Microeconometrics I – Problem Set 02

Vinícius de Almeida Nery Ferreira

August 27th, 2024

Question 1 – Standardization

(a) Costs and Benefits of Standardization

As said in the question, one of the main benefits of standardizing test scores is to ensure better comparability across tests from different years, as their difficulty and distribution can vary considerably.

Another benefit is that it is easier to identify outliers, as standardized scores are better in displaying *relative* performance. However, this is also a disadvantage, as standardized metrics lose the variable's *absolute* interpretation. Furthermore, the original scale itself may have some useful information, as the minimum grade required to pass the grade, for example.

In our specific case, we also have to account for the fact that the use of only control group information leads to some sampling uncertainty, as although, lotteries are fair, we may get an unlucky draw. In the 9th grade, we also have to account for the fact that enrollment is not only a function of winning the lottery, but also of the decision to participate in it, and so we may have differences between the treated and control groups depending on our target parameter of interest.

(b) Use of Control Statistics

The use of only the control statistics for the standardization is reasonable because then we are measuring all outcomes *relative* to a baseline that is not affected (or "contaminated") by treatment. Furthermore, since lotteries are fair, we expect that these statistics reflect, on average, what would have happened to treated units had they not enrolled in the school.

Question 2 – Balance Tests

(a), (c) Balance Test Table

Table 1: Balance Test Results

	Items (a) and (b)														Items (c)-(e)						
	"Treatment": Enrollment in 8th Grade							"Treatment": Enrollment in 9th Grade							"Treatment": Participated in Lottery B						
	Treated		Control		Difference T - C			Treated		Control		Difference T - C			Treated		Control		Difference T - C		
	Mean	SE	Mean	SE	Diff.	SE	***p***	Mean	SE	Mean	SE	Diff.	SE	***p***	Mean	SE	Mean	SE	Diff.	SE	***p***
Language Score (7th Grade)	5.00	0.16	4.97	0.10	0.03	0.19	0.88	5.31	0.22	5.32	0.17	-0.01	0.27	0.98	5.32	0.13	4.55	0.14	0.77	0.20	0.00**
Math Score (7th Grade)	5.01	0.12	5.08	0.08	-0.07	0.15	0.62	5.38	0.17	5.63	0.13	-0.25	0.21	0.24	5.53	0.10	4.54	0.11	1.00	0.15	0.00**
Std. Language Score (7th Grade)	0.01	0.09	0.00	0.05	0.02	0.10	0.88	0.18	0.12	0.18	0.09	0.00	0.15	0.98	0.18	0.07	-0.23	0.08	0.41	0.10	0.00**
Std. Math Score (7th Grade)	-0.03	0.08	0.02	0.05	-0.05	0.10	0.62	0.22	0.11	0.39	0.09	-0.17	0.14	0.24	0.32	0.07	-0.35	0.08	0.68	0.10	0.00**
Sex: Male	0.43	0.04	0.40	0.03	0.03	0.05	0.53	0.35	0.06	0.40	0.05	-0.05	0.07	0.52	0.38	0.04	0.42	0.04	-0.04	0.05	0.48
Sex: Female	0.57	0.04	0.60	0.03	-0.03	0.05	0.53	0.65	0.06	0.60	0.05	0.05	0.07	0.52	0.62	0.04	0.58	0.04	0.04	0.05	0.48
Race: White	0.24	0.04	0.27	0.02	-0.04	0.04	0.41	0.32	0.06	0.28	0.04	0.04	0.07	0.57	0.30	0.03	0.24	0.04	0.06	0.05	0.26
Race: Brown	0.51	0.04	0.53	0.03	-0.02	0.05	0.66	0.48	0.06	0.55	0.05	-0.07	0.08	0.38	0.52	0.04	0.55	0.04	-0.03	0.06	0.62
Race: Black	0.20	0.03	0.16	0.02	0.04	0.04	0.3	0.18	0.05	0.12	0.03	0.06	0.06	0.28	0.15	0.03	0.17	0.03	-0.02	0.04	0.59
Race: Yellow	0.02	0.01	0.03	0.01	-0.01	0.01	0.62	0.01	0.01	0.05	0.02	-0.03	0.03	0.19	0.03	0.01	0.02	0.01	0.01	0.02	0.45
Race: Indigenous	0.03	0.02	0.01	0.01	0.02	0.02	0.12	-	-	-	-	-	-	-	0.00	0.00	0.02	0.01	-0.02	0.01	0.08*
Mother's Education: <5th Grade	0.07	0.02	0.10	0.02	-0.03	0.03	0.21	0.08	0.03	0.15	0.03	-0.07	0.05	0.17	0.12	0.02	0.07	0.02	0.05	0.03	0.13
Mother's Education: 5th-9th Grade	0.21	0.03	0.23	0.02	-0.02	0.04	0.58	0.21	0.05	0.25	0.04	-0.04	0.06	0.50	0.24	0.03	0.23	0.03	0.01	0.05	0.87
Mother's Education: 9th Grade-HS	0.15	0.03	0.12	0.02	0.03	0.03	0.34	0.15	0.04	0.12	0.03	0.03	0.05	0.55	0.14	0.03	0.09	0.02	0.04	0.04	0.25
Mother's Education: HS-College	0.31	0.04	0.31	0.03	-0.01	0.05	0.87	0.30	0.05	0.31	0.05	-0.02	0.07	0.83	0.31	0.03	0.32	0.04	-0.02	0.05	0.71
Mother's Education: College	0.12	0.03	0.09	0.02	0.03	0.03	0.39	0.08	0.03	0.06	0.02	0.03	0.04	0.49	0.07	0.02	0.11	0.03	-0.05	0.03	0.15
Mother's Education: >College	0.07	0.02	0.06	0.01	0.01	0.02	0.7	0.07	0.03	0.05	0.02	0.02	0.04	0.53	0.06	0.02	0.06	0.02	0.00	0.03	0.87
Mother's Education: Unknown	0.08	0.02	0.09	0.02	0.00	0.03	0.87	0.10	0.04	0.06	0.02	0.04	0.04	0.32	0.07	0.02	0.10	0.02	-0.03	0.03	0.38
Father's Education: <5th Grade	0.11	0.03	0.11	0.02	0.00	0.03	0.94	0.11	0.04	0.15	0.03	-0.04	0.05	0.46	0.14	0.03	0.08	0.02	0.05	0.03	0.12
Father's Education: 5th-9th Grade	0.25	0.04	0.22	0.02	0.03	0.04	0.49	0.15	0.04	0.25	0.04	-0.09	0.06	0.14	0.21	0.03	0.24	0.04	-0.03	0.05	0.53
Father's Education: 9th Grade-HS	0.15	0.03	0.10	0.02	0.05	0.03	0.13	0.11	0.04	0.08	0.03	0.03	0.05	0.55	0.10	0.02	0.10	0.03	-0.01	0.03	0.86
Father's Education: HS-College	0.22	0.03	0.28	0.02	-0.05	0.04	0.21	0.31	0.06	0.26	0.04	0.05	0.07	0.52	0.28	0.03	0.27	0.04	0.01	0.05	0.84
Father's Education: College	0.05	0.02	0.07	0.01	-0.02	0.02	0.3	0.07	0.03	0.07	0.02	0.00	0.04	0.91	0.07	0.02	0.07	0.02	-0.01	0.03	0.81
Father's Education: >College	0.02	0.01	0.01	0.00	0.01	0.01	0.26	0.01	0.01	0.00	0.00	0.01	0.01	0.32	0.01	0.01	0.01	0.01	0.00	0.01	0.90
Father's Education: Unknown	0.20	0.03	0.21	0.02	-0.02	0.04	0.7	0.23	0.05	0.19	0.04	0.04	0.06	0.56	0.20	0.03	0.22	0.03	-0.02	0.05	0.65
Never Failed	0.68	0.04	0.74	0.02	-0.06	0.05	0.18	0.79	0.05	0.78	0.04	0.01	0.06	0.88	0.78	0.03	0.69	0.04	0.09	0.05	0.06*
Failed Once	0.21	0.03	0.21	0.02	0.00	0.04	0.9	0.17	0.04	0.18	0.04	-0.01	0.06	0.82	0.18	0.03	0.24	0.04	-0.07	0.05	0.15
Failed More than Once	0.11	0.03	0.05	0.01	0.06	0.03	0.05**	0.04	0.02	0.04	0.02	0.00	0.03	0.90	0.04	0.01	0.07	0.02	-0.03	0.03	0.28
Young (2019) RI P-Value							0.34							0.72							0.00**

