corinna.birner@stud-mail.uni-wuerzburg.de
max.mueller@stud-mail.uni-wuerzburg.de

Understanding Metrics based on Mastering Metrics

Corinna Birner & Max Müller

University of Würzburg

May 28, 2020

Chapter 2: Regression

Overview

Today we will continue our journey on the path from cause to effect. Therefore, we will discuss regressions as one possible way of finding causal relations in our data. Our topics will be the following:

Why Regressions?

A Tale of two Colleges

Regression

Appendix

Why Regressions?

- When the path to random assignment is blocked, we look for alternate routes to causal knowledge
- The most basic of these tools is regression, which compares treatment and control subjects who have the same observed characteristics.
- Regression based causal inference is predicated on the assumption that when key observed variables have been made equal across treatment and control groups, selection bias from the things we can't see is also mostly eliminated.

Distinction

- Whats important to understand is, that a regression is not an identification strategy to find causal relationships in data.
- It is a statistical tool to help our research design.
- Identification strategy (research design): E.g.: RCT, Regression Discontinuity Design, Difference in Difference
- Statistical tool: Regression

A Tale of two Colleges

To explain the basic framework of how regressions work, we look at an example from the college choice in America, or if the choice between public or private colleges does have an impact on economic returns.

Average pay in priv. colleges: 29.000, public: 9.000
An elite private education might be better in many ways: the classes smaller, the athletic facilities newer, the faculty more distinguished, and the students smarter.

But: is that worth 20000?

Harvard or U-Mass?

- In this Example we take a look at universities in Massachusetts
- The apples-to-apples question in this case asks how much a 40-yearold Massachusetts-born graduate of Harvard would have earned if he or she had gone to the University of Massachusetts (U-Mass) instead
- Comparisons of earnings between those who attended Harvard and U-Mass
- Comparison reflects the fact that Harvard grads typically have better high school grades and higher SAT scores, are more motivated, and perhaps have other skills and talents.

Harvard or U-Mass?

- Earnings comparisons across alma maters should be contaminated by selection bias.
- Selection bias is eliminated by random assignment
  - $\Rightarrow$ In that case there is no possibility for random assignment
- Two things must be accomplished: larger group comparisons to draw general lessons and the ceteris paribus idea (all things equal).

Comparisons in a fictious world

- Suppose the only things that matter in life are your SAT scores and where you go to school.
- Consider Uma and Harvey both 1,400 on the SAT.
- Uma went to U-Mass, while Harvey went to Harvard. We start by comparing Uma's and Harvey's earnings.
- Because we've assumed that all that matters for earnings besides college choice is the combined SAT score, Uma vs. Harvey is a ceteris paribus comparison and according to that everything Harvey is earning more is causal to his attendance at Harvard

Comparisons in the real world

- Of course life is more complicated.
- Uma is a young woman, and Harvey is a young man. Women with similar educational qualifications often earn less than men.
- The earnings gap could be due to discrimination or the superiority of Harvard, we don't know.
- We want to disentangle the pure Harvard effect.
- If we exchange Harvey with Hannah and compute the average earnings difference among Harvard and U-Mass students with the same gender and SAT score:
- This is an econometric matching estimator that controls for, or holds fixed—sex and SAT scores
- Estimator captures the average causal effect of a Harvard degree on earnings

Controlling for other Factors

- There is more to school choice than just sex, schools, and SAT scores.
- Since college attendance decisions aren't randomly assigned, we must control for all factors that determine both attendance decisions and later earnings.
- E.g.: student characteristics, like writing ability, diligence, family connections, and more.
- Control for such a wide range of factors is tough:infinite possibilities, characteristics hard to quantify.
- So what we do here is controlling for observables, which is a first step to make them more comparable
- Because it is easy to assume that, if they are comparable in observables, they might be in unobservables too.
- Shortcut: the characteristics of colleges to which students applied and were admitted.

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.

