Example: Health Insurance

- We're about to do our first reading from MM.
- You'll read about an observational dataset called the National Health Information Survey (NHIS).
  - The "treatment" is having health insurance.
  - The outcome is the overall health index (rated 1–5 on a survey).
  - The covariates are all the characteristics listed in panel B of Table 1.1.
- Reading: Chapter 1 of MM, through the second line of page six.

When Covariates Reveal Failure to Have Apples-to-Apples Comparison

- You just saw an example that fails a randomization check: Insured versus uninsured individuals had different values of covariates, from education to family size to income.
- Let's look at Table 1.1 on page 5 of MM and, in particular, the "family income" row near the bottom of the table.
  - There, we see that insured husbands have $60,000 more family income than uninsured husbands.
  - This is very statistically significant, because the standard error of the difference in column three is $1,355.
  - The difference between the insured and uninsured groups is not only economically large but also statistically significant: more than 40 standard errors.
  - We don't have an apples-to-apples comparison here; income could easily improve someone's health as much as insurance status does.

Example With Actual Randomization

- The health insurance data we saw just failed a covariate balance check.
  - Not surprising, since health insurance was not randomly assigned in the observational data of the NHIS.
- Next, we're going to read an example where health insurance status was randomly assigned in an experiment.
- Start with the second paragraph of page 16 of MM, and read through the first paragraph of page 22.
- Optional reading: Pages 6–16 of MM contain a review of potential-outcomes notation in a health insurance example, as well as a review of the concept of expectation.

Table 1.3 (Panel A)

Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible — catastrophic (2)	Coinsurance — catastrophic (3)	Free — catastrophic (4)	Any insurance — catastrophic (5)
		A. Demographic characteristics			
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)

Table 1.3 of MM (Panel B)

Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible — catastrophic (2)	Coinsurance — catastrophic (3)	Free — catastrophic (4)	Any insurance — catastrophic (5)
		B. Baseline health variables			
General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood	122	2.32	.91	1.12	1.39
pressure (mm Hg)	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

Experimental Design Lesson: Keep It Simple

- On page 18 of the reading, note that the HIE was complicated with many small treatment groups, spread over more than a dozen insurance plans.
- With just under 4,000 subjects, it isn't possible to get precise measurements of the results for each insurance plan.
- Confidence intervals are too wide to say anything valuable.
- In other words, we didn't have enough statistical power in this design, which is something worth thinking about when you do your experimental team project for class.
- It's possible to fix an experiment where you spread your observations too thin. In this case, researchers pool together a number of treatments into four broad categories of health insurance plans: catastrophic, deductible, coinsurance, and free.
- Note that if you find you have low statistical power in your project, you may want to do fewer experimental treatments, or have a plan that will allow you to pool together data across multiple treatments in order to get more power.

Results of the HIE and More Recent Health Insurance Experiment in Oregon

- The Oregon experiment wasn't intentionally designed by a researcher; it is what we call a **natural experiment**.
    - Oregon had limited budget available to expand health insurance for the poor, so they made it available via a lottery, for fairness.
    - This is almost as good as if someone had designed a research experiment.
- Think about the differences between the two experiments.
- How would you summarize the results of the two experiments?
- Read MM, from page 22 to the top of page 30.

Table 1.4

Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible—catastrophic (2)	Coinsurance—catastrophic (3)	Free—catastrophic (4)	Any insurance—catastrophic (5)
		A. Health care use			
Face-to-face visits	2.78	.19	.48	1.66	.90
	[5.50]	(.25)	(.24)	(.25)	(.20)
Outpatient expenses	248	42	60	169	101
	[488]	(21)	(21)	(20)	(17)
Hospital admissions	.099	.016	.002	.029	.017
	[.379]	(.011)	(.011)	(.010)	(.009)
Inpatient expenses	388	72	93	116	97
	[2,308]	(69)	(73)	(60)	(53)
Total expenses	636	114	152	285	198
	[2,535]	(79)	(85)	(72)	(63)
		B. Health outcomes			
General health index	68.5	−.87	.61	−.78	−.36
	[15.9]	(.96)	(.90)	(.87)	(.77)
Cholesterol (mg/dl)	203	.69	−2.31	−1.83	−1.32
	[42]	(2.57)	(2.47)	(2.39)	(2.08)
Systolic blood pressure (mm Hg)	122	1.17	−1.39	−.52	−.36
	[19]	(1.06)	(.99)	(.93)	(.85)
Mental health index	75.5	.45	1.07	.43	.64
	[14.8]	(.91)	(.87)	(.83)	(.75)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table reports means and treatment effects for health expenditure and health outcomes in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in Column (1). Standard errors are reported in parentheses in Columns (2)–(5); standard deviations are reported in brackets in Column (1).

