Problem Set 03

Aaron Graybill
Collaborators: John Cho, Thomas Li, Blair Moore

2025-10-5

Contents

If you're having trouble running my code, see https://github.com/aarongraybill/MTGECON_603. Also the spacing on the tables text are really not ideal, Rmarkdown and LaTeX are not playing nicely, sorry about that!

1 Problem 1

1.1 1.a

The results for parts 1.a through 1.e are summarized in table 1.

For part 1.a, we can see that vanilla bootstrapped confidence interval for both 1 and 4 year earnings are significantly larger than zero. Under the assumption that the bootstrap distribution is approximately normal under this large of a population and large number of bootstrap samples, 10^4, we can be reasonably sure that the treatment has a positive effect on the earnings and that is is very unlikely to be larger than 2 thousand dollars.

1.2 1.b.

The problem set only asks for bootstrap weights generated from the exponential distribution, but for my interest, I also generated Dirichlet weights where each of the bootstrap samples draws one realization from a Dirichlet distribution with parameter: $\alpha = (1 - (1/N), \ldots, 1 - (1/N))$. Where N is the total number of units. I then multiply these draws by N so that their sum is N. The $1 - (1/N)$ is to ensure that the variance of any individual observation matches the binomial variance of $1 - (1/N)$, though I'll omit the proof of this.

The confidence intervals from both the exponential and Dirichlet Bayesian bootstrap are quite similar to the vanilla bootstrap. So the interpretation is essentially identical. Any differences between the confidence intervals may simply be due to differences in realizations of the random sampling or due to slight differences in the nature of the estimator. Though it's tough to distinguish these effects in our context.

1.3 1.c.

We can also bootstrap the t-statistic instead of the actual estimated treatment effect. In this case, the effect is pretty minimal, the confidence interval is qualitatively identical. That said, if the estimator was much less normally distributed, we might expect the t-state to better capture the significance of the results.

For one-year earnings, I estimated the 5th and 95th percentiles of the t distribution to be: -1.6661874, 1.6510519. For four-year earnings, I estimated the 5th and 95th percentiles of the t distribution to be: -1.6488266, 1.6345872

Here, the significance of the results requires slightly different interpretation. Here, we are saying that we are 90% confident that the t-statistic is in the range specified in 1. So we have strong evidence to believe that the centered and rescaled variable is significantly different from zero and lies in the specified range.

1.4 1.d.

In this section, I compute the confidence intervals from a bootstrapped 74 unit without-replacement sample from the empirical distribution. Here, the confidence intervals for both variables are very modestly wider, reflecting the additional variation in smaller samples, although the overall effect is small. The interpretation of the confidence intervals is still the same as in the previous sections.

1.5 1.e.

For the custom bootstrap test, I have elected to use the difference in stratum means weighted by stratum size. This is the same metric discussed in question 2. This is a natural estimator because it allows us to control for between-stratum variation without imposing parametric assumptions on the relationship between high school completion and earnings. So in each iteration of the bootstrap, I randomly draw N rows from the original data with replacement. Then I split the data by highschool completion, then I compute the difference in means for treated and untreated units in those draws. To compute the test statistic, I take the weighted average of those differences in means, weighted by the number of units in that stratum for that bootstrap iteration. So the weights vary by the actual draws from the bootstrap.

To compute the 95% confidence interval, I follow parts 1a, 1b, and 1e in which I compute the compute the value of the test statistic in each of the bootstrap draws, then compute the empirical standard error across each of those bootstrap draws. To be precise, I compute

$$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}^{b}_{strat}-\hat{\tau}_{strat})^2}$$

where B is the number of bootstraps, $\hat{\tau}_{strat}$ is the value of the test statistic on the original data, and $\hat{\tau}^{b}_{strat}$ is the value of the test statistic for each draw from the bootstrap distribution. To compute the confidence interval, I used those standard errors, multiplied them by the 5th and 95th percentiles for a standard normal distribution, and added those to $\hat{\tau}_{strat}$.

Here the 95% confidence interval is wider than in previous questions. That is mostly mechanical, 95% confidence intervals must be wider than 90% confidence intervals given the same distribution.

