Problem Set 1, Question 4: Follow-up

Mitch Downey
Econometrics I

January 29, 2024

Original question: A researcher is interested in whether math teachers exert more effort when their classes have more boys. She finds a set of schools where students are randomly assigned to classes, and thinks this is a good opportunity to test her hypothesis because the gender composition of the class will be exogenously determined. Assume that the female fraction of students in each school is 1/2 and that students are indeed randomly assigned to classes. She observes effort exerted in N different classes, each of which has k students.

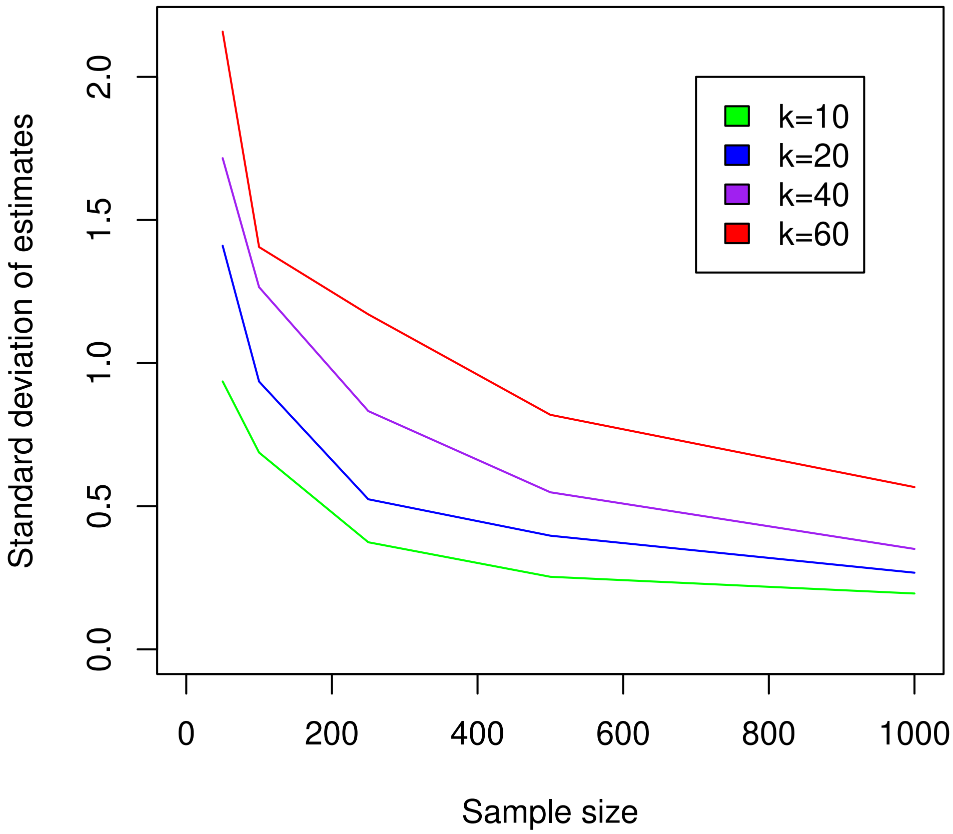
(a) Assume that the researcher's hypothesis is correct, and that the data generating process for the effort of the teacher of class i (written as y_i) is determined according to $y_i = \beta m_i + \varepsilon_i$ where m_i is the fraction of students in class i who are male and $\varepsilon_i \sim N(0, 1)$. Run 500 simulations: 100 for each $N \in \{50, 100, 250, 500, 1000\}$. In each case, simulate the data holding k fixed at 40 and $\beta = 1/3$, regress y_i on m_i, and save the estimate of $\hat{\beta}$. Calculate the standard deviation of $\hat{\beta}$ across each of the 100 iterations of your simulation. This gives you five different standard deviations from your five samples: $\sigma_{\hat{\beta}}$ for each $N \in \{50, 100, 250, 500, 1000\}$. Plot $\sigma_{\hat{\beta}}$ (y-axis) against N (x-axis).

(b) Run 1500 more simulations simulations with the same data generating process. Use the same vector of five values of N, and run the simulation for $k \in \{10, 20, 60\}$. For each combination of N and k, calculate $\sigma_{\hat{\beta}}$. Add these three additional series of $\sigma_{\hat{\beta}}$ (y-axis) against N (x-axis).

(c) One rule of thumb in applied research is that having a larger sample improves the reliability of estimates. Another rule of thumb is that having more variation improves the reliability of estimates. Relate these rules of thumb to your above answers.

(d) The researcher has currently collected data for a nationally representative sample of 200 classes which have, on average, 30 students. She has recently obtained funding to expand her sample. She can choose to use the funding in an urban area, where she could get more classes but they would be larger (another 100 40-student classes), or a rural area where she could get fewer classes but they would be smaller (another 50 15-student classes). Which would you recommend?