- Applications, admissions, and matriculation decisions for nine students
- E.g.: Group B: Admitted to same schools: Number 4 enrolled at Ivy, while number 5 chose Altered State.
- Earnings differential: $30,000(60 - 30 = 30)$.
- This gap suggests a substantial private school advantage.

All things equal?

- The Fact, that they were admitted at every school, and just chose different school suggests that they must be pretty similar in their abilities and potential and every earnings gap must be contributed to their school choice.
- Using only relevant groups A and B: construct a weighted average: $(\frac{3}{5} \times -5000) + (\frac{2}{5} \times 30000) = 9000$
- By emphasizing larger groups, this weighting scheme uses the data more efficiently and generates a statistically more precise summary of the private-public earnings differential.
- Important: apples-to-apples and oranges-to-oranges nature of the underlying matched comparisons:
  - $\rightarrow$ Apples in group A are compared to other group A apples, while oranges in group B are compared only with oranges.

The Regression Framework

- regression estimates are weighted averages of multiple matched comparisons
- key ingredients in the regression recipe are:
  - the dependent variable, in this case, student i's earnings later in life, also called the outcome variable (denoted by Y_i)
  - the treatment variable, in this case, a dummy variable that indicates students who attended a private college or university (denoted by P_i)
  - a set of control variables, in this case, variables that identify sets of schools to which students applied and were admitted.
- Dummies classify data into simple yes-or-no categories

The Regression Framework

- The regression model in this context is an equation linking the treatment variable to the dependent variable while holding control variables fixed by including them in the model.
- With only one control variable, Ai (if youre group A or not), the regression of interest can be written as:

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

- The regression parameters—called regression coefficients—are:
  - the intercept, ("alpha");
  - the causal effect of treatment, ("beta");
  - and the effect of being a group A student, ("gamma")

The Regression Framework

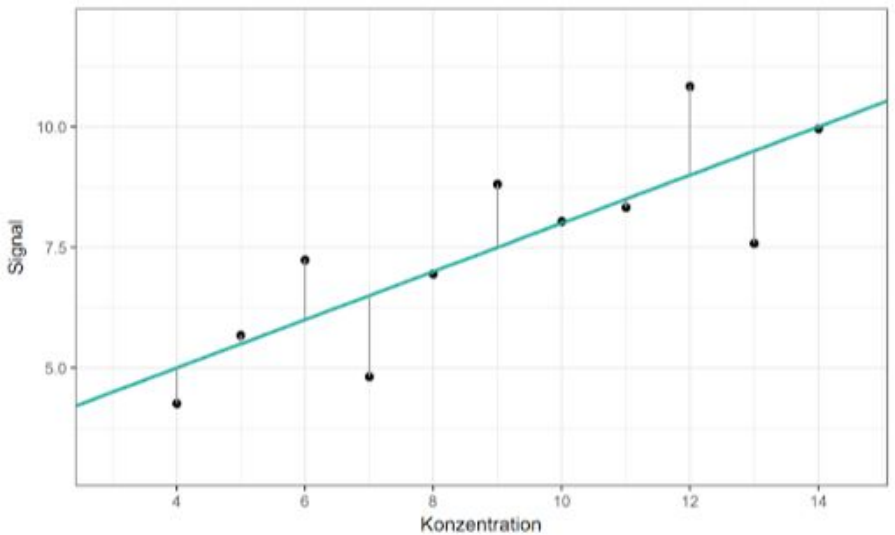
- And there is the residual, ϵ_i (also called an error term).
- Residuals are defined as the difference between the observed Yi and the fitted values generated by the specific regression model we have in mind.
- These fitted values are written as:

$$\hat{Y}_i = \alpha + \beta P_i + \gamma A_i$$

- Residuals are given by:

$$\epsilon_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \beta P_i + \gamma A_i)$$

Explaining the residual



- Difference between the observed value of the dependent variable (Y) and the predicted value ($\hat{Y}$) is the residual
- The data points usually don't fall exactly on this regression equation line; they are scattered around.
- A residual is the vertical distance between a data point and the regression line.

