Mastering Regression

Master Joshway

ASSA Continuing Ed: January 2020

Prerequisites

A World of Potentials

Acts demolish their alternatives, that is the paradox.

- James Salter (1975)

- The road without electronics leads to Y_{0i}
  - The road to distraction leads to Y_{1i}
- Let D_i indicate treatment or exposure to an intervention of interest
  - Like classroom electronics allowed ... or health insurance
- These are potential outcomes, only one is seen:

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

- $Y_{1i} - Y_{0i}$ is an unknowable individual electronics causal effect
- We seek, therefore, after average causal effects

RCT Theory

- We observe $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

$$\begin{aligned} &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{TOT} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}} \end{aligned}$$

  a fundamental causal conundrum (the attentive are anyway better; the insured anyway healthier)
- Models often assume constant effects: $Y_{1i} = Y_{0i} + \kappa$
- When D_i is randomly assigned,

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0],$$

  and

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

- *Random assignment removes selection bias (with or w/o constant effects)*

HEY, WHERE'D YA GO TO SCHOOL?

Regression Replaces Randomization

- We can't always run RCTs – and regressions are run more easily and faster! Yet, can regression really be causal?
- Define *potential outcomes* for two roads, public and private college

- American private schools are elite: expensive and selective (like MIT)
- Does private education pay? Write
  - Y_{1i} for graduate i's earnings having gone private ($P_i = 1$)
  - Y_{0i} for graduate i's counterfactual ($P_i = 0$)

- Get personal: most of my students are headed to Google & Goldman; yet, others there went to state schools
  - *Does MIT matter for you? (Michigan is cheaper!)*

Two College **Roads**

Regression and the CIA

- Private Y_{0i}'s are better (on average)
  - Regression reduces–maybe even eliminates–the resulting selection bias
- Let $Y_{0i} = \alpha + \eta_i$; assume $Y_{1i} - Y_{0i} = \beta$
- Though $E[\eta_i | P_i] \neq 0$, we assume *controls* X_i satisfy a conditional independence assumption (CIA):

$$E[\eta_i | P_i, X_i] = E[\eta_i | X_i] = \gamma' X_i$$

- This leads to

$$Y_i = \alpha + \gamma' X_i + \beta P_i + u_i,$$

  where
  - β is causal
  - γ is inconsequential
  - $E[u_i X_i] = 0$ by construction
- Note the asymmetry: in our design-based paradigm, regressors are not all created equal

Appraising Degrees (with Regression)

Appendix 1: School-Average SAT Score and Net Tuition of C&B Institutions

Institution	School-average SAT score in 1978	1976 Net tuition ($)
Barnard College	1210	3530
Bryn Mawr College	1370	3171
Columbia University	1330	3591
Denison University	1020	3254
Duke University	1226	3052
Emory University	1150	3237
Georgetown University	1225	3304
Hamilton College	1246	3529
Kenyon College	1155	3329
Miami University (Ohio)	1073	1304
Northwestern University	1240	3676
Oberlin College	1227	3441
Pennsylvania State University	1038	1062
Princeton University	1308	3613
Rice University	1316	1753
Smith College	1210	3539
Stanford University	1270	3658
Swarthmore College	1340	3122
Tufts University	1200	3853
Tulane University	1080	3269
University of Michigan (Ann Arbor)	1110	1517
University of North Carolina (Chapel Hill)	1080	541
University of Notre Dame	1200	3216
University of Pennsylvania	1280	3266
Vanderbilt University	1162	3155
Washington University	1180	3245
Wellesley College	1220	3312
Wesleyan University	1260	3368
Williams College	1255	3541
Yale University	1360	3744

- MM Chpt 2 (based on DK 2002) compares grads of these schools, *conditional on where they applied/were admitted*

Matchmaker, Matchmaker . . . Find Me a College!

Ambition and opportunity defined

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.

Regs Run in MM Chapter 2

- With one control variable, A_i, indicating group A in a sample containing A and B, an *alma mater* reg can be written:

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

- With many groups and a few other covs:

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + e_i \tag{2}$$

  This controls for 151 groups instead of two as in the example
  - Parameters γ_j, for $j = 1$ to 150, are the coefficients on 150 Barron's selectivity-group dummies, denoted $GROUP_{ji}$

- The CIA makes this causal:

$$E[Y_{0i} | \underbrace{P_i}_{poof!}; GROUP_i, SAT_i, \ln PI_i] = E[Y_{0i} | GROUP_i, SAT_i, \ln PI_i]$$

