Ec140 - Regression as Matching (Part II)

Fernando Hoces la Guardia
07/13/2022

Real Life Example: Regression and Causal Effects of Private College

- Dale and Krueger (2002) analyze data from college applications, admissions and final choice for individuals that apply

- The key idea of the paper is that instead of measuring all characteristics where treatment and control will differ, they argue that they have a measure that closely summarizes all those unobserved characteristics: college application and college decisions.

- Supposedly application information is a good proxy for motivation, and acceptance is a good proxy of capacity. In my view, this could have been a good argument 20 years ago, but not today (Harvard's Legacy+Athlete bonus, college admissions scandal, additional evidence). For the purpose of the example let's assume that these are good proxies for all other things.

Intuition Behind Control Strategy

TABLE 2.1
The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Intuition Behind Control Strategy: Notes 1/2

- Grouped by application and admission decision at the university level.
- Within a group there can be variation in final decisions.
- Within group variation for group A is negative (-5k). Group B has a positive difference (30k). There are many combinations of such university-application-decisions-groups.
- Group C and D have all private and all public respectively, so nothing to learn here in terms of private-public diffs (all treatment or all control).

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Intuition Behind Control Strategy: Notes 2/2

- Simple average (of within group differences) is a good estimate of causal effects (given our assumptions): $12,500, also another good estimate is the weighted average: 9,000. Giving more weight to more data makes more efficient use of information, leading to a more precise estimate.
- Comparing within groups we can argue that we are holding Y_0 (potential earnings if no treatment) constant.
- Simple group difference would estimate 19.5K (all) or 20K (just A and B) diff.
- Selection bias emerges when comparing across, instead of within, groups. Group A was

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Ready to Understand Regressions! 1/3

- Think of regression as an automated matcher: regression estimates are weighted averages of multiple matched comparisons (similar to groups A and B before).
- Regression ingredients. Right hand side (LHS):
  - Dependent variable, or outcome variable. In our example: earnings in 20 years after graduation.
    RHS:
  - Treatment variable, in our case, a binary variable indicating 1 for private and 0 for public.
  - A set of control variables, in our example variables that identify sets of schools to which students apply and were admitted too.
- Observations: C&D are excluded from our sample because they do not provide information regarding the relevant comparison we want to make.

Ready to Understand Regressions! 2/3

Regression equation:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

- All RHS variables are called regressors, explanatory or independent variables. The difference between A and P is conceptual, not formal. The research design justifies the role each variable plays. In our case, P plays a primary role, while A is secondary (not interested if it's actually measuring a causal relationship).
- Intercept/constant, α
- Causal effect of treatment β, and
- The effect of being a group A student, γ. (not relevant to us)
- The residual, e_i, defined as the difference between observed (Y_i) and fitted values ($\hat{Y}_i$). We will focus on this in *Regression as Line Fitting*.

Ready to Understand Regressions! 3/3

- What regression does: chooses α, β and γ, to minimize the sum of squared residuals. Executing this minimization is often called “Estimating” or “Running” a regression. We will explore a little of theory, and how to run regressions in a little. But first, let’s focus on the result of running a regression.

- Simple toy example (from table 2.1): β of 10,000 shows that the regression estimate is somewhere in between the simple group comparison (12.5k) and weighted group comparison (9K).

From Toy Example to Data

- Group by Barron's selectivity group-application-decisions instead of university-application-decisions-groups to increase sample size.

	Private			Public		
i	Ivy	Leafy	U3	All State	Tall State	Altered State
1		R	A		A	
2	R		A	A		
i	MC		HC	C	HC	
1	R		A	A		
2	R		A	A		

From Toy to Actual Regression

The simplified regression:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

Is operationalized in practice with:

$$lnY_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT + \delta_2 lnPI_i + e_i$$

Differences:

- lnY_i (not Y_i) $\Rightarrow \Delta\%$ interpretation
- 150 groups ($GROUP_{ji}$) instead of 1 (A_i)
- Additional controls: SAT, (Ln) Parental Income, plus others (not shown)
- Much closer to Other Things Equal!

First Read of Regressions Results! 1/5

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

First Read of Regressions Results! 1/5

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

First Read of Regressions Results! 1/5

- Focus on controls that appear in equation
- There are 6 regressions here
- Read from left to right (column 1 - 6)
- Each row contains estimates for the population parameters (α, β, δ). This estimates are usually refereed as $(\hat{\alpha}, \hat{\beta}, \hat{\delta})$, but following the book's

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a

First Read of Regressions Results! 2/5

- Column 1 represents a regression with only a constant and the treatment indicator: $lnY_i = \alpha + \beta P_i + e_i$
- In a regression with only one binary regressor on the RHS, its coefficient is the simple difference in groups between treatment and control $(\overline{Y_1} - \overline{Y_1})$
- This difference is close to 14% (0.135).
- Small SE suggests that this result is statistically different from zero.

