1. Introduction: regression and causality

LPO 8852: Regression II

Sean P. Corcoran

Regression

Regression is a technique used to estimate the *conditional expectation function* (CEF) for Y_i given the values of one or more other variables X_{ik}, $k = 1, ..., K$ $(X_{i1}, X_{i2}, ...X_{iK})$. That is, it seeks to estimate parameters of $E[Y_i|X_{ik}]$ that provide the mean of Y given specific values of X. Note:

- The conditional expectation function may not be *linear*
- The conditional expectation function may not be *causal*, but may be useful for *prediction* (in a statistical sense). More on this soon.

Note the CEF is a population concept, with a sample analog.

Regression - CEF

From Angrist & Pischke (2009): CEF of log weekly wages given years of completed schooling

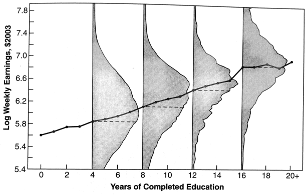

Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

Regression

We can decompose a random variable Y_i into two parts, the CEF and an error term: $Y_i = E[Y_i|X_i] + \epsilon_i$, where:

- ϵ_i is mean independent of X_i, that is $E[\epsilon_i|X_i] = 0$ and
- ϵ_i is uncorrelated with any function of X_i

This is not a statement about causality, but rather just a decomposition into a piece "explained by X_i" and a leftover orthogonal (uncorrelated) piece.

Linear regression

The population regression function is the *line* that best fits the population distribution of (Y_i, X_i) in that it minimizes the sum of the squared errors (in the population).

- Simple: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Multiple: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_3 X_{ki} + \epsilon_i$

The solutions to the least squares problem are familiar to you. In the simple regression case:

$$\beta_1 = \frac{Cov(Y_i, X_i)}{V(X_i)}$$

$$\beta_0 = E[Y_i] - \beta_1 E[X_i]$$

Linear regression

Why might we want to estimate the population regression function?

1. If the CEF happens to be linear, then the PRF is the CEF. This is unlikely in most real world-cases but true in two special cases: joint normality, and *saturated* regression models.
2. The PRF is the best linear predictor of Y_i given the X_i.
3. The PRF provides the least squares approximation to the CEF when the CEF is nonlinear.

Note: a *saturated* regression model is a regression model with discrete explanatory variables, where the model includes a separate parameter for every possible combination of values taken on by the explanatory variables.

Saturated regression models

Example: two dummy (0/1) explanatory variables X_1 and X_2

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 X_{1i} X_{2i} + \epsilon_i$$

There are four possible combinations of X_1 and X_2 and thus four possible predictions of $Y|X$:

X_1	X_2	$E(Y|X)$
0	0	β_0
1	0	$\beta_0 + \beta_1$
0	1	$\beta_0 + \beta_2$
1	1	$\beta_0 + \beta_1 + \beta_2 + \gamma_1$

The coefficients are *main effects* (β_1, β_2) and an *interaction term* (γ_1).

Regression - CEF

From Angrist & Pischke (2009): linear regression as an approximation to CEF

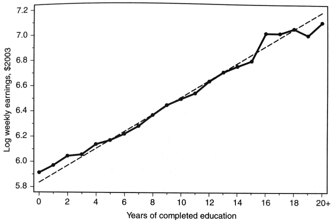

Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Regression and causality

The PRF is useful for several reasons, but its slope coefficients are *not necessarily causal*. So when will regression have a causal interpretation? "A regression is causal when the CEF it approximates is causal" (Angrist & Pischke, 2009). That is, the CEF needs to describe differences in average potential outcomes for a given reference population.