Notes: $*p < 0.1, **p < 0.05$. P-values calculated using a standard two-sided t-test allowing for different variance across groups.
The "enrollment in 9th grade" treatment includes only those that participated in the second lottery.
The "participated in Lottery B" treatment includes only those that didn't enroll in the 8th grade, i.e., those that didn't win lottery A.
See text for details on Young (2019) Randomization Inference (RI) procedure.

(b) Interpretation of Balance Table and Multiple Hypothesis Testing

A few covariates presenting statistically significant differences between treated and controls **does not indicate that the lottery was not fair**, only that the sample of winners and losers turned out to be unbalanced. We could also have **multiple hypothesis testing problems**, which we discuss further in the next few paragraphs.

In this situation, if we were interest on the average treatment effect, a good possibility would just be to control for the covariates, which reduces sampling uncertainty. We could also try alternative forms of inference, such as permutation tests/randomized inference.

This last point is specially helpful in dealing with problems of multiple hypothesis testing. As we have 28 outcomes, the chance of rejecting the null of no average difference of at least one of them when the null is true is far larger than the actual size of 5%. If they are independent, we have a $1 - 0.95^{28} \approx 76\%$ chance of rejecting at least one null even if they are false!

To deal with this, we implement Young (2019)'s randomization inference (RI) process.[1] Let $\hat{\beta}$ be a $k \times 1$ vector containing the average difference in means of our covariates.

We conducted permutation tests, where our "treatments" were winning the first lottery, the second or deciding to participate in the second (the same order of the panels in Table 1). We do this in such a way that preserves the number of "treated" units in each permutation; more details will be given in *Question 3b*. For each permutation and population of interest (described in the footnotes of Table 1), we calculate and store the difference in means or proportions across the "treated" and "control" for each covariate. We did 250 permutations. At the end, we have three $250 \times k$ matrices.

The implementation was rather cumbersome. As some covariates are linear combinations of one another (such as male and females), this lead to issues when inverting the variance-covariance matrices. To deal with this, we drop a level of each characteristic.[2]

Even then, we still had problems in the cases of winners of the second lottery due to the variance-covariance matrix being *computationally singular*, with the reciprocal condition being too small for the

[1] See *Question 4a* for more details.

[2] We dropped `sex_male`, `race_indigenous`, `mother_schooling_less_5th`, `father_schooling_less_5th` and `ever_failed_more_than_once`. We also dropped the standardized 7th grade scores.

matrix to be invertible.[3] For this "treatment", we calculate quadratic forms using an identity matrix instead, which is suboptimal in terms of power relative to using the variance option.

We then calculate the probabilities that the quadratic forms in the permutations are bigger than the value of the statistic using the whole sample. These are displayed in the last row of Table 1. We only have evidence against the null that all outcomes are the same on whether or not the student participated in lottery B. Furthermore, this confirms the evidence that lotteries A and B are indeed fair, at least in terms of observable characteristics.

(c) Lottery B Participants

We have evidence on the contrary, mainly due to the 7th grade language and math test scores: **students who participated in lottery B have higher baseline scores** than those who did not (among those that lost the first lottery). This could be a reflex of diligence, as they persevere and want to enter the military school. Note that the most of the covariates show no difference between groups.

(d) Proposed Estimators of Effect of Enrollment in Both Grades

Let Z_i^A indicate winning the first lottery, L_i^B indicate participating in the second lottery and Z_i^B winning it. Note that our dataset only has information on students that participated in the first lottery. In addition to perfect compliance, we assume that those that enrolled in the 8th grade stay enrolled in the 9th and that the 9th grade lottery is the only way of enrollment for those that lost the first one.

The proposed estimand is

$$\mathbb{E}\left[Y_i^9|Z_i^A=1\right]-\mathbb{E}\left[Y_i^9|Z_i^A=0,Z_i^B=0\right]$$

The argument in ***(c)*** would suggest that those that didn't participate in the second lottery are less persevering or applied when compared to those that participated in it. Since $\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,Z_i^B=0\right]$ encompasses both those that didn't participate in lottery B as well as those that participated in lottery B, we have that

$$\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,Z_i^B=0,L_i^B=0\right]<\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,Z_i^B=0\right]<\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,Z_i^B=0,L_i^B=1\right]$$

$$\mathbb{E}\left[Y_i^9(0,0)|L_i^B=0\right]<\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,Z_i^B=0\right]<\mathbb{E}\left[Y_i^9(0,0)|L_i^B=1\right],$$

since $Z_i^B=0$ if $L_i^B=0$ and $\left(\left(Z_i^A,Z_i^B\right)\perp Y(a,b)\right)|L_i^B=1$. By the LIE, this implies

$$\mathbb{E}\left[Y_i^9(0,0)|L_i^B=0\right]<\mathbb{E}\left[Y_i^9(0,0)\right]<\mathbb{E}\left[Y_i^9(0,0)|L_i^B=1\right],$$