Make Me a Match . . . Run Me a Regression

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.135	0.095	0.086	0.007	0.003	0.013
	(0.055)	(0.052)	(0.034)	(0.038)	(0.039)	(0.025)
Own SAT score/100		0.048	0.016		0.033	0.001
		(0.009)	(0.007)		(0.007)	(0.007)
Predicted log(Parental Income)			0.219			0.190
			(0.022)			(0.023)
Female			-0.403			-0.395
			(0.018)			(0.021)
Black			0.005			-0.040
			(0.041)			(0.042)
Hispanic			0.062			0.032
			(0.072)			(0.070)
Asian			0.170			0.145
			(0.074)			(0.068)
Other/Missing Race			-0.074			-0.079
			(0.157)			(0.156)
High School Top 10 Percent			0.095			0.082
			(0.027)			(0.028)
High School Rank Missing			0.019			0.015
			(0.033)			(0.037)
Athlete			0.123			0.115
			(0.025)			(0.027)
Selection Controls	N	N	N	Y	Y	Y

Notes: Columns (1)-(3) include no selection controls. Columns (4)-(6) include a dummy for each group formed by matching students according to schools at which they were accepted or rejected. Each model is estimated using only observations with Barron's matches for which different students attended both private and public schools. The sample size is 5,583. Standard errors are shown in parentheses.

Table 2.2: Private School Effects: Barron's Matches

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.212	0.152	0.139	0.034	0.031	0.037
	(0.060)	(0.057)	(0.043)	(0.062)	(0.062)	(0.039)
Own SAT Score/100		0.051	0.024		0.036	0.009
		(0.008)	(0.006)		(0.006)	(0.006)
Predicted log(Parental Income)			0.181			0.159
			(0.026)			(0.025)
Female			-0.398			-0.396
			(0.012)			(0.014)
Black			-0.003			-0.037
			(0.031)			(0.035)
Hispanic			0.027			0.001
			(0.052)			(0.054)
Asian			0.189			0.155
			(0.035)			(0.037)
Other/Missing Race			-0.166			-0.189
			(0.118)			(0.117)
High School Top 10 Percent			0.067			0.064
			(0.020)			(0.020)
High School Rank Missing			0.003			-0.008
			(0.025)			(0.023)
Athlete			0.107			0.092
			(0.027)			(0.024)
Average SAT Score of				0.110	0.082	0.077
Schools Applied to/100				(0.024)	(0.022)	(0.012)
Sent Two Application				0.071	0.062	0.058
				(0.013)	(0.011)	(0.010)
Sent Three Applications				0.093	0.079	0.066
				(0.021)	(0.019)	(0.017)
Sent Four or more Applications				0.139	0.127	0.098
				(0.024)	(0.023)	(0.020)

Note: Standard errors are shown in parentheses. The sample size is 14,238.

Table 2.3: Private School Effects: Average SAT Controls

REGRESSION THEORY

Population Regression

- Population regression solves a theoretical best linear prediction (BLP) problem. The $K \times 1$ regression slope vector, β, can be defined as:

$$\beta = \arg\min_b E\left[\left(Y_i - X_i' b\right)^2\right]$$

- Using the first-order condition,

$$E\left[X_i\left(Y_i - X_i' b\right)\right] = 0,$$

  the solution for b can be written

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right]$$

- By construction, $E\left[X_i\left(Y_i - X_i'\beta\right)\right] = 0$: the pop resid, defined as $Y_i - X_i'\beta = e_i$, is uncorrelated with the regressors, X_i

- e_i owes its meaning and existence to β

Three Reasons to Love Regression Fearlessly

1. Regression solves the population least squares problem: it's the MMSE BLP of Y_i given X_i
2. If the conditional expectation function (CEF) is linear, regression is it
3. Regression is the **best linear approximation** to the CEF:

$$\beta = \arg\min_b E\{(E[Y_i|X_i] - X_i' b)^2\}.$$

- What do these properties depend on?
  - Nothing!
  - If the regression you've got is not be the one you want, that's your fault

The CEF is All You Need (but weight!)