That said, the distribution of the test statistics should not be exactly identical under this new stratum level estimator. However, it turns out to be almost exactly the same as previous questions. Figure 1 plots kernel density estimates of the distributions of the test statistics under the various bootstrap regimes. The main takeaway is the subsampling approach introduces much more variation, that's mostly mechanical and is compensated for when computing confidence intervals. Otherwise, each of the methods produces nearly identical distributions of test statistics.

Table 1: Confidence intervals from various bootstraps. All reports, except for custom are 90 percent confidence intervals. The custom confidence interval is 95 percent.

method	variable	true_test_stat	lb	ub
vanilla	1 Year	1.136210	0.9134766	1.358944
vanilla	4 Year	1.232332	0.8232161	1.641448
bayesian_exponential	1 Year	1.136210	0.9143618	1.358059
bayesian_exponential	4 Year	1.232332	0.8220337	1.642631
bayesian_dirichlet	1 Year	1.136210	0.9144850	1.357935
bayesian_dirichlet	4 Year	1.232332	0.8169110	1.647753
boot_t_stat	1 Year	1.136210	0.9147402	1.359711
boot_t_stat	4 Year	1.232332	0.8256255	1.642582
subsample	1 Year	1.136210	0.9090065	1.363414
subsample	4 Year	1.232332	0.8189953	1.645669
custom	1 Year	1.136216	0.8750108	1.397421
custom	4 Year	1.232352	0.7501850	1.714518

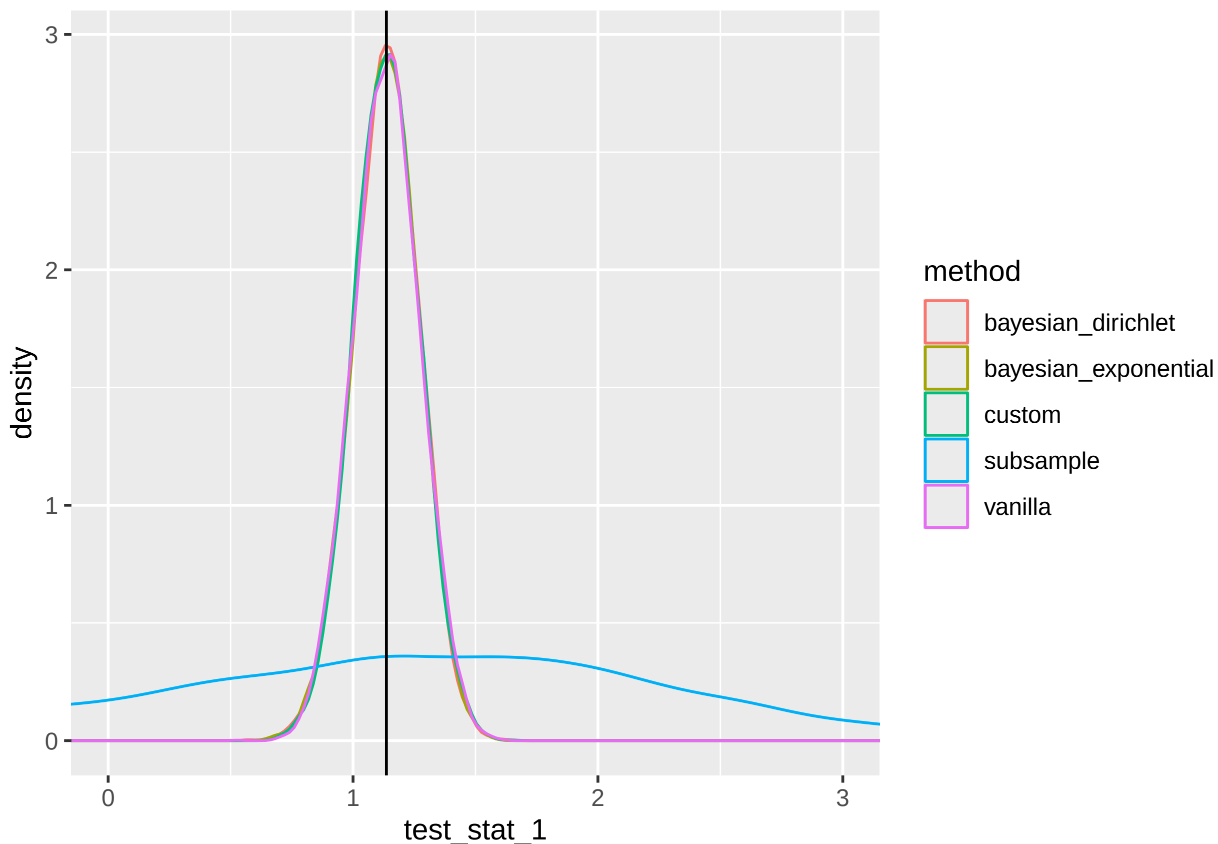

Figure 1: Kernel densities of bootstrapped test statistics under various bootstrap methodology.