Thus, the proposed estimand in general **does not recover the target parameter** $\tilde{\beta}:=\mathbb{E}[Y_i^9(1,1)-Y_i^9(0,0)]$. To see that, consider an extreme case where everyone with $L_i^B=1$ is admitted to the 9th grade through the lottery. Then,

$$\begin{aligned}\mathbb{E}\left[Y_i^9|Z_i^A=0,Z_i^B=0\right]&=\mathbb{E}\left[Y_i^9|Z_i^A=0,L_i^B=0\right]\\&=\mathbb{E}\left[Y_i^9(0,0)|Z_i^A=0,L_i^B=0\right]\\&=\mathbb{E}\left[Y_i^9(0,0)|L_i^B=0\right]\\&<\mathbb{E}\left[Y_i^9(0,0)\right],\end{aligned}$$

where we used that $Z_i^A\perp Y(a,b)$. Therefore, the estimand would be lower than the target parameter.

(e) Observables and Unobservables

If we didn't found significant differences in the covariates, it still wouldn't be a good idea to use the comparison described in previous item, as we may have **unobservable characteristics** that are only imperfectly related to observables (such as ability, persistence, and so on) and that are related to potential outcomes, which would make the difference of means biased for the target parameter.

[3]Note that this is a different error than the matrix being actually singular, which happened due to multicolinear covariates and was taken care of when this error happened. We speculate that this is due to some covariate being unaccounted for in the permutation, in the sense that in only a small portion of samples it had representatives in both the treated and control groups. We excluded `race_indigenous` to prevent this, but it was not enough.

Question 3 – Alternative Inference Procedures

Table 2: Alternative Inference Results

	Without Controls	With Controls
Estimate of Average Treatment Effect	0.402	0.393
Robust S.E.	0.150	0.137
P-Value	0.007	0.004
P-Value: Permutation Test with $\lvert\hat{\beta}\rvert$	0.008	0.003
P-Value: Permutation Test with $\frac{\lvert\hat{\beta}\rvert}{\hat{\sigma}_{\hat{\beta}}}$	0.009	0.004
P-Value: Wild Bootstrap with Null Imposed	0.010	0.004

The results for items ***(a)-(d)*** are presented in Table 2 (sorry, did two tables).

In all regressions, the absolute value of the 8th grade math scores were used in order to retain interpretability. Since we are not directly comparing scores of different years, this comes at no loss relative to using the standardized variables we created in Question 1.

The regression was estimated using robust standard errors (HC3), as our target parameter is the ATE of going to that specific military school. That is, we are estimating the causal effect given any potential heterogeneous treatment effects of that school, a distinction we explore further in Question 6.

The controls include extensive dummies for all levels – that is, with only one base group not include – of sex, race, mother and father schooling and if the student ever failed school. We also include 7th grade math and language test scores, which are also not standardized. This also comes at no loss, since we are interested only in controlling for their information, and not on their coefficients.

(a) Estimators and Traditional Inference

We see that the estimates with the specification without controls are very similar to the one including them, which is to be expected given that the lottery is fair and that we have perfect compliance. Using a *ceteris paribus* interpretation, going to this specific military school causes math test scores to increase by about 0.4 points in the 8th grade.

The robust standard errors are also very similar, but, as we expect, the one with controls is smaller. This is because the inclusion of controls reduces sampling uncertainty, which may cause imbalance in some covariates if not controlled for. In both cases, the p-values are well below 1%.

(b) Inference with Permutation Tests/Randomized Inference

We have 483 observations, so doing exhaustive permutations would be close to impossible – the number of possible combinations is $\binom{483}{150}$, as we have 150 students who won the first lottery. For this reason, we settle for 1000 alternative treatment allocations.

To assign the "placebo" treatments, we randomly shuffle the indexes of the data and, based on this shuffled list, allocate treatment to the first 150 indexes. This way, the number of treated units is constant across permutations. We don't check if an assignment has already been made, as the probability we have the exact same assignment on different draws is almost zero.

P-values are calculated as the proportion of test statistics ($|\hat{\beta}|$ or $\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}}$) that are greater than the ones calculated using the original allocation. We see in Table 2 that the p-values calculated using both test statistics are very similar to the ones using the traditional inference method, which looks at the asymptotic distribution of the t-statistic that, under the null of $\beta = 0$, will be a standard normal.

For visual evidence, we plot the distribution of $|\hat{\beta}|$ and $\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}}$ across the different permutations in Figures 1 and 2, respectively, which allows us to see that indeed the result with the treatment original allocation is far too high for it to be purely a result of randomness.

