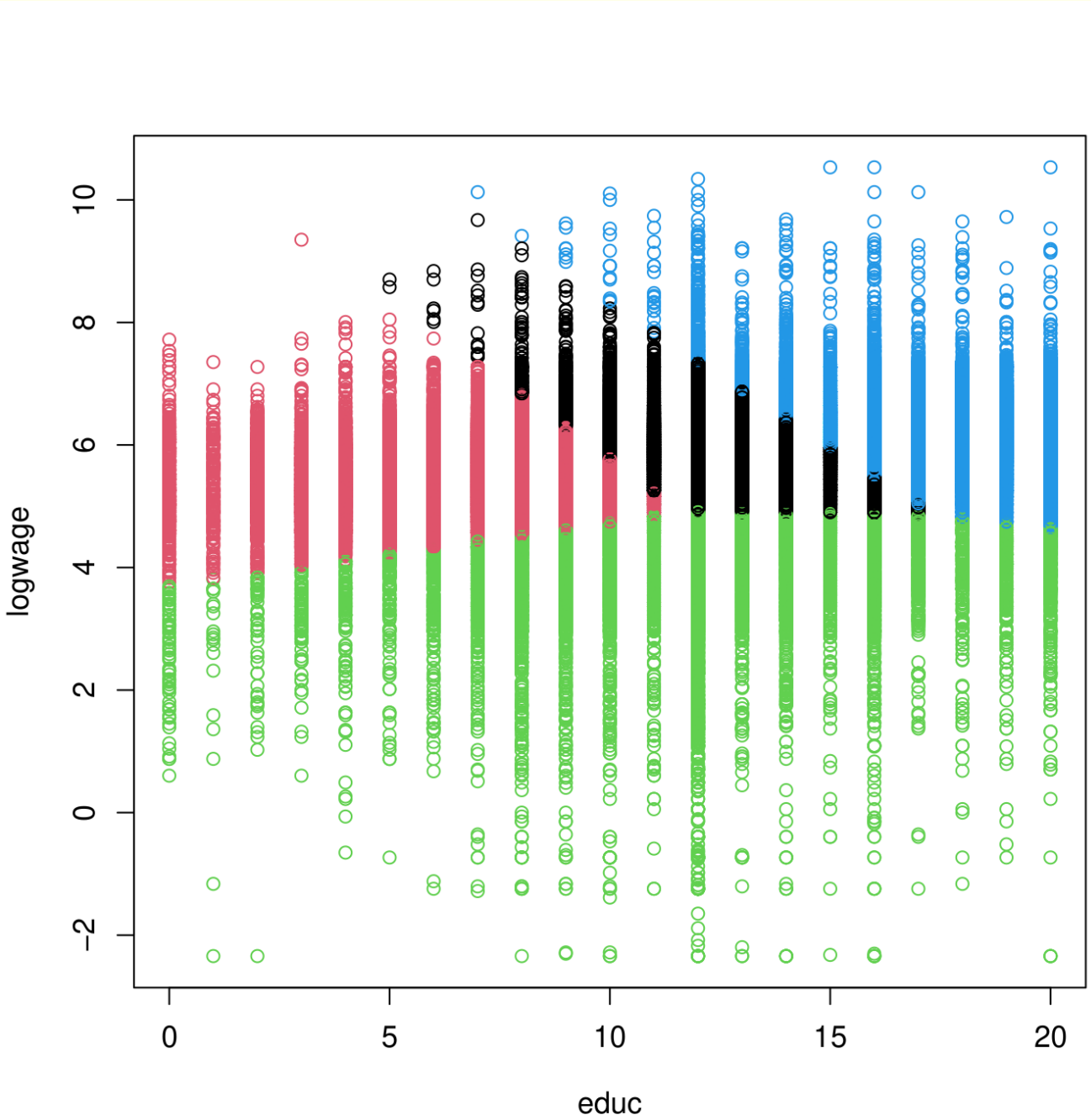
- But here, there is strong non-normality, and it suggests another dimension of specification search.


Normal Q-Q
Standardized residuals
5
0
-5
-10
-4
-2
0
2
4
Theoretical Quantiles
lm(logwage ~ as.factor(educ) + yob)

More indication of non-normality or non-linearity

- k-means is a simple, sub-optimal algorithm for organizing data into a finite number of groups.

- Applied to this data set's `logwage` and `educ` variables, with $k = 4$, it produces this.


logwage
−2
0
2
4
6
8
10
0
5
10
15
20
educ

Interpreting the groups

- Three correspond roughly to below-8, 9-12, and over 13 years of schooling, and their average `logwage` rises with the level of schooling.
- But the fourth group spans all the education levels, and has lower `logwage` than the other groups.
- This fourth group is smaller than the others, but large enough to affect estimates.

What to do about it

- The data for wages are in logs, and the raw data have been increased by a small constant to avoid infinite values. The existence of this low-wage group suggests we should check sensitivity to the constant and consider alternatives to the log transformation.

- A fairly straightforward way to model this would be to use a "mixture of regression models" approach. See Norets (2010).

- Does recognizing the existence of a group of people who have low incomes and education of all levels affect estimates of expected returns to education?

- This depends on how people end up in the low wage group — medical disaster, choice of rewarding but low-wage occupation,. . . ? This data set probably can't answer these questions.

*

References

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Angrist, J. D., and J. steffen Pischke (2009): *Mostly Harmless Econometrics*. Princeton University Press.

Card, D. (2022): "Design-Based Research in Empirical Microeconomics," *American Economic Review*, 112(6), 1773–1781.

Leamer, E. E. (1978): *Specification Searches: Ad Hoc Inference with Non Experimental Data*. John Wiley and Suns.

Liu, L. (2017): "Density Forecasts in Panel Data Models: A Semiparametric Bayesian Perspective," Discussion paper, University of Pennsylvania.

Norets, A. (2010): "Approximation of conditional densities by smooth mixtures of regressions," *Annals of Statistics*, 38(3), 1733–1766.