MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 6

Understanding Multivariate Regression

1 Regression, Causality, and Control

Does MIT matter? The Dale and Krueger (DK; 2002) study on our reading list looks at difference in earnings between graduates of more and less selective colleges, as measured by average SAT scores of those enrolled. We turn this into a Bernoulli treatment (here and in MM, Chpt 2) by considering the effect of graduation from a private college or university (which are more selective and expensive than public schools, on average). Two of my former Ph.D. students were admitted to Harvard yet attended their local state (public) schools. Today, these students are professors in top econ departments - not bad! But perhaps they would have done better if they had attended (private) Harvard instead. Who knows, they might even have found jobs on Wall Street!

These are just two data points, of course. In larger and more representative samples, however, comparisons between private and state school graduates consistently show higher earnings for those who went private. No surprise! *Something* must justify the hundreds of thousands of tuition dollars private schools collect.

Yet, part of the difference in earnings between private and public college grads is surely attributable to differences in the characteristics ($Y_{0i}'s$) of people who did and didn't attend private schools. Variables likely to determine potential outcomes include students' own SAT scores (which are correlated with their earnings), the selectivity of schools they applied to (which says something about students' own judgements of their ability) and family income (which is also correlated with later earnings).

- We'd like to hold these things constant, that is, to control for them when comparing groups of students who went to different types of schools
- We hope this controlled (matched) comparison brings us one giant step closer to the average causal effect revealed by a hypothetical experiment that randomly assigns private attendance

1.1 The Payoff to Private College

The DK (2002) research design, as implemented in Chapter 2 of MM, compares public and private graduates who applied to and were admitted to schools of similar selectivity.

- Consider a hypothetical set of applicants, all of whom applied to one or more schools among three public (All State, Tall State, and Altered State) and three private (Ivy, Leafy, and Smart).
- The matching matrix these students face appears below:

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.

- Five of nine students (numbers 1,2,4,6,7) attended private schools. Average earnings in this group are $92,000. The other four, with average earnings of $72,500, went to a public school. The roughly$20,000 gap between these two groups suggests a large private school advantage.

The hypothesis motivating a DK-style analysis is that, conditional on the identity (or selectivity) of schools to which an applicant applies, and the identity (or selectivity) of schools to which an applicant has been admitted, comparisons of students who went to different schools (say, one to public and one to private) are likely to be "apples to apples." In other words, we uncover the effects of private school attendance by:

- Comparing students 1 and 2 with student 3 in group A and by comparing student 4 and student 5 in Group B
- Discarding students in groups C and D (why?)
- The average of the -5 thousand dollars gap for group A and the 30,000 gap dollars for group B is $12,500. This is a good estimate of the effect of private school attendance on average earnings because it controls (at least partially) for applicants' ambition and ability (a weighted average reflecting the fact that 3/5 of applicants are in Group A is $9,000)
- Note that overall average earnings in Group A are much higher than overall average earnings in group B. Our within-group matching estimates of 12,500 or 9,000 eliminate this source of selection bias in public-private comparisons

Instead of manually averaging these group-specific contrasts, regress!

- Limit the analysis to Groups A an B. With only one control variable needed, A_i (a dummy for those in Group A), the regression of interest can be written:

$$Y_i = \alpha + \beta P_i + \gamma A_i + \varepsilon_i \tag{1}$$

- The distinction between the causal variable, P_i, and the control variable, A_i, in equation (1) is conceptual, not formal: Stata can't tell the difference.

- Using data for the five students in Groups A and B generates $\beta = 10,000$ and $\gamma = 60,000$. The private school coefficient in this case is $10,000$, close to the estimate obtained by averaging the public-private contrasts within groups A and B and well below the raw public-private difference of almost 20,000.

Public-Private Face-Off

The *College and Beyond* (C&B) data set used in the DK study includes over 14,000 college graduates who attended one of 30 schools.

- Only about 2,000 graduates can be classified as belonging to groups of two or more who applied to and were accepted by the same schools – this is the sample for which we have an exact match.

The number of useful comparisons is increased by deeming schools to be "matched" if they are equally selective instead of insisting on exact matches.