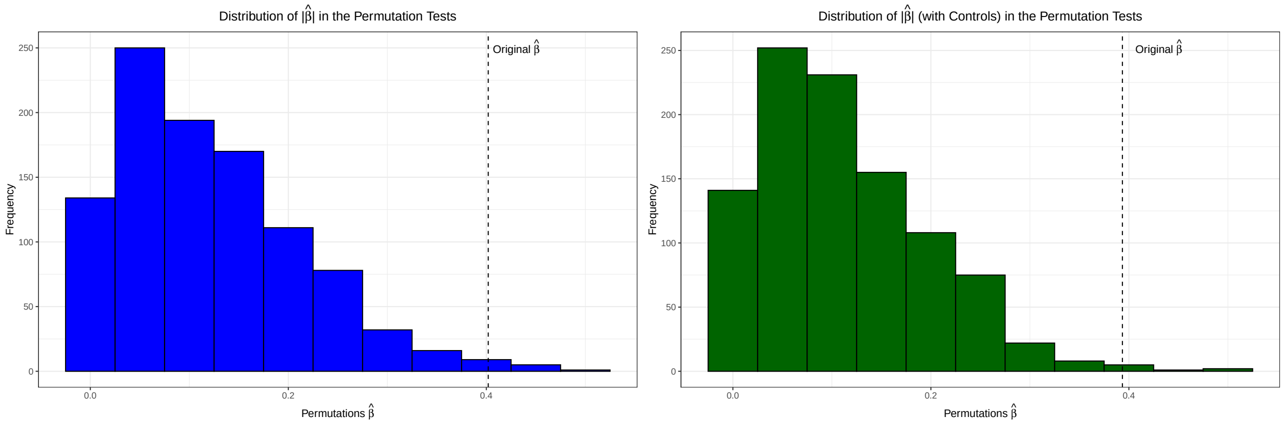

Figure 1: Permutation Distributions of $|\hat{\beta}|$

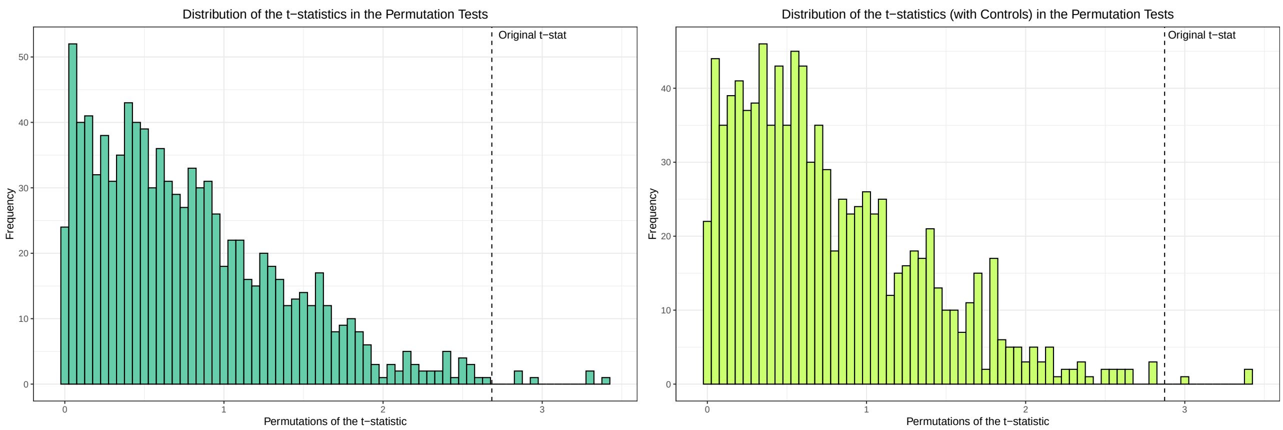

Figure 2: Permutation Distributions of $\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}}$

(c) Test Statistics

In the bivariate OLS regressions, $|\hat{\beta}|$ is simply the difference-in-means of the outcome between treated and controls. Given that we are interested in the null $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, this is asymptotically equivalent to using a t-test with non-robust standard errors, as we saw in class.

Thus, if we have heterogeneous treatment effects – or, in general, if treatment affects the dispersion of the outcome –, we violate the homoskedasticity assumption behind this standard error. Recall that we can *approximate* the distribution of $\hat{\beta}$ by

$$\hat{\beta} \sim N\left(\beta, \frac{\text{Var}(Y(1))}{N_T} + \frac{\text{Var}(Y(0))}{N_C}\right)$$

If the treatment increases the outcome's variability, we are likely to underestimate $\hat{\beta}$'s standard errors, which leads to overrejection.

On the other hand, the studentized statistic with robust standard errors on each permutation ($\frac{|\hat{\beta}|}{\hat{\sigma}_{\hat{\beta}}}$) is asymptotically equivalent to using a t-test with robust standard errors. Under the previous situation, this is less likely to lead to overrejection in finite samples under $H_0 : \beta = 0$ (and is a valid test asymptotically). Thus, the t-statistic is preferred when using permutations tests.

(d) Inference with Wild Bootstrap with Null Imposed

For the wild bootstrap with null imposed procedure, we first estimate the specifications with and without controls imposing that the average treatment effect is zero, storing the generated residuals $\hat{\varepsilon}_i^R$ and the coefficients $\hat{\beta}^R$. Let $x_i'^R$ be the appropriate vector of covariates without the treatment variable.

We did 1000 bootstrap samples. In each of them, for each observation i, we randomly allocate the residual of observation j_i, where j_i was determined by sampling from a Uniform$(1, N)$ and rounding to the nearest integer. This ensures that sampling of residuals is done with replacement.

Then, for each observation, we sample from a Uniform$(0, 1)$ to determine if the residual will be multiplied by -1 or not: if the sampled number is less than .5, we multiply it by -1 and allocate the residual. If the sampled number is larger than .5, we allocate the residual $\hat{\varepsilon}_{j_i}$ as is.

We denote the result of this process by $g_i \hat{\varepsilon}^R_{j_i}$, with $g_i \in \{-1, 1\}$ with equal probability. Note that the randomization of j_i and g_i was done separately for the specification with and without controls, so as to not create any dependence between the results.

We then create the DGP $y_i^B = x_i'^R \hat{\beta}^R + g_i \hat{\varepsilon}^R_{j_i}$ for both specifications using the appropriate variables. Finally, we run the regression of y_i^B on the treatment indicator (and controls, if applicable) and store the estimated coefficients for the ATE. P-values were calculated as the proportion of estimates whose absolute value is bigger than the absolute obtained from the original sample.[4]

The result of this process gives similar p-values to the traditional inference method and the permutation test, as seen in the last row of Table 2. We plot the distribution of the ATEs of the wild bootstrap with null imposed in Figure 3.

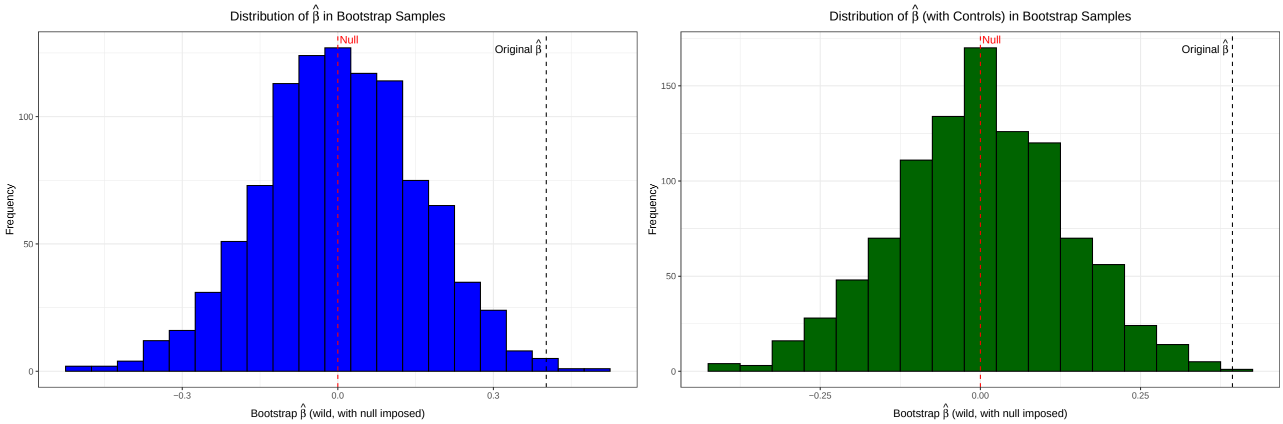

Figure 3: Wild Bootstrap Distributions

(e) Estimator and Bootstrap Distribution of Alternative Parameter

In *Question 7c* of *Problem Set 01*, we showed that the parameter $\tilde{\beta} := \mathbb{E}\left[Y_i^9(1,1) - Y_i^9(0,0)\right]$ (with Y_i^9 being absolute 9th grade math test scores) for those that participated in the first lottery is identified by the estimand