If we have sufficiently conditioned on the right X_i's, it might. I.e., *design matters*

Potential outcomes

Suppose there is one dichotomous explanatory variable D_i, where $D_i = 1$ is "treated" and $D_i = 0$ is "not treated." For every individual i there are two potential outcomes:

- $Y_i(1)$ or Y_{i1} = outcome when $D = 1$
- $Y_i(0)$ or Y_{i0} = outcome when $D = 0$

These are referred to as "potential outcomes" since individuals are not observed in more than one state (the "fundamental problem of causal inference"). If we could observe these, the *treatment effect* of D for individual i would be $Y_i(1) - Y_i(0)$. We could average these in the population to get an *average treatment effect*: $E(Y_i(1) - Y_i(0))$.

A *counterfactual* is the outcome for an individual in a different state. E.g., the counterfactual for a treated i here would be $Y_i(0)$.

Potential outcomes

A conditional expectation function for this simple example is the following:

$$E[Y_i|D_i] = \beta_0 + \beta_1 D_i$$

Pretty simple and linear—two possible values for D_i and two conditional expectations.

- Key question: can this CEF be interpreted as *causal*?
- Does it describe differences in *potential outcomes* for a given reference population?
- When economists use the term *ceteris paribus* (all else equal) they are usually thinking of differences in potential outcomes: the change in (expected) outcomes across states, holding all else equal.

Example

Does attending a highly-selective private college result in higher earnings?

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.

Source: *Mastering Metrics ch. 2*

Example

The simple regression in column (1) estimates a CEF (the model is fully saturated). But does it describe differences in potential outcomes?

$$\beta_1 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Is this the same as:

$$E[Y_i(1) - Y_i(0)]?$$

Only if:

$$E[Y_i|D_i = 1] = E[Y_i(1)]$$
$$E[Y_i|D_i = 0] = E[Y_i(0)]$$

Example

Attendance at a private college is not randomly assigned; we should be concerned that the CEF does not describe differences in average *potential* outcomes. It may be that students attending selective private colleges are better qualified on a number of dimensions than students not attending such colleges.

If the CEF we are estimating does not describe differences in average potential outcomes, we say the causal effect is not *identified*.

Another example: class size

Omitted variables bias

Suppose instead that potential outcomes are described by the following "long" regression, where Y_i is (log) earnings, P_i is an indicator variable for private college attendance and A_i is a measure of "ability":

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

The "short" regression estimated in column (1) above is:

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

We can estimate the "short" regression, but if the true model of potential outcomes is the "long" regression ($\gamma \neq 0$), we may have *omitted variables bias*. The error term in the "short" regression is: $e_i^s = \gamma A_i + e_i^\ell$.

Omitted variables bias

There is a formal (and mechanical) link between β^s and β^ℓ:

$$\beta^s = \beta^\ell + \pi_1 \gamma$$

Where:

- γ comes from the long regression: it is the relationship between A_i and Y_i (conditional on P_i).
- π_1 comes from an "auxiliary" regression of the omitted variable (A_i) on the included variable (P_i).

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

Example

Auxiliary regressions where A_i is the student's SAT score (in hundreds):

Table 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212 (.060)	.152 (.057)	.139 (.043)	.034 (.062)	.031 (.062)	.037 (.039)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)		.036 (.006)	.009 (.006)
Log parental income			.181 (.026)			.159 (.025)
Female			−.398 (.012)			−.396 (.014)
Black			−.003 (.031)			−.037 (.035)
Hispanic			.027 (.052)			.001 (.054)
Asian			.189 (.035)			.155 (.037)
Other/missing race			−.166 (.118)			−.189 (.117)
High school top 10%			.067 (.020)			.064 (.020)
High school rank missing			.003 (.025)			−.008 (.023)
Athlete			.107 (.027)			.092 (.024)
Average SAT score of schools applied to ÷ 100				.110 (.024)	.082 (.022)	.077 (.012)
Sent two applications				.071 (.013)	.062 (.011)	.058 (.010)
Sent three applications				.093 (.021)	.079 (.019)	.066 (.017)
Sent four or more applications				.139 (.024)	.127 (.023)	.098 (.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.