- To fatten up the selectivity categories, call schools comparable if they fall into the same Barron's selectivity group

- 9,202 students have Barron's matches, that is, they can be put into groups of two or more who applied to and were accepted by sets of schools in the same Barron's categories.

- Because we're interested in public-private comparisons, our Barron's matched sample is also limited to matched applicant groups that contain both public and private school graduates. This leaves 5,583 matched applicants for analysis. These matched applicants fall into 151 different selectivity groups containing both public and private graduates.

The operational regression model for the Barron's selectivity-matched sample includes many control variables, while the stylized 6-student example controls only for the dummy variable A_i, indicating students in group A. The key controls in the operational model consist of many dummy variables indicating all Barron's matches represented in the sample (with one group left out as a reference category). These controls capture applicant ambition and ability as measured by the selectivity of the schools to which they applied and were admitted in the real world, where many combinations of schools are possible.

The resulting regression model looks like this:

$$\ln Y_i = \alpha + \beta P_i + \underbrace{\sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i}_{controls, X_i'\gamma} + \varepsilon_i$$

- The parameter β in this model is the coefficient of interest, an estimate of the causal effect of attendance at a private school. The controls, collected in $X_i'\gamma$, include Barron's selectivity group dummies as well as applicant ability (SAT_i) and family background ($\ln PI_i$)

- This model controls for 151 groups instead of the two groups in our stylized example. The parameters γ_j, for $j = 1$ to 150, are the coefficients on 150 selectivity-group dummies, denoted $GROUP_{ji}$

- The fully-controlled model includes a few more covariates we haven't bothered to write out. The table below reports key findings:

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.

- Selectivity controls kill the private collegge effect!
- Control at what cost? From an original N of around 14,000, here we're down to about 5,500

- Perhaps it's enough to control linearly for the average SAT scores of the schools to which I've applied, as well as the number of schools to which I've applied:

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)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
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.

- This buys us a larger sample and doesn't change the results much

- What about school selectivity effects as in DK, instead of a public/private comparison?

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)
Female			−.403			−.396
			(.015)			(.014)
Black			−.023			−.034
			(.035)			(.035)
Hispanic			.015			.006
			(.052)			(.053)
Asian			.173			.155
			(.036)			(.037)
Other/missing race			−.188			−.193
			(.119)			(.116)
High school top 10%			.061			.063
			(.018)			(.019)
High school rank missing			.001			−.009
			(.024)			(.022)
Athlete			.102			.094
			(.025)			(.024)
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.

- Pity my poor parents, whom I made even poorer by attending Oberlin, a pricey private college. It seems I could just as well have gone to Penn State!

2 When and How Controls Matter: OVB

2.1 One vs. Two

Suppose you'd like to regress log wages (Y_i) on years of schooling (S_i), controlling for ability (A_i):

$$Y_i = \alpha + \rho S_i + \gamma A_i + \varepsilon_i \tag{2}$$

You seek this regression in the hope that controlling for ability mitigates selection bias in estimates of the economic returns to schooling. Alas, you don't observe ability, and must therefore make do with the short regression on schooling alone:

$$Y_i = \alpha^* + \rho^* S_i + \upsilon_i$$

- Substituting (2) into the formula for a bivariate regression slope reveals that:

$$\underbrace{\rho^*}_{short} = \frac{C(Y_i, S_i)}{V(S_i)} = \underbrace{\rho}_{long} + \underbrace{\gamma\delta_{AS}}_{OVB},$$

  where δ_{AS} is the regression of A_i on S_i

- Neat formula! We say:

  Short equals long -plus-
  the effect of omitted -times- the regression of omitted on included

  - This *omitted variables bias* (OVB) *formula* is regression's golden rule (the OVB term is $\gamma\delta_{AS}$)

- In a wage equation like (2) where the omitted variable is ability, labor economists refer to OVB as *ability bias*

  - Ponder the sign of ability bias: Is the short reg ρ^* too big or too small relative to the long-regression ρ that you seek?