$$\mathbb{E}\left[Y_i^9 | Z_i^A = 1\right] - \left\{\mathbb{P}(L_i^B = 0 | Z_i^A = 0) \cdot \mathbb{E}[Y_i^9 | Z_i^A = 0, L_i^B = 0] + \mathbb{P}(L_i^B = 1 | Z_i^A = 0) \cdot \mathbb{E}[Y_i^9 | Z_i^A = 0, L_i^B = 1, Z_i^B = 0]\right\},$$

which can be estimated using the plug-in principle:

$$\begin{aligned}
\hat{\tilde{\beta}} = & \frac{1}{N_1} \sum_{i=1}^{N} Y_i \cdot \mathbb{1}\left\{Z_i^A = 1\right\} \\
& - \left(\frac{N_2}{N - N_1}\right) \frac{1}{N_2} \sum_{i=1}^{N} Y_i \cdot \mathbb{1}\left\{Z_i^A = 0, L_i^B = 0\right\} \\
& - \left(1 - \frac{N_2}{N - N_1}\right) \frac{1}{N_3} \sum_{i=1}^{N} Y_i \cdot \mathbb{1}\left\{Z_i^A = 0, L_i^B = 1, Z_i^B = 0\right\},
\end{aligned}$$

where $N_1 = \sum_{i=1}^{N} \mathbb{1}\left\{Z_i^A = 1\right\}$, $N_2 = \sum_{i=1}^{N} \mathbb{1}\left\{Z_i^A = 0, L_i^B = 0\right\}$ and $N_3 = \sum_{i=1}^{N} \mathbb{1}\left\{Z_i^A = 0, L_i^B = 1, Z_i^B = 0\right\}$.

[4] That is, our test statistic is $|\hat{\beta}^R|$. Note that we are only able to calculate p-values because we are imposing the null on the wild bootstrap procedure.

We implement this estimator using our data. To estimate its distribution, we use a standard non-parametric bootstrap. We again use 1000 bootstrap samples.

Similarly to ***(d)***, our bootstrap sample was constructed by sampling indexes from a Uniform$(1, N)$, which allows for replacement. That is, for each unit i^B in our bootstrap sample, we sample the index i of the original sample from (the integer of) the above uniform, which allows for replacement. We then construct the estimator based on the bootstrap sample and store its value.

The results are in Table 3, where τ_q is the value of the q-th quantile of the bootstrap distribution of $\hat{\tilde{\beta}}$. The distribution of the estimated average treatment effects can be seen in Figure 4.

The 95% confidence interval constructed using the bootstrap distribution of the estimator does not include 0. Therefore, we have evidence that there is a significant effect on average math test scores of enrolling in the military schools in both grades, relative to not being enrolled in any grade.

Table 3: Statistics of the Estimator of $\tilde{\beta} \coloneqq \mathbb{E}\left[Y_i^9(1,1) - Y_i^9(0,0)\right]$

$\hat{\tilde{\beta}}$	Standard Error	$\tau_{2.5\%}$	$\tau_{97.5\%}$
0.470	0.217	0.057	0.898

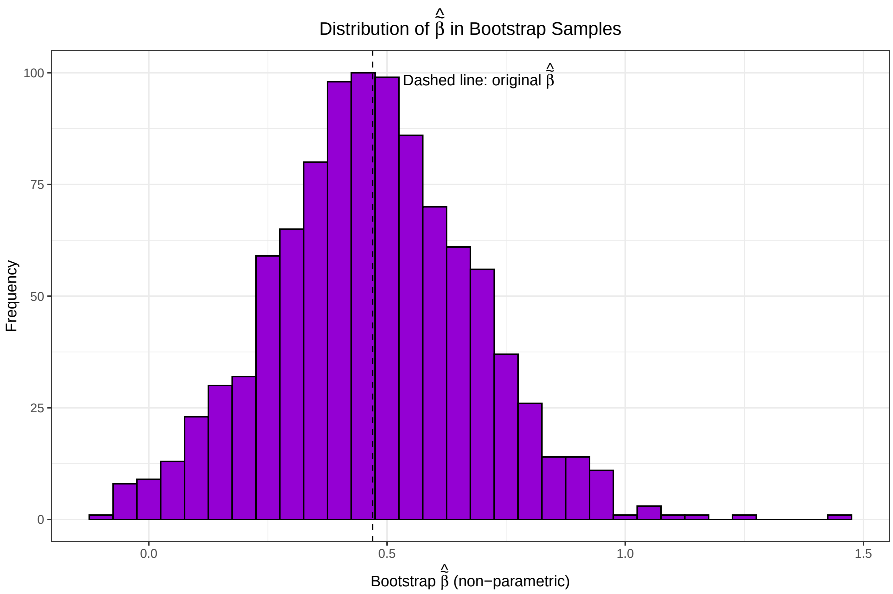

Figure 4: Bootstrap Distribution of $\hat{\tilde{\beta}}$

If we were to assume that $\hat{\tilde{\beta}}$ is asymptotically normal, its t-stat would be 2.166. This gives a p-value of about 0.03, and so we also reject the null of no average treatment effect using this approach. We note that, since we are not imposing the null, we are not able to calculate p-values directly using the bootstrap data.

Question 4 – Multiple Outcomes

(a) Multiple Outcomes and Possible Procedures

Suppose now that we are in interest in the ATE of enrolling in 8th grade in both math and language scores in the 8th grade. We show the estimated coefficients of the regressions in Table 4, which uses the absolute scores. Table 5 shows the results using the standardized scores as a robustness check.

Table 4: ATE Estimates for Math and Language Scores – 8th Grade

	Math - Without Controls	Math - With Controls	Language - Without Controls	Language - With Controls
Enrollment in 8th Grade	0.402	0.393	0.211	0.222
Standard Errors	[0.150]	[0.137]	[0.197]	[0.184]
P-Values	0.008	0.004	0.286	0.228
Observations	483	465	483	465
R^2	0.014	0.296	0.002	0.247

HAC robust standard errors used (HC3). Dependent variable is the absolute value of the test scores. Controls are the same as those of *Question 3*.