2 Problem 2

2.1 2.a

I would probably advocate for the stratified analysis. While we don't have information on the dependent variable, I will assume that this is a dependent variable that is correlated with age—a common feature in a variety of social science contexts. In this case, the post-hoc stratified estimation should have a lower expected variance than the unstratified version because it is able to account for the between-age variance in the dependent variables in a way that the unstratified analysis does not. That said, since the number of observations is relatively small, if age is not correlated with the outcome, the post-hoc stratification may add additional variance to the estimator.

2.2 2.b

We've already shown in lecture that the difference in sample treated and control means is an unbiased estimator for the average treatment effect. But here's a quick re-derivation. First note that when we're not conditioning on the number of treated in each group, we know that $\mathbb{E}[W_i] = \frac{M}{N}$. Using that we have

$$\begin{aligned}
\mathbb{E}[\hat{\tau}_{dif} - \tau] &= \mathbb{E}[\tau_{dif} - \tau] \\
&= \mathbb{E}[\hat{\tau}_{dif}] - \mathbb{E}[\tau] \\
&= \mathbb{E}[\hat{\tau}_{dif}] - \tau \\
&= \mathbb{E}[\hat{\tau}_{dif}] - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{N} W_i Y_i(T) - \frac{1}{N-M}\sum_{i=1}^{N}(1-W_i)Y_i(C) \right] - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= \left(\frac{1}{M}\sum_{i=1}^{N} \mathbb{E}[W_i Y_i(T)] - \frac{1}{N-M}\sum_{i=1}^{N}\mathbb{E}[(1-W_i)Y_i(C)] \right) - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= \left(\frac{1}{M}\sum_{i=1}^{N} \mathbb{E}[W_i] Y_i(T) - \frac{1}{N-M}\sum_{i=1}^{N}\mathbb{E}[(1-W_i)]Y_i(C) \right) - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= \left(\frac{1}{M}\sum_{i=1}^{N} \frac{M}{N} Y_i(T) - \frac{1}{N-M}\sum_{i=1}^{N}\frac{N-M}{N}Y_i(C) \right) - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i(T) - \frac{1}{N}\sum_{i=1}^{N} Y_i(C) \right) \\
&= 0
\end{aligned}$$

Now we can show that the (weighted) aggregate of the difference in means by strata, τ_{strat} is an unbiased estimator for τ. I think the easiest way to approach this is by conditioning on the number of old and young assignments. So denote $N_O + N_Y = N$ to be the old and young total populations respectively. Next denote, $\tilde{M}_O, \tilde{M}_Y$ as the stochastic number of treated units for old and young respectively (this is stochastic since we're doing post-hoc stratificiation).

Notice however, that N_O, N_Y are not stochastic and that $\tilde{M}_O + \tilde{M}_Y = M$ is not stochastic (in our specific case, there are always 50 treated, only the breakdown of old versus young is stochastic). Further dentoe A_i to be the non-stochastic indicator for whether or not a unit is old. $A_i = 1$ implies that unit i is old. Conditioning on specific values of $\tilde{M}_O, \tilde{M}_Y, A_i$ we see that:

$$\mathbb{E}[W_i | \tilde{M}_O = M_O, \tilde{M}_Y = M_Y, A_i = 1] = \frac{\binom{N_O - 1}{M_O - 1}\binom{N_Y}{M_Y}}{\binom{N_O}{M_O}\binom{N_Y}{M_Y}} = \frac{M_O}{N_O}$$

The denominator in the intermediate fraction is the total number of ways to choose the M_O treated old individuals from the N_O options multiplied by each of the ways to do the same for the young person. For the numerator, we have number of ways to choose the remaining $M_O - 1$ treatments from the other $N_O - 1$ mutliplied by the same amount of combinations from the young individuals since their combinations are not affected by whether or not a specific old person is treated or untreated (once you've conditioned on M_Y).