2.2 Two vs. Four

Suppose now that your long regression includes four regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \underbrace{\beta_3 X_{3i} + \beta_4 X_{4i}}_{additional\ controls} + \varepsilon_i \qquad E[\varepsilon_i X_{ji}] = 0;\ j = 1,\ 2,\ 3,\ 4 \tag{3}$$

- You'd like to estimate (3), the long regression of your dreams. Alas, you're missing data on X_{3i} and X_{4i}. So, you settle for the short regression with only two:

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \nu_i \qquad E[\nu_i X_{ji}] = 0;\ j = 1,\ 2 \tag{4}$$

- What's the relationship between β_1^* and β_1? Between β_2^* and β_2? The regression anatomy formula formula for β_1^* gives

$$(short)\ \beta_1^* = \frac{C(Y_i, \tilde{x}_{1i})}{V(\tilde{x}_{1i})}, \tag{5}$$

  where this *auxiliary regression:*

$$X_{1i} = \pi_{10} + \pi_{11} X_{2i} + \tilde{x}_{1i},$$

  partials out (removes) the influence of X_{2i} on X_{1i}.

- To interpret this, substitute the long reg for Y_i in (5):

$$\begin{aligned}\beta_1^* &= \frac{C(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i,\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \frac{C(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \beta_1 \frac{C(X_{1i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \frac{C(\beta_3 X_{3i} + \beta_4 X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\ &= \beta_1 + \beta_3 \frac{C(X_{3i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \beta_4 \frac{C(X_{4i},\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})}\end{aligned}$$

Write this as:

$$\beta_1^* = \beta_1 + \beta_3 \delta_{31.2} + \beta_4 \delta_{41.2},$$

where

$$\delta_{31.2} = \frac{C(X_{3i,}\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \qquad \text{(Regression of } X_{3i} \text{ on } X_{1i} \text{ in a model that includes } X_{2i})$$

$$\delta_{41.2} = \frac{C(X_{4i,}\, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \qquad \text{(Regression of } X_{4i} \text{ on } X_{1i} \text{ in a model that includes } X_{2i})$$

Likewise:

$$\beta_2^* = \beta_2 + \beta_3 \delta_{32.1} + \beta_4 \delta_{42.1},$$

where

$$\delta_{32.1} = \frac{C(X_{3i,}\, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \qquad \text{(Regression of } X_{3i} \text{ on } X_{2i} \text{ in a model that includes } X_{1i})$$

$$\delta_{42.1} = \frac{C(X_{4i,}\, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \qquad \text{(Regression of } X_{4i} \text{ on } X_{2i} \text{ in a model that includes } X_{1i})$$

- OVB, same as it ever was:

  Short equals long plus {effect(s) of omitted in long -times- regression(s) of omitted on included}, with auxiliary regressions computed in a models maintaining the set of controls included in both short and long.

2.3 Sample Short and Long

- OVB holds in your data!
  - Let $\hat{\beta}_1^*$ be the OLS estimate of β_1^* in (4) and let $\hat{\beta}_2^*$ be the corresponding estimate of β_2^*. Then, we have:
$$\hat{\beta}_1^* = \frac{\sum Y_i \tilde{x}_{1i}}{\sum \tilde{x}_{1i}^2} = \hat{\beta}_1 + \hat{\beta}_3 \hat{\delta}_{31.2} + \hat{\beta}_4 \hat{\delta}_{41.2},$$
    where hats denote estimates and $\tilde{x}_{1i}$ is the sample residual from a regression of X_{1i} on X_{2i}.
  - Likewise,
$$\hat{\beta}_2^* = \frac{\sum Y_i \tilde{x}_{2i}}{\sum \tilde{x}_{2i}^2} = \hat{\beta}_2 + \hat{\beta}_3 \hat{\delta}_{32.1} + \hat{\beta}_4 \hat{\delta}_{42.1},$$
    where $\tilde{x}_{2i}$ is the sample residual from a regression of X_{2i} on X_{1i}.