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

Omitted variables bias: example

Assessing omitted variables bias:

- $\hat{\beta}^s = 0.212$
- $\beta^s = \beta^\ell + \pi_1\gamma$
- What do you think the signs of π_1 and γ are?
- The estimated $\widehat{\pi_1} = 1.165$ (the difference in SAT scores between private and public college students) and $\hat{\gamma} = 0.051$
- So, $0.212 = \beta^\ell + (1.165 * 0.051)$. Our estimator of β using β_s is likely biased upward.
- $\hat{\beta}^\ell = 0.152$ (compare to column (2))

Example

Of course, a model with two explanatory variables is probably not sufficient in this example: it alone is unlikely to describe differences in average potential outcomes. Column (3) of Table 2.3 includes additional student covariates, such as log parental income, gender, race/ethnicity, athlete, and HS top 10%. The reduction in $\hat{\beta}$ suggests the estimator used in column (2) was still biased upward.

In a setting like this, one should still be concerned about *unobserved* omitted variables.

Example

In an attempt to address these, columns (4) - (6) represent what might be called a "self-revelation" model. They include the number and characteristics of schools to which students *applied*. This behavior might proxy for unobserved differences that are related to both private college attendance and earnings.

Example

Table 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212 (.060)	.152 (.057)	.139 (.043)	.034 (.062)	.031 (.062)	.037 (.039)
Own SAT score ÷ 100		.051 (.008)	.024 (.006)		.036 (.006)	.009 (.006)
Log parental income			.181 (.026)			.159 (.025)
Female			−.398 (.012)			−.396 (.014)
Black			−.003 (.031)			−.037 (.035)
Hispanic			.027 (.052)			.001 (.054)
Asian			.189 (.035)			.155 (.037)
Other/missing race			−.166 (.118)			−.189 (.117)
High school top 10%			.067 (.020)			.064 (.020)
High school rank missing			.003 (.025)			−.008 (.023)
Athlete			.107 (.027)			.092 (.024)
Average SAT score of schools applied to ÷ 100				.110 (.024)	.082 (.022)	.077 (.012)
Sent two applications				.071 (.013)	.062 (.011)	.058 (.010)
Sent three applications				.093 (.021)	.079 (.019)	.066 (.017)
Sent four or more applications				.139 (.024)	.127 (.023)	.098 (.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.

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

Example

In columns (4) - (6) the estimated coefficient on private school shrinks and becomes statistically insignificant.

Interestingly, the correlation between *own* SAT score and private school enrollment is eliminated once application behavior has been controlled for (the self-revelation model). See column (3) of Table 2.5.

Ceteris paribus?

Even with rich controls we may remain concerned that the CEF we are estimating is not a description of how potential outcomes relate to our explanatory variable of interest. Example of Dinardo & Pischke (1997) on the returns to computer use on the job.

The techniques covered in this course are methods that have been developed to address this concern, in the absence of a randomized experiment.

Designs for causal inference

Abadie & Cattaneo (2018) provide a review of designs for causal identification in program evaluation.

1. Randomized experiments
2. Matching estimators
3. Difference-in-differences and synthetic controls
4. Instrumental variables
5. Regression discontinuity

Regression anatomy

The "regression anatomy" formula is a useful algebraic property of regression. Suppose X_1 is a causal variable of interest (e.g., prvaite college attendance) and X_2 is a control (e.g., SAT score). Then:

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

where $\tilde{X}_{1i}$ is the *residual* from a regression of X_{1i} on X_{2i}:

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

Intuitively, "purge" X_{1i} of its covariance with X_{2i}, and regress Y_i on the residual.

Regression anatomy

This extends to models with more than 2 regressors:

$$\beta_K = \frac{Cov(Y_i, \tilde{X}_{Ki})}{V(\tilde{X}_{Ki})}$$

where $\tilde{X}_{Ki}$ is the *residual* from a regression of X_{Ki} on *all other* covariates.

Also known as the Frisch-Waugh-Lovell theorem.