Error Term or Residual?

- We must make one important distinction very clear.
- The error term (causal language) and the residual (statistical language) are closely related, but not the same.
- An error term is the difference between the observed value and the true value, which is unobserved.
- So an Error term is a theoretical concept that can never be observed, it is all that explains Y_i and is not included in our regression.
- Important: Nothing in our error term can correlate with the treatment, otherwise our causal effect β would be confounded by something in the error term.
- A residual is the difference between the observed value and the predicted value by our model.
- So the residual is a real world value that is calculated for each time a regression is done.

The Regression Framework

- Regression analysis assigns values to model parameters (α, β *and* γ) so as to make $\hat{Y}_i$ as close as possible to Y_i.
- This is accomplished by choosing values that minimize the sum of squared residuals, leading to ordinary least squares (OLS) for the resulting estimates.
- This linear regression works under the foundation that that our world can be explained by linear relationships
- Important: A regression itself can just tell us something about the statistical correlation, the research design is crucial for our causality.
- Regression estimates (and the associated standard errors used to quantify their sampling variance) are readily constructed using computers and econometric software.

Public-Private Face-Off

- Back to our example of public and private schools: 14,000 former students in the CB Dataset made comparable by using the matching matrix
- Because were interested in comparing returns from public and private schools: leaves 5,583 matched students for analysis. These matched students fall into 151 similar selectivity groups containing both public and private students.
- New Regression Model:

$$lnY_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_i GROUP_{ji} + \delta_1 SAT_i + \delta_2 lnPI_i + \epsilon_i$$

Public-Private Face-Off

- Two Changes:
  - log of earnings on the left-hand side $\rightarrow$ logged dependent variable allows regression estimates to be interpreted as a percent change.
  - includes many control variables
- The parameter β in this model is still the treatment effect of interest, an estimate of the causal effect of attendance at a private school.
- Y_j, for j $= 1$ to 150, are the coefficients on 150 selectivity-group dummies, denoted $GROUPji$.
- 150 because we need one group as a reference group.
- Addition of two further control variables: individual SAT scores (SAT_i) and the log of parental income (PI_i),

Regressions Run

- We start with regression estimates of the private school earnings advantage from models with no controls.
- The coefficient from a regression of log earnings (in 1995) on a dummy for private school attendance, with no other regressors gives the raw difference in log earnings between those who attended a private school and everyone else
- Private school students are estimated to have earnings about 14% higher than the earnings of other students.
- Standard errors quantify the statistical precision of the regression estimates reported here.

Private school effects

	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.

- Standard error in column (1) is .055. The Estimate .135 is more than twice the size of the standard error → unlikely a chance finding.
- The private school coefficient is statistically significant.
- Interesting descriptive fact, but, due to selection bias.

Private school effects

- Column (2) of Table 2.2: Every 100 points of SAT achievement are associated with a 5 perc. point earnings gain.
- Controlling for other factors: brings the private school premium down a little further, to a still substantial and statistically significant .086, reported in column (3) of the table.
- Column (4) reports estimates from a model with no controls, but the dummy for each matched college selectivity group in the sample.
- The Premium then falls to almost 0, columns (5) and (6) show that the premium moves little when controls for ability and family background are added to the model.
- Private university attendance seems unrelated to future earnings once we control for selection bias.

Another Approach

	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.

- Perhaps our focus on public-private comparisons misses the point.
- Students may benefit from attending schools like Ivy, Leafy, or Smart because their classmates are so much better.
- The synergy generated by a strong peer group may be the feature that justifies the private school price tag.

Another Approach

- (3) of Table 2.4 show that students who attended more selective schools do markedly better in the labor market, with an estimated college selectivity effect on the order of 8% higher earnings for every 100 points of average selectivity increase.
  - $\rightarrow$ 8% more earnings per 100 average SAT Score Points
- Selection bias due to the greater ambition and ability of those who attend selective schools.
- Show average college selectivity to be essentially unrelated to earnings

Ceteris Paribus

- Regression is a way to make other things equal, but equality is generated only for variables included as controls on the right-hand side of the model.
- Failure to include enough controls or the right controls still leaves us with selection bias.
- The regression version of the selection bias generated by inadequate controls is called omitted variables bias (OVB).