Table 5: ATE Estimates for Math and Language Standardized Scores – 8th Grade

	Math - Without Controls	Math - With Controls	Language - Without Controls	Language - With Controls
Enrollment in 8th Grade	0.257	0.251	0.110	0.116
Standard Errors	[0.096]	[0.088]	[0.103]	[0.096]
P-Values	0.008	0.004	0.286	0.228
Observations	483	465	483	465
R^2	0.014	0.296	0.002	0.247

HAC robust standard errors used (HC3). Dependent variable is the standardized value of the test scores.
Controls are the same as those of *Question 3*, but now using the standardized 7th grade scores.

We saw in *Question 3* that the treatment has a significant impact on the average math score. Now, we see that it doesn't quite have the same impact on the average of language scores, with both specifications finding insignificant results. Furthermore, we see that results are the same when using the standardized variables – with a change only in the magnitude of the coefficients –, and so we continue the analysis using only the absolute scores from now on.

Since we have multiple outcomes, we can have problems associated with multiple hypothesis testing, specially if we are interested in if enrollment affects *at least one* of them. We have a few alternatives to account for that.

The first one was explored in *Question 2b*: Young (2019)'s RI process, which uses permutations tests on all of the outcomes and tests them all using quadratic forms. Note that the null in this test is that all outcomes are the same across all units.

To be more specific, let $\hat{\beta} = (\hat{\beta}_{\text{math}}, \hat{\beta}_{\text{language}})'$ be the 2×1 vector of estimated coefficients for each specification (with and without controls). Denote by $\hat{\beta}(T_E)$ the one obtained in the original treatment assignment, which is in Table 4.

Let $\hat{\beta}(T_n)$ be the vector of coefficients estimated under permutation T_n. Denoting the variance-covariance matrix of $\hat{\beta}(T_n)$ by $\text{Var}\left(\hat{\beta}(T_n)\right)$. The p-value of the test is the probability that

$$\hat{\beta}(T_n)' \left[\text{Var}\left(\hat{\beta}(T_n)\right)\right]^{-1} \hat{\beta}(T_n) > \hat{\beta}(T_E)' \left[\text{Var}\left(\hat{\beta}(T_E)\right)\right]^{-1} \hat{\beta}(T_E)$$

In practice, we use the same variance matrix for $\hat{\beta}(T_n)$ and $\hat{\beta}(T_E)$, which is estimated by the variance-covariance matrix of the coefficients in the permutation test. In fact, we can use any matrix, but $\left[\text{Var}\left(\hat{\beta}(T_n)\right)\right]^{-1}$ maximizes power.

We can also use Anderson (2008)'s procedure, which tests the overall effect of treatment on a family of outcomes.

Assuming that a positive treatment effect indicates a better outcome – as is this case here –, we can standardize each outcome using the control group mean and standard deviation to make them all on the same scale. As in *Question 1*, we don't use treated statistics because they can be affected by treatment (or the expectation of it). Let $\tilde{Y}_i$ be the vector of transformed outcomes for unit i.

We then create a summary measure s_i by doing a weighted average of $\tilde{Y}_i$. This makes it so that we are considering correlations between different outcomes. We then regress s_i on treatment status to test whether the ATE on all outcomes is different from zero.

Since we are using an average, we reduce the variance of outcome, which increases power of tests. However, we don't know what outcome is driving the (potentially significant) ATE. Furthermore, we lose interpretability of the coefficients, as we are using an average of standardized variables.

The final methods consist on adjusting the $m = 2$ individual p-values to control for the *family-wise-error rate* (FWER). We saw two methods for that.

The first was is Bonferroni's procedure. In it, we only reject H_s, $s \in \{1, ..., m\}$, if its p-value is such that $p_s < \frac{\alpha}{m}$, where α is the individual significance level. This controls for the worst case scenario, which makes the test underpowered.

The second was Holm's. Order the individual p-values from lowest to highest and let k be the minimal index such that

$$p_{(k)} > \frac{\alpha}{m + 1 - k}$$

Only the nulls $H_{(1)}, ..., H_{(k-1)}$ are rejected. Note that, if $k = 1$, none of the hypothesis are rejected in this "sequential" procedure, which has higher power when compared to Bonferroni's.

(b) Implementation of Two Solutions

Since we already used Young (2019) in *Question 2b*, we will used Anderson (2008) to test the overall effect of enrollment in test scores of both subjects. To account for FWER, we will test each outcome individually using Bonferroni's procedure.

Anderson (2008)'s procedure results is shown in Table 6, where we used the simple averages of standardized 8th grade math and language test scores. The coefficients on both specifications are the same, with p-values equal to or less than 1%, which tells us that the overall treatment effect on both outcomes is significantly different from zero. According to Tables 4 and 5, the results are driven by the effects on math test scores.

Table 6: ATE Estimates on All Outcomes – Anderson (2008)

	Without Controls	With Controls
Enrollment in 8th Grade	0.184	0.184
Standard Errors	[0.071]	[0.063]
P-Values	0.010	0.004
Observations	483	465
R^2	0.014	0.353

HAC robust standard errors used (HC3).

Dependent variable is the simple average of standardized 8th grade math and language test scores.

Controls are the same as those of *Question 3*.

Bonferroni's procedure is simpler, as we just have to multiply the p-values of Table 4 by two. We still reject the null hypothesis of $H_0 : \beta = 0$ on math test scores, with adjusted p-values of 0.016 and 0.008 on the specifications without and with controls, respectively. We are now further away from rejecting the null in the case of language scores, with adjusted p-values of 0.572 and 0.456.

Question 5 – Attrition

(a) Causal Effect of Enrolling in Military-School on Exam-Taking

Out of the 483 students who participated in the 8th grade lottery, 320 of them took the voluntary exam. Out of the 150 lottery winners, 109 took the exam. We then have that $\frac{109}{150} = 72.6\%$ of those enrolled in the military school took the exam, while $\frac{320-109}{483-150} = 63.3\%$ of those that didn't win the lottery chose to take the exam.

To estimate the causal effect of enrollment in the military school on the probability of taking the voluntary exam, we leverage the fact that the lottery is fair and that there is perfect compliance. This allows us to estimate this effect using a linear probability model.

Results are shown in Table 7. We see that enrollment increases the probability of taking the exam by 8 to 9%, although the result loses significance at the 5% when including controls. In broad terms, we have evidence of different attrition rates by treatment group.

Table 7: Enrollment in 8th Grade and Probability of Taking the Voluntary Exam

	Without Controls	With Controls
Enrollment in 8th Grade	0.093	0.082
Standard Errors	[0.045]	[0.048]
P-Values	0.040	0.088
Observations	483	465

HAC robust standard errors used (HC3).

Dependent variable is the indicator of taking the voluntary exam.

Controls are the same as those of *Question 3*.