Table 1.6

OHP effects on health indicators and financial health

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
	A. Health indicators			
Health is good	.548	.039		
		(.008)		
Physical health index			45.5	.29
				(.21)
Mental health index			44.4	.47
				(.24)
Cholesterol			204	.53
				(.69)
Systolic blood pressure			119	−.13
(mm Hg)				(.30)
	B. Financial health			
Medical expenditures			.055	−.011
> 30% of income				(.005)
Any medical debt?			.568	−0.32
				(.010)
Sample size	23,741		12,229	

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on health indicators and financial health. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

Summary of Results

- Increasing coverage causes increased spending on health care.
  - Downward-sloping demand
- Increasing coverage does not produce large improvements in physical health.
- Increasing coverage in the Oregon experiment shows increases in mental health and in financial status.
- Unfortunately, we don't see any evidence of decreased emergency-room utilization.

An Example of Noncompliance

- In the middle of page 29, we see that only 25% of the sample of lottery winners actually obtained insurance.
    - Winning the Oregon lottery merely qualified you to apply for the insurance but was no guarantee that you would get it.
- This is an example of noncompliance: not everyone we intended to treat actually received the treatment.
- This indicates that the effects of the Oregon insurance are four times larger for each person who actually received insurance, than the "intent to treat" estimates given in Tables 1.5 and 1.6.
- In the noncompliance chapter, we'll talk about this kind of problem in more detail.

Reading: Checking for Covariate Balance

- You're going to read GG, Section 4.3.
- On page 107, the authors suggest using an F test to do a joint test of whether multiple covariates are different from each other across treatments.
    - To give you some motivation for this idea, consider MM Table 1.3. Recall that we looked at 11 different covariates in this table. We looked at individual t-tests for each covariate.
- This is usually sufficient, but in case the covariates are correlated to each other, one might want to do a joint test of all covariates at once across two treatments.
    - To do this, we could regress the binary treatment variable (0 for catastrophic, 1 for deductible) on the whole list of 11 covariates and then look at the regression F statistic to test the hypothesis that all 11 differences are jointly equal to zero.

Covariate Imbalance Problem From Ad Effectiveness Research (Lewis & Reiley)

- We had a third-party data firm who matched up individual Yahoo! advertising data with the retailer's sales data, using name and address information in both databases, then de-identified each observation before sending it to us for analysis.
- Over 100 columns of data:
  - Number of ad views each week
  - Online and offline purchases each week, with 52 weeks of history
- Problem: When we got back the data and did a quick calculation of the treatment effect, we found that the control group was purchasing significantly less than the treatment group.
  - This was unexpected.

Checking Covariates to Determine What Went Wrong

- Our most important covariate: pretreatment purchases.
- The treatment group started a gradual decline in sales relative to the control group, weeks before the experiment began.
  - But if randomization worked correctly, we should have had pretreatment purchases identical between the two groups. We clearly had a problem.
- Another clue: lower sales for those who actually received ads in the treatment group (versus those untreated).

Checking Covariates to Determine What Went Wrong (cont.)

- Another clue: online sales much lower than expected in both treatment and control groups, and almost exactly zero in the treatment group.
- Eventually, we realized that the data-matching firm had a buffer-overflow problem that they never noticed.
- This is a great example of an administrative error detected by paying careful attention to the covariates in the data.

Next: A Brief History of Experiments in the Social Sciences

Read MM, pages 30–33.

Milestones in Experimentation

- The prophet Daniel proposed an experiment on a vegetarian diet.
- James Lind (1742) experimented with citrus in the diet and demonstrated that it cured scurvy.
    - Note that while he didn't randomize, he chose his 12 subjects with covariate balance in mind.
- Charles Peirce (1885): first recorded use of random assignment.
- Sir Ronald Fisher (1935): detailed theory of randomized experiments.
- Because of their focus on social science, Angrist and Pischke are missing one key historical reference on experimentation: Galileo (1500s).