Similarly, for a young unit we have:

$$\mathbb{E}[W_i|\tilde{M}_O = M_O, \tilde{M}_Y = M_Y, A_i = 0] = \frac{\binom{N_O}{M_O}\binom{N_Y-1}{M_Y-1}}{\binom{N_O}{M_O}\binom{N_Y}{M_Y}} = \frac{M_Y}{N_Y}$$

Linearity of expectations or more combinatorics gives: $\mathbb{E}[(1 - W_i)|\tilde{M}_O = M_O, \tilde{M}_Y = M_Y, A_i = 1] = \frac{N_O - M_O}{N_O}$ and $\mathbb{E}[(1 - W_i)|\tilde{M}_O = M_O, \tilde{M}_Y = M_Y, A_i = 0] = \frac{N_Y - M_Y}{N_Y}$

Now we proceed to compute $\mathbb{E}[\hat{\tau}_{strat} - \tau]$ using the law of total expectation. Law of total expectation gives that $\mathbb{E}[\hat{\tau}_{strat} - \tau] = \mathbb{E}[\mathbb{E}[\hat{\tau}_{strat} - \tau|M_O, M_Y]]$. If every interior expectation is zero, then it must be that the outer expectation is zero, so if we can show that $\mathbb{E}[\hat{\tau}_{strat} - \tau|M_O, M_Y] = 0$ for all M_O, M_Y, then we know that the estimator is unbiased unconditionally. We proceed as follows:

$$
\begin{aligned}
\mathbb{E}[\hat{\tau}_{strat} - \tau | M_O, M_Y] = \mathbb{E}\Bigg[& \left(\frac{N_O}{N} \left(\frac{1}{M_O} \sum_{A_i=1} W_i Y_i(T) - \frac{1}{N_O - M_O} \sum_{A_i=1} (1 - W_i) Y_i(C) \right) + \right. \\
& \left. \frac{N_Y}{N} \left(\frac{1}{M_Y} \sum_{A_i=0} W_i Y_i(T) - \frac{1}{N_Y - M_Y} \sum_{i=0} (1 - W_i) Y_i(C) \right) \right) - \\
& \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} \right) | M_O, M_Y \Bigg] \\
= & \left(\frac{N_O}{N} \left(\frac{1}{M_O} \sum_{A_i=1} \mathbb{E}[W_i Y_i(T) | M_O, M_Y, A_i = 1] - \frac{1}{N_O - M_O} \sum_{A_i=1} \mathbb{E}[(1 - W_i) Y_i(C) | M_O, M_Y, A_i = 1] \right) \right. \\
& \left. \frac{N_Y}{N} \left(\frac{1}{M_Y} \sum_{A_i=0} \mathbb{E}[W_i Y_i(T) | M_O, M_Y, A_i = 0] - \frac{1}{N_Y - M_Y} \sum_{A_i=0} \mathbb{E}[(1 - W_i) Y_i(C) | M_O, M_Y, A_i = 0] \right) \right) \\
& \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} \right) \\
= & \left(\frac{N_O}{N} \left(\frac{1}{M_O} \sum_{A_i=1} \frac{M_O}{N_O} Y_i(T) - \frac{1}{N_O - M_O} \sum_{A_i=1} \frac{N_O - M_O}{N_O} Y_i(C) \right) + \right. \\
& \left. \frac{N_Y}{N} \left(\frac{1}{M_Y} \sum_{A_i=0} \frac{M_Y}{N_Y} Y_i(T) - \frac{1}{N_Y - M_Y} \sum_{A_i=0} \frac{N_Y - M_Y}{N_Y} Y_i(C) \right) \right) - \\
& \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} \right) \\
= & \left(\frac{1}{N} \left(\sum_{A_i=1} Y_i(T) - \sum_{A_i=1} Y_i(C) \right) + \right. \\
& \left. \frac{1}{N} \left(\sum_{A_i=0} Y_i(T) - \sum_{A_i=0} Y_i(C) \right) \right) - \\
& \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} \right) \\
= & \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} Y_i(C) \right) - \\
& \left(\frac{1}{N} \sum_{i=1}^{N} Y_i(T) - \frac{1}{N} \sum_{i=1}^{N} \right) \\
= & \, 0
\end{aligned}
$$