(b) Balance Table for Non-Attriters

We define as attriters those that enrolled in the 8th grade and did not take the voluntary exam. Results of the balance tests are shown in Table 8.

The left panel includes only those that took the voluntary exam and checks whether there is balance across covariates of lottery winners and losers. We find some evidence of covariate imbalance, but this can be due to multiple hypothesis problems. Indeed, the Young (2019)'s procedure gives a p-value of 0.45, so we are not able to reject the null that all outcomes are equal for all units. This further confirms the fairness of the lottery.

The right panel includes all students and checks whether there is balance across covariates of those that did and didn't take the voluntary exam. We have strong evidence of imbalance: those that took the test have higher baseline scores and seem to have more educated parents. Young (2019)'s p-value is 0, indicating that these differences are not due to problems of multiple hypothesis testing.

(c) Causal Effect of Enrolling in Military-School on Exam Score

The results of the previous item are problematic for point identification of the effect of enrolling in the military school on voluntary test scores, mainly due to the results of the right panel of Table 8.

This is because attrition is not random: those who take the voluntary exam have higher baseline scores and more educated parents, and so are likely to fair better in this exam. Since treatment (winning the lottery) increases the probability of taking the exam, $Z_i \perp V_i$ does not hold and we are likely to have selection bias, even if the original lottery is fair.

Let V_i be an indicator variable on whether the student took the voluntary exam and Y_i^V the score on it. We expect that

$$\mathbb{E}\left[Y_i^V | V_i = 1\right] > \mathbb{E}\left[Y_i^V\right] \quad \therefore \quad \mathbb{E}\left[Y_i^V | Z_i, V_i = 1\right] > \mathbb{E}\left[Y_i^V | Z_i\right]$$

Table 8: Balance Test – Attrition

	"Treatment": Win Lottery A							"Treatment": Took Voluntary Exam						
	Treated		Control		Difference T - C			Treated		Control		Difference T - C		
	Mean	SE	Mean	SE	Diff.	SE	**p**	Mean	SE	Mean	SE	Diff.	SE	**p**
Language Score (7th Grade)	5.09	0.19	5.10	0.12	-0.02	0.23	0.94	5.10	0.10	4.74	0.14	0.36	0.18	0.05**
Math Score (7th Grade)	5.12	0.15	5.32	0.09	-0.20	0.17	0.26	5.25	0.08	4.68	0.12	0.58	0.14	0.00**
Sex: Male	0.39	0.05	0.41	0.03	-0.02	0.06	0.77	0.40	0.03	0.41	0.04	0.00	0.05	0.96
Sex: Female	0.61	0.05	0.59	0.03	0.02	0.06	0.77	0.60	0.03	0.59	0.04	0.00	0.05	0.96
Race: White	0.25	0.04	0.30	0.03	-0.04	0.05	0.4	0.28	0.03	0.22	0.03	0.06	0.04	0.12
Race: Brown	0.47	0.05	0.52	0.03	-0.05	0.06	0.4	0.50	0.03	0.57	0.04	-0.06	0.05	0.21
Race: Black	0.22	0.04	0.13	0.02	0.09	0.05	0.06*	0.16	0.02	0.19	0.03	-0.03	0.04	0.44
Race: Yellow	0.02	0.01	0.03	0.01	-0.01	0.02	0.41	0.03	0.01	0.02	0.01	0.01	0.01	0.49
Race: Indigenous	0.04	0.02	0.01	0.01	0.02	0.02	0.26	0.02	0.01	0.01	0.01	0.02	0.01	0.13
Mother's Education: <5th Grade	0.08	0.03	0.08	0.02	0.00	0.03	0.93	0.08	0.02	0.11	0.02	-0.02	0.03	0.41
Mother's Education: 5th-9th Grade	0.18	0.04	0.22	0.03	-0.04	0.05	0.42	0.20	0.02	0.27	0.04	-0.07	0.04	0.12
Mother's Education: 9th Grade-HS	0.15	0.03	0.11	0.02	0.04	0.04	0.34	0.12	0.02	0.13	0.03	-0.01	0.03	0.82
Mother's Education: HS-College	0.31	0.05	0.34	0.03	-0.03	0.06	0.57	0.33	0.03	0.27	0.04	0.06	0.04	0.16
Mother's Education: College	0.14	0.03	0.10	0.02	0.04	0.04	0.32	0.12	0.02	0.06	0.02	0.05	0.03	0.05**
Mother's Education: >College	0.08	0.03	0.06	0.02	0.01	0.03	0.68	0.07	0.01	0.05	0.02	0.02	0.02	0.45
Mother's Education: Unknown	0.06	0.02	0.08	0.02	-0.03	0.03	0.39	0.07	0.01	0.11	0.02	-0.03	0.03	0.24
Father's Education: <5th Grade	0.15	0.03	0.08	0.02	0.07	0.04	0.09*	0.11	0.02	0.12	0.03	-0.01	0.03	0.64
Father's Education: 5th-9th Grade	0.22	0.04	0.22	0.03	0.00	0.05	0.99	0.22	0.02	0.26	0.03	-0.04	0.04	0.32
Father's Education: 9th Grade-HS	0.12	0.03	0.10	0.02	0.03	0.04	0.49	0.11	0.02	0.13	0.03	-0.03	0.03	0.39
Father's Education: HS-College	0.27	0.04	0.29	0.03	-0.02	0.05	0.70	0.29	0.03	0.21	0.03	0.08	0.04	0.06*
Father's Education: College	0.05	0.02	0.08	0.02	-0.03	0.03	0.28	0.07	0.01	0.06	0.02	0.01	0.02	0.66
Father's Education: >College	0.02	0.01	0.01	0.01	0.01	0.01	0.54	0.01	0.01	0.01	0.01	0.01	0.01	0.47
Father's Education: Unknown	0.17	0.04	0.22	0.03	-0.05	0.05	0.26	0.20	0.02	0.22	0.03	-0.01	0.04	0.79
Never Failed	0.70	0.05	0.76	0.03	-0.07	0.05	0.23	0.74	0.02	0.69	0.04	0.05	0.04	0.23
Failed Once	0.20	0.04	0.19	0.03	0.01	0.05	0.76	0.19	0.02	0.25	0.03	-0.05	0.04	0.18
Failed More than Once	0.10	0.03	0.05	0.02	0.05	0.03	0.13	0.07	0.01	0.07	0.02	0.00	0.02	0.94
Young (2019) RI P-Value							0.45							0.00**

Notes: $^{*}p < 0.1, ^{**}p < 0.05$. P-values calculated using a standard two-sided t-test allowing for different variance across groups.
The "win lottery A" treatment includes only those that took the voluntary exam.
Includes only those that took the voluntary exam. See *Question 2b* for details on Young (2019) Randomization Inference (RI) procedure.