Optional Statistics Review

Depending on how solid you feel your statistics background to be, you may also optionally wish to review the appendix to MM Chapter 1, which contains material on the following topics:

- Sample means and being correct on average
- Variance, standard deviation, and standard error
- How far off the sample mean can be (CLT)
- Comparing two means
- Confidence intervals and t-statistics

Next: Regression

- Checking for covariate balance is just one use of covariates.
- As Gerber and Green noted on page 109 of your previous reading:
  - When treatment and control have slightly different covariate values, we may wish to control for these covariates.
  - We do so using regression.
- Your next reading will therefore be the regression chapter of MM.
- Please read MM Chapter 2, pages 47–55, on the concept of matching.

Motivating Regression: Matching Observations

- We want to know whether more expensive private colleges cause their alumni to earn more income than they would if they'd gone to public universities.
- Recall the problem of selection bias:
  - People who go to Harvard often have other advantages besides going to a prestigious university, so those students would end up with higher incomes on average even if they had gone to public universities.
- One proposal to fix this problem is to compare two students who are accepted to exactly the same set of colleges, but one matriculates private while another matriculates public.

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.

- Group A: Estimated value of private college is $105,000 – $110,000 = –$5,000.
- Group B: Estimated value of private college is $60,000 – $30,000 = +$30,000.
- Weighted average: 0.6(–$5,000) + 0.4(+30,000) = +$9,000.

The Danger of Selection Bias

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.

- If we naively compare the mean for people who attend private colleges with the mean for those who attend public colleges, we get a much larger answer: mean difference of $20,000.
- By comparing Group A to Group B, we can see that students who apply to more private schools tend to earn higher incomes. So the differences don't depend merely on where you attended but also on the types of people who choose to submit more applications to private schools.

The Danger of Selection Bias (cont)

- Why do we use both application and acceptance to create our matching groups? Why not just acceptance decisions?
  - My answer: A person who applies to lots of selective schools but got into only one may be less talented than the average person admitted to the selective school.

When the Observational Matching Strategy Fails

- The matching strategy intends to compare apples to apples by matching students who have identical application and acceptance decisions.
    - This is an observational strategy rather than an experiment, so we know it's still possible to get the wrong answer even with this matching.
- Can you think of a story under which the results of the matching exercise are still plagued by selection bias?

When the Observational Matching Strategy Fails (cont.)

- My story: Consider two people who get in to the same pair of schools, but Alice chooses the private school while Bob chooses the public school.
- Perhaps Alice knows she is likely to benefit a lot from the private school, given her intellectual interests in ornithology and computer programming, while Bob knows that for him, the increased benefit of the private school is not worth the cost.
- We still have selection effects: We would overestimate the benefit of switching someone from a public to a private school, because people who choose to go to private school benefit from it more than the people who choose public schools.

Regression: An Elegant Method to Control for Covariates

- With a pure matching estimator, we need a match from the untreated group for each member of the treated group.
  - In this case, private versus public college students
- Regression gives us a more convenient way to summarize the results of such an effort, by specifying a single equation simultaneously containing both the treatment variable and the covariates we care about accounting for.

MM, Section 2.2

- A few pointers:
  - Equation 2.1 applies to the simulated data in Table 2.1. P_i is a dummy variable equal to 1 if the student attended a private school and 0 if the student attended a public school.
  - A_i is a dummy variable equal to 1 if the individual is in Group A and 0 if the individual is in Group B.
    - Remember that we have to drop from our analysis any individuals in Groups C and D, because we don't get any public-private comparisons out of those groups.
- Now read MM, Section 2.2, page 55 just to the middle of page 59, before moving on.

Regression

- Understanding regression results really requires only simple arithmetic.
- If needed, please review the regression content in weeks 10–14 of the Exploring and Analyzing Data course.

Equation 2.1

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

- This is a multivariate regression, because it includes the two variables P_i (for private college) and A_i (for application/acceptance group).
- Which one is the treatment variable, and which one is the covariate?
  - P_i is treatment, and A_i is covariate.
- Both are dummy variables.

Equation 2.1 (cont.)

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

- If we dropped A_i from the regression so that P_i is the only variable left on the right-hand side, this would be equivalent to computing the simple difference in means between private and public schools.
  - In the regression, the coefficient β tells us the average increase in Y when we go from $P_i = 0$ to $P_i = 1$.
  - The coefficient and standard error we get in this regression on a single dummy variable is equivalent to what we would get for the difference in sample means.
- Adding A_i to the regression allows the mean income to be different for group A than group B, which helps us correct selection bias, because students in each group may earn different incomes for reasons other than their school being private versus public.
- Including this covariate in the regression accomplishes a matching function just like we saw earlier but now in the regression context.

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.