- Show this at home

2.4 When Short Equals Long

Two scenarios yield short=long:

1. Omitted variables have coefficients of zero in long (in which case, they're not really "omitted")
2. Omitted variables are uncorrelated with included variables, that is, the coefficient on included is zero in a regression of omitted on included plus any covariates maintained in both short and long

- Important ideas! Let's see how they work in practice

3 Empirical OVB

Immigrant and native wages (working men aged 40-49 in the 2016 ACS)

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	67,179	44.59836	2.843473	40	49
wagp	67,179	77017	70468.34	4	665000
wkhp	67,179	44.73532	9.786132	1	99
uhe	67,179	33.90219	27.61486	.0016	201.5789
loguhe	67,179	3.264571	.7410341	-6.437752	5.306181
immig	67,179	.2214829	.4152479	0	1
yearsEd	67,179	13.83362	3.240573	0	21
hsgrad	67,179	.9243365	.264461	0	1
somecol	67,179	.4721565	.4992279	0	1
colgrad	67,179	.3860581	.4868478	0	1
asianpac	67,179	.0833147	.2763594	0	1
white	66,790	.7721216	.4194669	0	1
married	67,179	.7207461	.4486359	0	1

```
58 .
59 . ***short vs long***
60 .
61 . reg loguhe immig
```

Source	SS	df	MS	Number of obs	=	67,179
				F(1, 67177)	=	570.52
Model	310.655619	1	310.655619	Prob > F	=	0.0000
Residual	36578.9048	67,177	.544515307	R-squared	=	0.0084
				Adj R-squared	=	0.0084
Total	36889.5604	67,178	.549131567	Root MSE	=	.73791

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1637641	.0068562	-23.89	0.000	-.1772022	-.1503259
_cons	3.300842	.0032267	1022.99	0.000	3.294518	3.307166

```
62 .         gen beta_short=_b[immig]

63 . reg loguhe immig yearsEd
```

Source	SS	df	MS	Number of obs	=	67,179
				F(2, 67176)	=	8096.70
Model	7165.30841	2	3582.6542	Prob > F	=	0.0000
Residual	29724.252	67,176	.442483208	R-squared	=	0.1942
				Adj R-squared	=	0.1942
Total	36889.5604	67,178	.549131567	Root MSE	=	.66519

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.0346034	.0062671	-5.52	0.000	-.0468869	-.02232
yearsEd	.0999527	.0008031	124.46	0.000	.0983786	.1015267

```
64 .         gen beta_long=_b[immig]

65 .         gen gamma_long=_b[yearsEd]

66 .
67 . **Regression of omitted on included (aux reg)**
68 .
69 . reg yearsEd immig
```

Source	SS	df	MS	Number of obs	=	67,179
				F(1, 67177)	=	1893.82
Model	19342.5545	1	19342.5545	Prob > F	=	0.0000
Residual	686114.857	67,177	10.2135382	R-squared	=	0.0274
				Adj R-squared	=	0.0274
Total	705457.411	67,178	10.5013161	Root MSE	=	3.1959

yearsEd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-1.292219	.0296939	-43.52	0.000	-1.350419	-1.234019
_cons	14.11983	.0139745	1010.40	0.000	14.09244	14.14722

```
70 .         gen delta=_b[immig]

71 .
72 . **check OVB formula**
73 .
74 .         gen short_chk = beta_long + delta*gamma_long

75 .
76 .         sum short_chk beta_short beta_long gamma delta
```

Variable	Obs	Mean	Std. Dev.	Min	Max
short_chk	67,179	-.1637641	0	-.1637641	-.1637641
beta_short	67,179	-.1637641	0	-.1637641	-.1637641
beta_long	67,179	-.0346034	0	-.0346034	-.0346034
gamma_long	67,179	.0999527	0	.0999527	.0999527
delta	67,179	-1.292219	0	-1.292219	-1.292219

```
77 .
78 . ***repeat with maintained controls***
79 .
80 . cap drop delta short_chk beta_short beta_long gamma_long delta

81 .
82 . reg loguhe immig married age
```

Source	SS	df	MS	Number of obs	=	67,179
				F(3, 67175)	=	1261.06
Model	1966.79504	3	655.598348	Prob > F	=	0.0000
Residual	34922.7653	67,175	.519877415	R-squared	=	0.0533
				Adj R-squared	=	0.0533
Total	36889.5604	67,178	.549131567	Root MSE	=	.72103