Ceteris Paribus

- Illustration for OVB: Back to the Example of Group A and B:
- The "long regression" here includes the dummy variable, Ai, which indicates those in group A.
- We write the regression model that includes A_i as:

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + \epsilon_i^l$$

- Does the inclusion of A_i matter for estimates of the private school effect in the regression above?
- Suppose we make do with a short regression with no controls. This can be written as:

$$Y_i = \alpha^s + \beta^s P_i + \epsilon_i^s$$

Ceteris Paribus

- The Difference between β^s and β^l is the OVB due to omission of A_i in the short regression.
- Here, OVB amounts to \$10,000, a figure worth worrying about.
- Why such a big effect: In part from the fact that the mostly private students in group A have higher earnings anyway, regardless of where they enrolled.
- Inclusion of the group A dummy in the long regression controls for this difference.

The OVB Formula

- Connection between short and long regression coefficients has two components:
  - The relationship between the omitted variable (A_i) and the treatment variable (P_i); regression of the omitted variable A_i on the private school dummy.

$$A_i = \pi_0 + \pi_1 P_i + u_i$$

  - The relationship between the omitted variable (A_i) and the outcome variable (Y_i). This is given by the coefficient on the omitted variable in the long regression, γ
- So we can write the following OVB formula:

$$OVB = Effect\ of\ P_i\ in\ short\ -\ Effect\ of\ P_i\ in\ long$$
$$= \beta^s - \beta^l = \pi_1 \times \gamma$$

OVB

- Back to our example:
- π_1 in our five-student example is therefore .1667.
- Calculation of the OVB:

$$\begin{aligned} OVB &= Short \ - \ Long \\ &= \beta^s - \beta^l \\ &= 20.000 - 10.000 = 10.000 \end{aligned}$$

and/or

$$\begin{aligned} OVB &= Regression\ of\ omitted\ on\ included \times Effect\ of\ omitted\ in\ long \\ &= \pi_1 \times \gamma = .1667 \times 60.000 = 10.000 \end{aligned}$$

OVB

- OVB formula is a mathematical result that explains differences between regression coefficients in any short-versus-long scenario, irrespective of the causal interpretation of the regression parameters.
- The OVB formula is a tool that allows us to consider the impact of control for variables we wish we had.
- This in turn helps us assess whether ceteris is indeed paribus.
- We can't use data to check the consequences of omitting variables that we don't observe, but we can use the OVB formula to make an educated guess for the consequences of their omission.

Regression Sensitivity Analisis

- We might have missed some control variables in our analisis to completely eliminate selection bias
- therefore we check how robust our treatment effect is
- this means that our treatment effect is insensitive to adding or dropping a particular control variable
- as we have seen in the columns (4) - (6) the private school coefficients were insensitve to SAT scores and parental income
- this can be explained by the OVB Formula

Regression Sensitivity Analisis

- short model: regression of log wages on P_i with no controls
- long model: adds individual SAT scores
- OVB = Short - Long = .212 - .152 = .06 (from Table 2.3)
- OVB = Regression of omitted on included x Effect of omitted in long = 1.165 x .051 = .06

Regression Sensitivity Analisis

	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.