Figure 1: Sample standard deviation of $\hat{\beta}$: $\sigma_{\hat{\beta}}$



Notes:

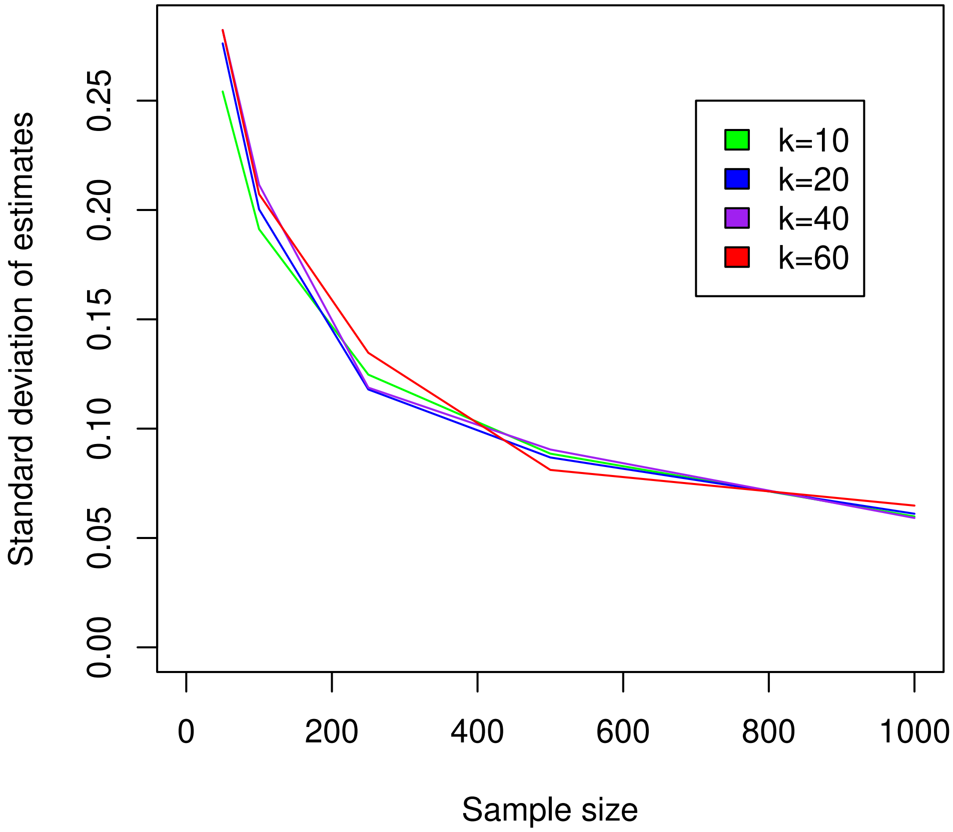
Follow-up discussion: (Note: Throughout this discussion, I use m_i and x_i interchangeably; sorry.) Figure 1 shows the original solution to parts a and b. It shows that the variance of the estimates is increasing in k, because the WLLN implies that the variation in m_i is progressively killed off as class sizes get larger.

Several students noticed that if you run the regression without a constant, then you do not get this pattern. Also interesting, you find that the standard deviation is incredible low. This is shown in Figure 2 (note that the y-axis is much smaller). Why?

This is an artifact of the fact that running a regression without an intercept mechanically forces $\hat{y}_{m_i=0}$ to be zero. Then most of the data is actually located quite far from $m_i = 0$, and so you're basically fitting a line through two points: $(0, 0)$ and $(\bar{m}_i, \bar{y})$.