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1844893	.0067131	-27.48	0.000	-.197647	-.1713317
married	.3467611	.0062125	55.82	0.000	.3345845	.3589376
agep	.0071519	.0009788	7.31	0.000	.0052334	.0090704
_cons	2.736543	.0439402	62.28	0.000	2.65042	2.822666

```
83 .         gen beta_short=_b[immig]

84 . reg loguhe immig yearsEd married age
```

Source	SS	df	MS	Number of obs	=	67,179
				F(4, 67174)	=	4753.57
Model	8138.3322	4	2034.58305	Prob > F	=	0.0000
Residual	28751.2282	67,174	.428011257	R-squared	=	0.2206
				Adj R-squared	=	0.2206
Total	36889.5604	67,178	.549131567	Root MSE	=	.65423

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.055513	.0061851	-8.98	0.000	-.0676358	-.0433901
yearsEd	.095556	.0007958	120.08	0.000	.0939963	.0971157
married	.2638873	.0056791	46.47	0.000	.2527563	.2750183
agep	.0086363	.0008882	9.72	0.000	.0068954	.0103772
_cons	1.379621	.0414398	33.29	0.000	1.298399	1.460843

```
-------------+------------------------------------------------------------

. gen beta_long=_b[immig]

. gen gamma_long=_b[yearsEd]

.
. **Regression of omitted on included (aux reg)**
.
. reg yearsEd immig married age

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(3, 67175)     =    979.45
       Model |  29564.8411         3  9854.94703   Prob > F        =    0.0000
    Residual |   675892.57    67,175  10.0616683   R-squared       =    0.0419
-------------+----------------------------------   Adj R-squared   =    0.0419
       Total |  705457.411    67,178  10.5013161   Root MSE        =     3.172

------------------------------------------------------------------------------
     yearsEd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -1.349747   .0295329   -45.70   0.000    -1.407631   -1.291862
     married |   .8672798   .0273309    31.73   0.000     .8137113    .9208482
        agep |  -.0155344   .0043061    -3.61   0.000    -.0239744   -.0070944
       _cons |   14.20029   .1933065    73.46   0.000     13.82141    14.57917
------------------------------------------------------------------------------

. gen delta=_b[immig]

.
. **check OVB formula**
.
. gen short_chk = beta_long + delta*gamma_long

.
. sum short_chk beta_short beta_long gamma_long delta

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   short_chk |     67,179   -.1844894           0  -.1844894  -.1844894
  beta_short |     67,179   -.1844893           0  -.1844893  -.1844893
   beta_long |     67,179    -.055513           0   -.055513   -.055513
  gamma_long |     67,179     .095556           0    .095556    .095556
       delta |     67,179   -1.349747           0  -1.349747  -1.349747

.
. log close
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/2020/1432apps/LN8log.smcl
  log type:  smcl
 closed on:   2 Mar 2020, 14:32:28
-------------------------------------------------------------------------------
```

- works again - phew!

Private college redux

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)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
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.

- Why does the private wage premium fall as we move from column 1 to column 2?

TABLE 2.5
Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

Notes: This table describes the relationship between private school attendance and personal characteristics. Dependent variables are the respondent's SAT score (divided by 100) in columns (1)–(3) and log parental income in columns (4)–(6). Each column shows the coefficient from a regression of the dependent variable on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

- Apply OVB: Take *Short* to be the bivariate coefficient reported in col 1 of MM T2.3; *Long* is the coefficient in the model that adds individual SAT scores, reported in col 2. We see that:

$$Short - Long = OVB = .212 - .152 = .06.$$

The effect of SAT in the long regression is .051, while MM T2.5 (above) shows the regression of SAT (omitted in short) on a private school dummy (included in short) produces a coefficient of 1.165. This confirms *OVB=Reg of omitted on included* × *Effect of omitted in Long = 1.165*×*.051=.06.* Phew!

- Why are SAT_i and other controls irrelevant for the private premium in cols 4-6 of MM T2.2-2.4? Must be like "no OVB"!