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

First Read of Regressions Results! 3/5

- Column 2 represents the following regression: $lnY_i = \alpha + \beta P_i + \delta_1 SAT_i + e_i$.
- The control SAT is divided by 100, hence the coefficient, which represents an increment in one unit, represents the (percent) increase in earnings **associated** with an increase of 100 points in the SAT.

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

First Read of Regressions Results! 4/5

- The value of 0.048, means that additional 100 pts in the SAT are **associated** with an increase of 5% in earnings 20 years in the future. Also statistically significant.
- More important: the (apparent) causal effect of private school fell to 10% (0.95) after controlling for SAT.
- Column 3 expands on this approach by adding more observables to the regression. The effect of private drops to 9% (0.86).

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

First Read of Regressions Results! 5/5

- Now add the selection controls (move to cols 4-6). Column 6 represents the regression specified in slide 10.
- Effect of private school goes to zero (0.007 - 0.013).
- Effect of adding more control is now irrelevant.
- This suggests that the "selectivity controls" are measuring a significant amount of information for observables and unobservables.

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

Second Read of Regressions Results

- Now let's repeat the exercise but with a different measure of selectivity: average SAT in schools that applied to, and binaries for number of schools applied to.
- This gives us the full sample from C&B (before we only had 5,583)
- A similar pattern emerges: controlling for observables diminishes the effect, but it remains substantial (in economic terms).; adding "selectivity controls" drops the effect to zero.

TABLE 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Third Read of Regressions Results 1/2

- Finally, what if private/public school selectivity is not the right treatment to analyze? What if its how much "better" your classmates are (at taking the SAT)
- A similar story seems to emerge: some effect when looking at simple differences or controlling by some observable characteritics.
- But effect goes away when controlling for the SAT selectivity proxy

TABLE 2.4
School selectivity effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109	.071	.076	−.021	−.031	.000
	(.026)	(.025)	(.016)	(.026)	(.026)	(.018)
Own SAT score ÷ 100		.049	.018		.037	.009
		(.007)	(.006)		(.006)	(.006)
Log parental income			.187			.161
			(.024)			(.025)
Average SAT score of schools applied to ÷ 100				.138	.116	.089
				(.017)	(.015)	(.013)
Sent two applications				.082	.075	.063
				(.015)	(.014)	(.011)
Sent three applications				.107	.096	.074
				(.026)	(.024)	(.022)
Sent four or more applications				.153	.143	.106
				(.031)	(.030)	(.025)

Notes: This table reports estimates of the effect of alma mater selectivity on earnings. Each column shows coefficients from a regression of log earnings on the average SAT score at the institution attended and controls. The sample size is 14,238. Standard errors are reported in parentheses.

Third Read of Regressions Results 2/2

- This evidence seems less credible and should be treated with much more skepticism. The entire exercise was meant to justify using one or another set of controls to answer a specific policy question (effect private or public school on future earnings). Changing the policy question (to effect of selectivity of class mates, measured as average SAT, on earnings) and extrapolating the validity of the former exercise into the latter is a good example of overextending the validity of a research design.

TABLE 2.4
School selectivity effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109	.071	.076	−.021	−.031	.000
	(.026)	(.025)	(.016)	(.026)	(.026)	(.018)
Own SAT score ÷ 100		.049	.018		.037	.009
		(.007)	(.006)		(.006)	(.006)
Log parental income			.187			.161
			(.024)			(.025)
Average SAT score of schools applied to ÷ 100				.138	.116	.089
				(.017)	(.015)	(.013)
Sent two applications				.082	.075	.063
				(.015)	(.014)	(.011)
Sent three applications				.107	.096	.074
				(.026)	(.024)	(.022)
Sent four or more applications				.153	.143	.106
				(.031)	(.030)	(.025)

Notes: This table reports estimates of the effect of alma mater selectivity on earnings. Each column shows coefficients from a regression of log earnings on the average SAT score at the institution attended and controls. The sample size is 14,238. Standard errors

Wrapping up the Example: Why Regression is Great

Four reasons:

- Clear conceptually interpretation: as difference in matched sub-groups.
- Good benchmark to compare against other methods.
- Under specific circumstances, it's an unbiased the most efficient estimator we can use to measure the causal effect of the intervention (these “specific circumstances” used to take 2-4 classes to explained).
- Computationally feasible: tractable minimization problem (will discuss more next class about this).

Acknowledgments

- Ed Rubin's Undergraduate Econometrics II
- XQCD
- BITSS
- ScPoEconometrics
- XQCD
- MM
- Matt Hollian