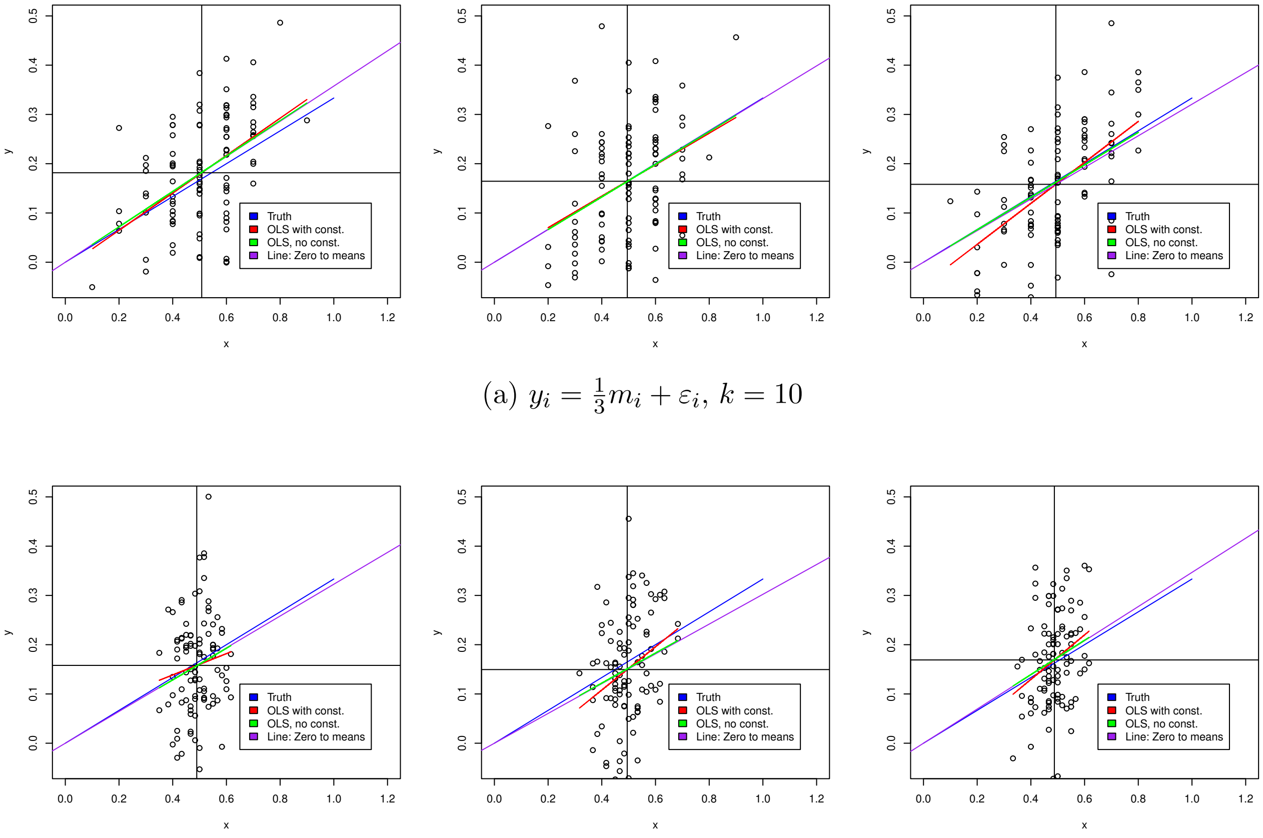
To see what's going on here, Figure 3 plots the underlying relationship between y_i and m_i for three iterations with $k = 10$ and three iterations with $k = 60$. In each case, relative to the problem set, I reduced the variance of ε_i to make it a little easier to see what's going on.

Figure 2: Sample standard deviation of $\hat{\beta}$: $\sigma_{\hat{\beta}}$ (suppressing the constant)



Notes:

Figure 3: Scatterplots of the relationship identifying $\hat{\beta}$



(a) $y_i = \frac{1}{3}m_i + \varepsilon_i$, $k = 10$

(b) $y_i = \frac{1}{3}m_i + \varepsilon_i$, $k = 60$

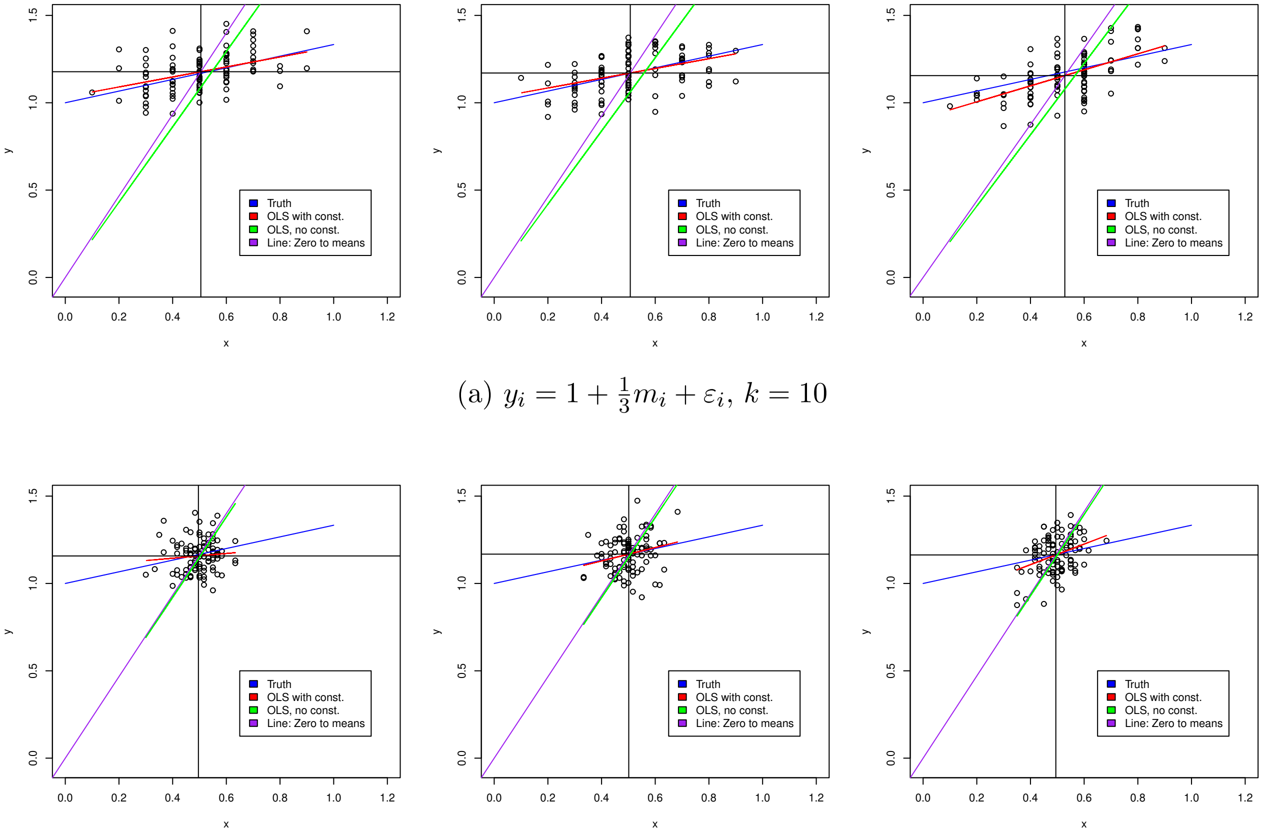
Each panel shows four lines of fit. The blue is the truth ($y = x/3$). The red line is the OLS with a constant. Looking at the red lines, it's clear that the bottom panels just have way less variation in x, and that's the point of the exercise. So then even though the slope is the same as in the top panels, because you have very little range of observed x values to infer this slope from, you'll end up with much noisier estimates. You see a bit of this in the red lines (which are maybe more bouncy around the blue line in the $k = 60$ case in the bottom panel than the $k = 10$ case in the top), but even if it's not clear in these 3 iterations, it's clear how and why this will happen.

But the green line isn't simply fitting the dots. The green line also mechanically goes through (0,0). In this case, the truth (the blue line) does too, so that's ok. In fact, it's helpful because imposing true assumptions usually buys you power.

OLS mechanically goes through the mean. This is a consequence of the estimated residuals being mean zero because $y_i = x_i'\beta + \hat{\varepsilon}_i \ \forall \ i \Rightarrow \frac{1}{n}\sum_i y_i = \frac{1}{n}\sum x_i'\beta + \frac{1}{n}\sum \hat{\varepsilon}_i \Rightarrow \bar{y} = \bar{x}'\beta$. This property is only true if your model has a constant (otherwise, the estimated residuals need not be mean zero), but it gives good intuition. The green line (which omits a constant) roughly goes through the mean of the data (denoted by vertical and horizontal black lines), as well as (0,0). To see this, the purple line is just a line segment connecting (0,0) and $(\bar{x}, \bar{y})$. The green line almost perfectly corresponds to the purple line in every case (they don't literally correspond to one another; they are different at the fourth decimal point). But since the true relationship also goes through (0,0), this is kind of fine. But be clear about what this is doing: The green line (without a constant) is not so much based on fitting the observed relationship between x and y that we see in the data, it's instead based on drawing a line connecting (0,0) – an externally imposed constraint – with the means of the data. So this will give a very stable estimate of $\hat{\beta}$, but it's very stable because it basically has nothing to do with the actual data you observe.

The easiest way to see the problem here is to change the data generating process to $y_i = 1 + \frac{1}{3}m_i + \varepsilon_i$ where the distribution of ε_i is exactly the same as before, and the student-level random assignment mechanism that generates m_i is the same, too. This is done in Figure 4. It's still true that the purple line is connecting (0,0) with the mean, and the green line is approximately doing that. The green line is a bit off of the purple because it's somewhat influenced by the slope of the dots, while the purple line isn't at all, and the green line deviates from the purple more when there's more variation in m than when there's less (i.e., when $k = 60$ and you have very little identifying variation in the data, that's when your externally imposed assumptions have the most influence). But now, going through (0,0) and $(\bar{m}, \bar{y})$. Still, you'll note how stable the slope of that green line is. So the variance is very low, but that doesn't mean it's a *good* estimate; it's just a stable one. The red line bounces around more (and especially when there's not so much variation in x to identify the slope) but is much close to the blue line of truth.

Figure 4: Scatterplots of the relationship identifying $\hat{\beta}$ with a constant



(a) $y_i = 1 + \frac{1}{3}m_i + \varepsilon_i$, $k = 10$

(b) $y_i = 1 + \frac{1}{3}m_i + \varepsilon_i$, $k = 60$