4 OVB and Selection Bias

- The OVB formula is algebra, true for any short-vs-long regression comparison
- Even so, we use the OVB formula to investigate selection bias – the math behind this is simple but the idea is subtle:
  1. Why do we *care* to go long? Because the Y_{0i}'s of those who attend a private college are likely better (on average)
  2. Regression with the right controls reduces, maybe even eliminates, selection bias arising from imbalanced Y_{0i}
  3. Verily, a regression coefficient so blessed should have no OVB in the sense that, once key controls are included, it matters not whether we add more
- A constant-effects causal effects model helps formalize this argument:
  - Let $Y_{0i} = \alpha + \eta_i$, where $E[Y_{0i}] = \alpha$; assume $Y_{1i} - Y_{0i} = \rho$, so we can write:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})P_i = \alpha + \rho P_i + \eta_i \tag{6}$$

- Private college isn't randomly assigned, so,

$$E[\eta_i|P_i] \neq 0$$

  - Because P_i is Bernoulli, the regression of Y_i on P_i produces

$$E[Y_i|P_i = 1] - E[Y_i|P_i = 0] = \rho + \underbrace{\{E[\eta_i|P_i = 1] - E[\eta_i|P_i = 0]\}}_{selection\ bias} \neq \rho$$

- **The causal interpretation of regression estimates is based on a key claim, called a *conditional independence assumption* (CIA)**
  - The CIA in this case asserts that, conditional on X_i, mean Y_{0i} does not depend on P_i:

$$E[Y_{0i}|\underbrace{P_i}_{poof!}; X_i] = E[Y_{0i}|X_i] = \alpha + \gamma' X_i \tag{7}$$

  Equivalently,

$$Y_{0i} = \alpha + \gamma' X_i + u_i,$$

  where $E[u_i|X_i] = 0$ by *definition* of u_i and $E[u_i|P_i] = 0$ by virtue of the CIA
- Plugging (7) into (6) then leads to a causal regression model:

$$Y_i = \alpha + \gamma' X_i + \rho P_i + u_i, \tag{8}$$

where X_i and P_i are both uncorrelated with u_i and coefficient ρ is the causal effect we seek

  - The CIA-satisfying X_i in DK02 and MM Chpt. 2 contains dummies for Barrons selectivity groups. DK02 argues that the CIA holds given these variables; suppose for the moment these key controls are also the only controls.

- It'd be nice to check the CIA by asking whether Y_{0i} is is correlated with P_i conditional on X_i. If only we could regress Y_{0i} on P_i and X_i...
  - Alas, Y_{0i} is unobserved
  - We do see variables that are surely correlated with Y_{0i} conditional on X_i, like family income and SAT scores. Call these variables W_i.
  - We can therefore check the CIA by running the auxiliary regression:

$$W_i = \pi_0 + \pi_1' X_i + \pi_2 P_i + \xi_i$$

- Columns 3 and 6 in MM T2.5 report estimates of π_2 in this model, where W_i is log parental income and own SAT scores. Sure enough, these estimates are small, and not significantly different from zero.
- By the OVB formula, therefore, we should expect estimates of ρ_1 in the long regression that adds controls for W_i, written:

$$Y_i = \alpha_1 + \gamma_1' X_i + \gamma_2' W_i + \rho_1 P_i + u_{1i}, \tag{9}$$

  to be the same as estimates of short-regression ρ in (8)
  - Remarkably, once we know where they applied and were admitted, applicants who go public and private have similar ability and family background
  - In other words, no OVB!

5 Beware Bad Control: An OVB Conundrum

- MHE Table 3.2.1 compares schooling coefficients estimated with controls added sequentially for family background, AFQT scores (a measure of ability), and occupation

Table 3.2.1: Estimates of the returns to education, males

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupational dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey)
The number in the first row is the coefficient on years of education in a weighted least squares regression of education on wages with the indicated controls. The number in parentheses is the associated standard error. The sample is restricted to males, weighted by NLSY sampling weights, and the sample size is 2434.
* Additional controls are mother's/father's years of education, and dummy variables for race and Census region.

- Controls here seem to matter, but that alone doesn't ensure that more control is better as we navigate a path around selection bias
- Some controls are *bad controls*, meaning they're better left omitted in efforts to mitigate selection bias. This tricky point is tackled in MM Chapter 6.1 and MHE Chapter 3.2