- Relationship between private school attendance and personal characteristics
- Regression of omitted on included
- Regression of omitted SAT_i on P_i

Regression Sensitivity Analisis

- short model: regression of log wages on P_i with no controls
- long model: adds individual SAT scores
- OVB = Short - Long = .212 - .152 = .06 (from Table 2.3)
- OVB = Regression of omitted on included x Effect of omitted in long = 1.165 x .051 = .06
- when including self-revelation controls, we get
- OVB = Short - Long = .034 - .031 = .003
- OVB = .066 x .036 = .0024 (instead of .003 due to rounding errors)
- students who chose private and public schools aren't very different (at least in SAT scores)

In a nutshell

- for causal comparisons we have to compare like with like
- good comparisons eliminate systematic differences that are associated with outcomes
- the method of matching sorts individuals into groups with the same values of control variables and matched comparisons are averaged
- the method of regression is an automated matchmaker
- the regression estimate is an average of within-group comparisons
- OVB is the difference between short and long regression coefficients

Regression and the CEF

- conditional expectation tell us how the population average of one variable changes as we move the conditioning variable (over the values this variable can assume)
- the collection of all such averages is called conditional expectation function (CEF)
- the CEF with K conditioning variables is written

$$E[Y_i|X_{1i}, ..., X_{Ki}]$$

- we saw that private school attendance seems to be unrelated to average earnings once certain controls are held fixed
- we now suppose that the CEF of log wages is a linear function of these certain conditioning variables, so that:

$$E[lnY_i|P_i, GROUP_i, SAT_i, lnPI_i]$$

$$= \alpha + \beta P_i + \sum \gamma_j Group_{ij} + \delta_1 SAT_i + \delta_2 lnPI_i$$

Regression and the CEF

- when CEF of ln Y_i is a linear funciton of these variables, the regression of ln Y_i recovers this linear funcion
- linear models help us to understand regression but regression is even more flexible due to theoretical properties:
- If $E[Y_i|X_{1i},...,X_{Ki}] = a + \sum_{k=1}^{K} b_k X_{ki}$ for some constants a and $b1,...,b_{Ki}$ the regression of Y_i on $X_{1i}....,X_{Ki}$ has intercept a and slope $b1,...,b_k$
- if the CEF of $Y_i on X_{1i},...,X_{Ki}$ is linear, the regression is it
- If $E[Y_i|X_{1i},...,X_{Ki}]$ is a nonlinear funcion of the conditioning variables, the rgression of Y_i on $X_{1i},...,X_{Ki}$ gives the best linear approximation to this nonlinear CEF
- if the CEF is linear, regression finds it and if not, regression finds a good approximation to it

Bivariate Regression and Covariance

- regression is closely related to the concept variance
- the covariance between two variables X_i and Y_i is

$$C(X_i, Y_i) = E[(X_i - E[X_i])(Y_i - E[Y_i])]$$

- the covariance has three properties:
  - covariance of a variable with itself is its variance: $C(X_i, X_i) = \sigma_x^2$
  - if the expectation of X_i or Y_i is 0, the covariance is the expectation of their product: $C(X_i, Y_i) = E[X_i Y_i]$
  - the covariance between linear functions of X_i and Y_i ($W_i = a + bX_i$ and $Z_i = c + dY_i$) is: $C(W_i, Z_i) = bdC(X_i, Y_i)$

Bivariate Regressoin and Covariance

- a bivariate regression is a regression with one regressor, X_i, and an intercept
- the bivariate regression slope and intercept are the values of a and b that minimize the residual sum of squares:

$$RSS(a, b) = E[Y_i - a - bX_i]^2$$

- the solution for the bivariate case is

$$b = \beta = \frac{C(Y_i, X_i)}{V(X_i)}$$

$$a = \alpha = E[Y_i] - \beta E[X_i]$$

- when two variables are uncorrelated (covariance of 0), the regression of one on the other generates a slope coefficient of 0!