Assignment

Now read MM from page 59 to the bottom of page 64 to see our second example of a regression.

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.

Reading a Table of Regression Results

- In our first regression example, we had just two regressors, both of them discrete. This time we have some continuous regressors as well. Which of the covariates are continuous?
  - Own SAT score, parental income
- What does it accomplish to have the left-hand side be log(earnings) rather than just earnings?
  - The coefficient tells us the percentage change in Y when we increase P_i from 0 to 1.
- What does it mean that the treatment effect changes as we move from column one to column three but does not change much as we move from column four to column six?

Next: Tennessee STAR Experiment on Kindergarten Class Size

- The supplemental reading comes from Chapter 2 of Angrist and Pischke's previous book, *Mostly Harmless Econometrics*.
- We now move to an example of regression not with observational data but with an experiment.
- Read from the last paragraph of page 16 to the end of page 24.
- Note: The last two paragraphs of page 21 discuss a technique called **regression discontinuity** that can be used with observational data to try to get valid causal inference; we'll discuss this later in the week on observational data.
- Pay particular attention to making sure you can read the regression results in Table 2.2.2.

Tennessee STAR Experiment

TABLE 2.2.2

Experimental estimates of the effect of class size on test scores

Explanatory Variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian	—	—	8.35	8.44
			(1.35)	(1.36)
Girl	—	—	4.48	4.39
			(.63)	(.63)
Free lunch	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Teacher master's degree	—	—	—	−0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Notes: Adapted from Krueger (1999), Table V. The dependent variable is the Stanford Achievement Test percentile score. Robust standard errors allowing for correlated residuals within classes are shown in parentheses. The sample size is 5,681.

Tennessee STAR Experiment (cont.)

- Note that the treatment effect does change a bit, from 4.82 to 5.37, when we add school fixed effects to the regression. Why?
- Randomization happened within each school, not necessarily consistently across schools.
    - For example, urban schools might have had 1/4 small classes while rural schools had 1/2 small.
    - If the treatment group disproportionately contains rural schools with lower test scores, this gives us downward bias in the ATE.
- Fixed effects let us estimate the effect within each school, then pool across schools.
    - Same as our "selectivity groups" in the private-school example
    - Also reduces noise in the outcome due to school-level factors (similar to blocking)

Next: Difference in Differences Reading

- Read FE, Section 4.1.
- DID analysis is a special case of regression with covariates.
- I find "rescaling" a confusing term and prefer to think in terms of differences.

DID: Redefines Outcome Variable to Before-After Change

- Instead of the original Y, we use the change in Y.
- DID is very similar to inserting the past (lagged) value of the outcome into the regression as a covariate.
- What's the main difference between lagged-dependent-variable covariate and DID?
  - DID constrains the slope to be 1.
- Why might we do this?
  - E.g., Lewis and Reiley (2014), Sections 3.3–3.4.
  - We get additional power (similar to the case of blocking).

Reading: Adjusting for Covariates Using Regression

Next, read FE Section 4.2. This reading should reinforce the principles we have already learned by looking at the Tennessee STAR experiment together.

Key Points to Review

- With randomized treatment assignment, we know that treatment is uncorrelated with everything else: both observable covariates and unobservables we can't measure. So in an experiment, we don't have to worry about omitted-variable bias, because we should get approximately the same answer no matter how many covariates we include.
- What including covariates does for us in an experiment is explain some of the residual variance in the regression, allowing us to shrink the standard error on the treatment effect. You can see this in Table 2.2.2 from the supplemental reading on the Tennessee STAR experiment.

Next Reading: Omitted Variable Bias in the Observational Study of Private Schools

- Remember that in MM, we're using observational data, and we're going to use this to illustrate the concept of omitted variable bias.
- The previous analysis we looked at was able to examine only 5,600 out of 14,000 student observations, because we had to drop all the observations that didn't have exact matches where another student had the same acceptances but matriculated differently.
- This reading looks at a way to make use of all 14,000 observations.
- In MM, now read from the bottom of page 64 to the middle of page 68.

Observational Analysis: Earnings Benefits of Private Colleges

- We simplify the model in order to be able to look at a larger sample size.
- Previously, we included only those 5,600 observations for whom we could measure effects within a matching selectivity group.
- Now we replace the 150 selectivity group dummies with just four variables:
  - Average SAT score of schools applied to
  - Sent two applications
  - Sent three applications
  - Sent four or more applications
  - (Omitted category: sent a single application)
- This is one continuous variable and three discrete variables. (Note that we have effectively transformed number of applications from a continuous variable into three binary variables.)