And so the comparison of winners and losers of the first lottery might be invalid, as we are estimating $\mathbb{E}\left[Y_i^V | Z_i^A = 1, V_i = 1\right] - \mathbb{E}\left[Y_i^V | Z_i^A = 0, V_i = 1\right]$ instead of $\mathbb{E}\left[Y_i^V | Z_i^A = 1\right] - \mathbb{E}\left[Y_i^V | Z_i^A = 0\right]$, which identifies the target parameter $\mathbb{E}\left[Y_i^V(1)\right] - \mathbb{E}\left[Y_i^V(0)\right]$.

The left panel of Table 8 suggests that, for the students who took the exam, there is not much of a difference between lottery winners and losers. The direction of the bias is thus unclear at first, depending on if

$$\mathbb{E}\left[Y_i^V | Z_i^A = 1, V_i = 1\right] - \mathbb{E}\left[Y_i^V | Z_i^A = 1\right] \lesseqgtr \mathbb{E}\left[Y_i^V | Z_i^A = 0, V_i = 1\right] - \mathbb{E}\left[Y_i^V | Z_i^A = 0\right]$$

We can think that, as enrollment increases the probability of exam taking and those that it are expected to better students, going to the military school increases the difference between $\mathbb{E}\left[Y_i^V | V_i = 1\right]$ and $\mathbb{E}\left[Y_i^V\right]$, as there may be exam-specific training or other forms of gaming. This would lead to an upward bias in the naive comparison between lottery winners and losers.

If, on the other hand, the school incentivizes students that otherwise wouldn't take the exam due to not being interested, we may have that the comparison be biased downwards.

(d) Possible Solutions and Assumptions

To solve this attrition issue, we could impose strong assumptions on how selection is occurring. For example, we could impose that attrition is random across treatment groups, or use some selection model such as Heckman's.

A preferred alternative is to go on a partial identification rout, using bounds for treatment effects such as Manski's or Lee's. A common assumption between these methods is that treatment be randomly assigned in the original sample, which our fair lottery satisfies.

Manski's bounds only impose that the outcome variable Y_i^V be bounded – which exam scores are –, estimating the bounds based on the worse and best case scenario of values for attritors. This can lead to some uninformative bounds, which lose precision the bigger the range of the variable of interest.

Lee bounds[5] can provide sharper bounds, but rely on stronger monotonicity assumptions. That is, we need that the latent attrition status be such that

$$V_i(Z_i^A) : \mathbb{P}(V_i(1) \geq V_i(0)) = 1$$

In words, treatment can only increase the probability of remaining in the sample (no one is observed on the final sample only when assigned to the control group). That is, every control student who took the voluntary exam would also have taken it had they enrolled in the military school, while there are some treated students that only take the exam because they were enrolled.

We get some confirmation of this assumption on item *(a)*, where we see that enrollment increases the probability of taking the exam. This is an average, however, and not a result that holds for all units. Although not confirming the validity of the assumption, this result gives some reassurance of it holding.

(e) Implementation of Solution

We will implement Lee Bounds using the `leebounds` R package and homonimous function. For comparison, we also estimate OLS regressions of exam score on enrollment in 8th grade. All results are shown in Table 9.

Table 9: Enrollment in 8th Grade and Score on Voluntary Exam

	Without Controls	With Controls	Lee Bounds
Enrollment in 8th Grade	0.284	0.313	[-0.094, 0.750]
Standard Errors	[0.210]	[0.200]	
P-Values	0.177	0.120	
Observations	483	465	483

HAC robust standard errors used (HC3).

Dependent variable is the indicator of taking the voluntary exam.

Controls are the same as those of *Question 3*.

Despite increasing the probability of exam taking, it seems that enrollment in the 8th grade does not lead to better scores on the voluntary exam. OLS results – which are likely to be biased upwards if we believe the first story of item *(c)* – are insignificant even at the 10% level. Lee Bounds don't exclude the possibility of null treatment effect.[6]

[5] This blog post from World Bank has a great discussion. Lee's original paper can be found here.

[6] Worth noting that the package does not allow for covariates.

Question 6 – External Validity

(a) Robust Standard Error, Target Parameters and Unobserved Shocks

When using robust standard errors, we are estimating the average treatment effect *conditional on* the realization of school-specific shocks or heterogeneous treatment effects, such as the existence of a large construction site nearby.

This is because the robust standard errors assume independence between the errors of different observations, and thus do not capture uncertainty related to school-specific shocks, only due to sampling. In this specific case, the construction is likely to be detrimental to students' learning, and so the target parameter we would be estimating would be lower than the average treatment effect for "a" generic school.

(b) J Military Schools and Robust vs. Clustered Standard Errors

If we have J military schools with lotteries, the choice of the standard error implicitly changes the parameter our estimators recover.

As said in the previous item, the use of robust standard errors makes it so that our OLS estimator recovers the causal effect[7] *conditional on* the realization of any heterogeneous treatment effect at the school level. Therefore, if we are interested in the ATE for a specific school, we should use robust SEs. This increases the internal validity of the experiment.

On the other hand, if we are interested in the ATE *netted out* of school-level shocks, we should cluster the standard errors. This is because they take into account any within-school dependence between students. Therefore, if we are interesting in scaling up the number of military schools, so that we want the ATE for a generic school, we should cluster our standard errors. This increases the external validity of the experiment.

When $J \to \infty$, the recovered parameters are the same, as school-level heterogeneous effects cancel each other out with probability 1.

(c) Few Clusters

If J was very small, the two parameters discussed are more likely to be different. Furthermore, with few clusters, the cluster robust standard errors are likely to be biased downwards, and so we should correct for degrees of freedom.

If, as is the case of most of this Problem Set, we have $J = 1$, we can't cluster our standard errors, and so can only recover the ATE conditional on the realization of the specific school shock. This is because, by the OLS FOC,

$$\frac{1}{NJ}\sum_j \sum_i \hat{\varepsilon}_{ij} = \frac{1}{N}\sum_i \hat{\varepsilon}_{i1} = 0$$

where we assume N students at the school and use that $J = 1$. This implies that the plug-in estimator for the cluster-robust standard error is identically zero:

$$\frac{1}{J}\sum_j \left(\frac{1}{N}\sum_i \hat{\varepsilon}_{ij}\right)^2 = \left(\frac{1}{N}\sum_i \hat{\varepsilon}_{i1}\right)^2 = 0^2 = 0$$

[7] The OLS estimand identifies the causal effect due to lotteries being fair and because of perfect compliance.