Fits and Residuals

- regression breaks any dependent variable into two pieces: $Y_i = \hat{Y}_i + e_i$
- the fitted value $\hat{Y}_i$ is the part of Y_i explained by the model and e_i is the residual
- regression residuals and regressors are uncorrelated
- suppose α and $\beta_1, ..., \beta_K$ are the intercept and the slope coefficients of a regression, the fitted values are

$$\hat{Y}_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki}$$

and the regression residuals are

$$e_i = Y_i - \hat{Y}_i = Y_i - \alpha - \sum_{k=1}^{K} \beta_k X_{ki}$$

Fits and Residuals

- properties of residuals:
- have expectation 0: $E[e_i] = 0$
- are uncorrelated with all the regressors and with the corresponding fitted values $E[X_{k\,i}\, e_i] = 0$ and $E[\hat{Y}_i\, e_i] = 0$

Regression for Dummies

- special case: bivariate regression with a dummy variable
- the conditional expectation of Y_i given a dummy Z_i can be

$$E[Y_i|Z_i = 0] = \alpha$$

and

$$E[Y_i|Z_i = 1] = \alpha + \beta$$

so that

$$\beta = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

- we can therefore write:

$$E[Y_i|Z_i] = E[Y_i|Z_i = 0] + (E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0])Z_i$$

$$= \alpha + \beta Z_i$$

- $E[Y_i|Z_i]$ is a linear function with slope β and intercept α!

Regression and OVB

- A lot of times regressions are multiple and include more control variables
- suppose the causal variable is X_{1i} and the control variable is X_{2i}
- the coefficient on X_{1i} in a regression controlling for X_{2i} can be written as:

$$\beta_1 = \frac{C(Y_i, \tilde{X}_{1i})}{V(\tilde{X}_{1i})}$$

with $\tilde{X}_{1i}$ being the residual from a regression of X_{1i} on X_{2i} :

$$X_{1i} = \pi_0 + \pi_1 X_{2i} + \ X_{1i}$$

Regression and OVB

- OVB in multiple regression
- call the coefficient on X_{1i} in a multivariate regression controlling for X_{2i} the long coefficient β^l:

$$Y_i = \alpha^l + \beta^l X_{1i} + \gamma X_{2i} + e_i^l$$

- call the coefficient in a bivariate regression (without X_{2i}) β^s:

$$Y_i = \alpha^s + \beta^s X_{1i} + e_i^s$$

- the OVB would be $\beta^s = \beta^l + \pi_{21}\gamma$ where γ is the coefficient on X_{2i} in the long regression and π_{21} is the coefficient on X_{1i} in a regression of X_{2i} on X_{1i}
- short equals long plus the effect of the ommited times the regression of omitted on included!

Models with Log

- Let's use a bivariate regression to explain why we use logs instead of just Y_i:

$$lnY_i = \alpha + \beta P_i + e_i$$

  with P_i being a dummy for private school attendance.
- regression in this case fits the CEF perfectly
- now we create a ceteris paribus change in P_i for student i:

$$lnY_{0i} = \alpha + e_i$$

$$lnY_{1i} = \alpha + \beta + e_i$$

- the difference in potential outcomes is therefore $lnY_{1i} - lnY_{0i} = \beta$

$$\beta = ln\frac{Y_{1i}}{Y_{0i}} = ln(1 + \frac{Y_{1i} - Y_{0i}}{Y_{0i}})$$
$$= ln(1 + \Delta\%Y_p) \approx \Delta\%Y_p$$

Models with Log

- therefore, we can answer the question of why we use logs in regressions as follows:
- the regression slope in a model with ln Y_i gives the approximate percentage change in Y_i generated by changing the corresponding regressor

Regression Standard Errors and Confidence Intervals

- the standard error of the slope estimate in a bivariate regression $\hat{\beta}$ is similar to the one we looked at in Chapter 1: $SE(\hat{\beta}) = \frac{\sigma_e}{\sqrt{n}} x \frac{1}{\sigma_X}$
- σ_e is the standard deviation of the regression residuals and σ_X is the standard deviation of the regressor
- this formula assumes homoskedasticity meaning that the variance of residuals is unrelated to regressors

Regression Standard Errors and Confidence Intervals

- in case that the homoskedasticity might not be satisfied, meaning we have hetereoskedascity, we need to use robust standard errors:

$$RSE(\hat{\beta}) = \sqrt{\frac{V(\tilde{X}_{ki} e_i)}{n(\sigma^2_{X_k})^2}}$$

Thank you very much for listening