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				.110	.082	.077
schools applied to ÷ 100				(.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.

Reading Regression Results

- Having log(earnings) as the dependent variable allows us to interpret the difference as a percentage of earnings (Y).
    - Note that with continuous variables, sometimes we might choose to take the log both of the Y variable and the X variable before regressing.
    - That coefficient would be interpreted as an elasticity: What percentage increase in Y do we see as X increases by 1%?
- Notice that we are measuring the difference between "four or more applications" and those in the default (omitted) category. The default group submitted only a single college application.

Omitted Variable Bias

In observational data, we always have the potential for omitted variable bias.

- Note the direction of bias when we omit the "own SAT score" variable. We overestimate the effect in column one, because SAT score is positively correlated both with earnings and those who go to private school. The coefficient shrinks as we move to columns two and three and four, with more covariates.
- If we had randomized private school assignment, we would have had private school uncorrelated with SAT, so the coefficient would not have changed much.
- As before, the takeaway is that the correlation between earnings and private school is spurious: The causal effect appears to be approximately zero.

Omitted Variable Bias (cont.)

- Omitted variable bias looks much better once we have controlled for selectivity of colleges applied to.
- The treatment coefficient doesn't change much as we move across columns four through six. The treatment coefficient stabilizes.
- The authors take this as evidence that with the four selectivity control variables, we are approaching the good causal inference of an experiment. However, even here, we could have an OVB problem.
- As noted earlier, suppose that choosing a private school is more likely to be done by people who predict they will especially benefit from private school. We can't measure this unobserved tendency to benefit from private school, but it could always be lurking there.

Next: Omitted Variable Bias Theory

- We're going to read MM, Section 2.3, pages 68–78.
- Some tips before you read:
  - By **short**, the authors refer to a regression with a short list of regressors.
  - By **long**, they denote a regression with a longer list of regressors, including previously omitted ones.
- What you'll see in Table 2.5 are results for what's indicated in the equations by the phrase "regression of omitted on included."
- What we're doing here is taking Table 2.3 as a starting point. There we see what happens when we omit the variables "own SAT score" and "log(parental income)": OVB makes us overestimate the true effect of private college.
- In Table 2.5, we regress these omitted variables on the other included variables in the regression, to see how they are correlated.

Direction of Bias in Observational Regressions With Omitted Variables

- Suppose Y is a function of both X_1 and X_2:

$$Y = X_1 + X_2$$

- Instead of this long regression, we instead omit X_2 and run the short regression:

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

- When we do this, any effects of X_2 can influence Y only through X_1 in the equation.

Example

- Suppose we regress earnings on years of education, and we omit IQ.
- Suppose people with higher IQ are likely to get more school.
- This means we are going to overestimate the effects of schooling in our short regression, because the schooling variable is going to capture the effects of schooling PLUS the fact that higher schooling is associated with higher IQ, which itself causes higher earnings.

Direction and Magnitude of Bias

Depend on two things:

1. How correlated is X_2 with Y? If the omitted variable has no effect of its own, there's no bias.
2. How correlated is X_2 with X_1? If the omitted variable is not correlated with the included variable, then there's no bias.

- We just talked about overestimation bias for earnings when IQ is omitted.
    - Both correlations are positive, and the amount of bias is therefore positive.

Direction and Magnitude of Bias (cont.)

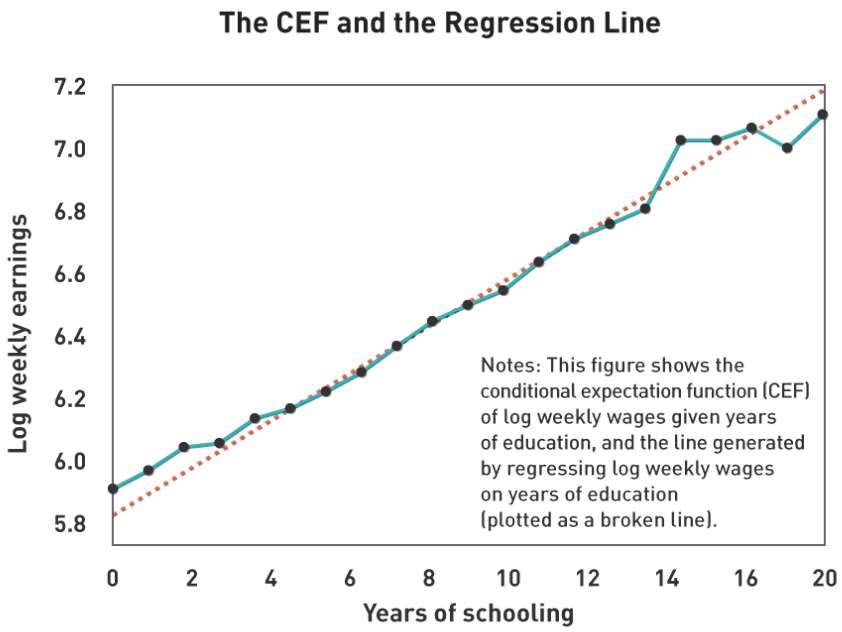
- We can also get underestimation bias. For example, consider the omitted variable of rock-music talent.
- Assume this is positively correlated with earnings and negatively correlated with years of school, because great rock musicians may choose to skip college and start a band out of high school.
- We underestimate the returns to schooling in this case because we miss the fact that the greatest musicians earn large amounts of money and are less likely to go to college than the rest of the population.
- Notice that we can never be sure we've gotten rid of all omitted variable bias in observational data. But in experimental data, we can guarantee that the treatment X_1 is uncorrelated with everything else, and thereby remove omitted variable bias.

Points to Remember

- Regression was developed in the late 1800s by Galton and Yule. Note that the calculations were much harder in the absence of computers.
- The most useful item is to learn where the word **regression** came from.
    - Galton's first linear model (1886) showed "regression to the mean" when he regressed sons' heights on their fathers' heights.
    - The slope was less than 1, indicating that each person's height is, on average, intermediate between his father's height and the mean height.
    - As you know, the term **regression** has now come to mean least-squares fitting of a model.
    - Galton's regression had only one variable on the right-hand side.
- George Yule developed multiple regression (1899), with more than one variable on the right-hand side.
    - His application was whether various English antipoverty programs increased or decreased the number of people in poverty.

Next: Theory of Regression

- Next, I'd like you to take a look at the appendix to Chapter 2, on the theory of regression.
- Right now, I'd just like you to read about the conditional expectation function on pages 82–85.
- You can optionally read pages 86–94, if you are interested, on the theory of:
    - Covariance
    - Residuals
    - Dummy variables
    - Derivation of the OVB bias formula
    - Interpretation of logarithmic models
- We'll save the rest of the appendix for next week.
- Before continuing, please read MM, pages 82–85.



The Conditional Expectation Function

- Linear regression of log(earnings) on years of schooling gives us the best linear approximation to this curve.
- To measure the curve in full generality, we would need to use a "saturated" model, where instead of including "years of schooling" as a linear regressor on the right side, we'd instead use a full set of dummy variables, one for every possible number of years of schooling between zero and 20.
- Next, read another excerpt from Angrist and Pischke's other book, *Mostly Harmless Econometrics*, which we've made available here. This time, you'll be reading Section 3.1.4 on saturated models, pages 48–51.

Choosing Covariates and Functional Form in a Regression

- What is a saturated model?
  - Each value of the covariate gets a different dummy variable to represent it.
  - With multiple covariates, we also have to include all possible interactions between these dummy variables (to estimate a two- or three-dimensional surface).
- Why might a researcher choose a model that is not fully saturated?
  - Linear in a continuous variable (not a full set of dummies).
  - Exclude interaction effects.
  - Answer: Worry about too little data per cell to get identification and precision. If you have only 500 observations, you can't estimate a full 10×10×10 model. But you could try a coarser model.

Choosing Covariates and Functional Form in a Regression (cont.)

- What is good about using lots of covariates?
  - Increased statistical precision
- What is bad about using lots of covariates?
  - "Fishing expeditions" happen quite frequently with observational data.
  - Not much downside in an experiment, if we have made sure that X is independent of all covariates.

Lessons to Remember

- Checking for covariate balance can help us identify execution problems in experiments.
- A multivariate regression can contain both discrete and continuous regressors.
- Most observational data are studied using multivariate regression. Because it's observational, we can never be sure the coefficients give us true causal effects.
- Omitted variable bias—learn to tell stories about the expected direction of bias.

Lessons to Remember (cont.)

- Experiments give us independence between treatment and potential outcomes, so we don't have to worry about OVB. So why use regression to analyze experiments?
    - Increased statistical power. Measuring and including covariates can help us gain precision of our experimental treatment estimates.
    - Covariates soak up residual variance and shrink standard errors on the treatment effect.