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS              Number of obs =  409435
-------------+------------------------------           F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793           Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019           R-squared     =  0.1071
-------------+------------------------------           Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893           Root MSE      =  .67879
-------------+----------------------------------------------------------------
             |              Robust                   Old Fashioned
    earnings |      Coef.   Std. Err.      t           Std. Err.        t
-------------+----------------------------------------------------------------
      school |   .0674387   .0003447    195.63          .0003043    221.63
      const. |   5.835761   .0045507   1282.39          .0040043   1457.38
------------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =  540.31
       Model | 1.16077332      1  1.16077332           Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358           R-squared     =  0.9660
-------------+------------------------------           Adj R-squared =  0.9642
       Total | 1.20159212     20  .060079606           Root MSE      =  .04635
-------------+----------------------------------------------------------------
     average |              Robust                   Old Fashioned
   _earnings |      Coef.   Std. Err.      t           Std. Err.        t
-------------+----------------------------------------------------------------
      school |   .0674387   .0040352     16.71         .0029013     23.24
      const. |   5.835761   .0399452    146.09         .0381792    152.85
------------------------------------------------------------------------------
```

Figure 3.1.3: Micro-data and grouped-data estimates of returns to schooling. Source: 1980 Census - IPUMS, 5 percent sample. Sample is limited to white men, age 40-49. Derived from Stata regression output. Old-

Regression for Dummies

- The CEF for any dummy D_i takes on two values:

$$E\left[Y_i|D_i=0\right] = \alpha \quad (3)$$

$$E\left[Y_i|D_i=1\right] = \alpha+\beta \quad (4)$$

- This CEF is linear in D_i, so regression fits it perfectly:

$$E\left[Y_i|D_i\right] = E\left[Y_i|D_i=0\right]+\beta D_i = \alpha+\beta D_i$$

  where

$$\beta = E\left[Y_i|D_i=1\right] - E\left[Y_i|D_i=0\right]$$

- Now add controls: consider (2) w/group dummies only. Here, the private coefficient is

$$E\left\{\left(E\left[Y_i|P_i=1, GROUP_i\right] - E\left[Y_i|P_i=0, GROUP_i\right]\right) w(GROUP_i)\right\}$$

  for weights $w(GROUP_i)$ proportional to $V(P_i|GROUP_i)$

Regression Anatomy Lesson

- Bivariate reg recap: $\beta_1 = \frac{Cov(\mathrm{Y}_i, x_i)}{V(x_i)}$; $\alpha = E[\mathrm{Y}_i] - \beta_1 E[\mathrm{X}_i]$
- With multiple regressors, the k-th slope coefficient is:

$$\beta_k = \frac{Cov(\mathrm{Y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}, \tag{5}$$

  where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates
- Each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialing out" other variables in the model
- Verify regression-anatomy by subbing

$$\mathrm{Y}_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + ... + \beta_{\mathrm{K}} x_{\mathrm{K}i} + e_i$$

  in the numerator of (5) to find that $Cov(\mathrm{Y}_i, \tilde{x}_{ki}) = \beta_k V(\tilde{x}_{ki})$

Omitted Variables Bias

- The omitted variables bias (OVB) formula connects regression coefficients in models with different controls
- Go long: regress wages on schooling, S_i, controlling for ability (A_i)

$$\mathrm{Y}_i = \alpha + \rho \mathrm{S}_i + A_i' \gamma + \varepsilon_i \tag{6}$$

- Ability is hard to measure. What if we omit it? The result is

$$\frac{Cov(\mathrm{Y}_i, \mathrm{S}_i)}{V(\mathrm{S}_i)} = \rho + \gamma' \delta_{As},$$

  where δ_{As} is the vector of coefficients from regressions of the elements of A_i on S_i . . .
  - *Short equals long plus the effect of omitted times the regression of omitted on included*
- Short equals long when omitted and included are uncorrelated
  - when included is a dummy, "no OVB"="covariate balance"

OVB in a Wage Equation

TABLE 3.2.1
Estimates of the returns to education for men in the NLSY

Controls:	(1) None	(2) Age Dummies	(3) Col. (2) and Additional Controls*	(4) Col. (3) and AFQT Score	(5) Col. (4), with Occupation Dummies
	.132	.131	.114	.087	.066
	(.007)	(.007)	(.007)	(.009)	(.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.

*Additional controls are mother's and father's years of schooling, and dummy variables for race and census region.

Checking the CIA in the DK Design: no OVB

	Dependent Variable					
	Own SAT score/100			Predicted log(Parental Income)		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	1.165	1.130	0.066	0.128	0.138	0.028
	(0.196)	(0.188)	(0.112)	(0.035)	(0.037)	(0.037)
Female		-0.367			0.016	
		(0.076)			(0.013)	
Black		-1.947			-0.359	
		(0.079)			(0.019)	
Hispanic		-1.185			-0.259	
		(0.168)			(0.050)	
Asian		-0.014			-0.060	
		(0.116)			(0.031)	
Other/Missing Race		-0.521			-0.082	
		(0.293)			(0.061)	
High School Top 10 Percent		0.948			-0.066	
		(0.107)			(0.011)	
High School Rank Missing		0.556			-0.030	
		(0.102)			(0.023)	
Athlete		-0.318			0.037	
		(0.147)			(0.016)	
Average SAT Score of Schools Applied To/100			0.777			0.063
			(0.058)			(0.014)
Sent Two Application			0.252			0.020
			(0.077)			(0.010)
Sent Three Applications			0.375			0.042
			(0.106)			(0.013)

BAD CONTROL

(finish up)

When More Isn't Better

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

- An algebraic fact, devoid of causal content
- Bad control creates selection bias
- A parable
  - College is randomly assigned: simple college comparisons are causal
  - College boosts earnings by \$500/week
  - College allows some to get better jobs, specifically, to move from blue to white collar employment
- College changes the conditional-on-occ composition of the workforce
  - *The white collar group of non-college grads includes only AW*
  - *The white collar group of college grads includes some BW's, who are weaker than the AWs*
    - Flip it for the blues: blue non-college include some who could be white

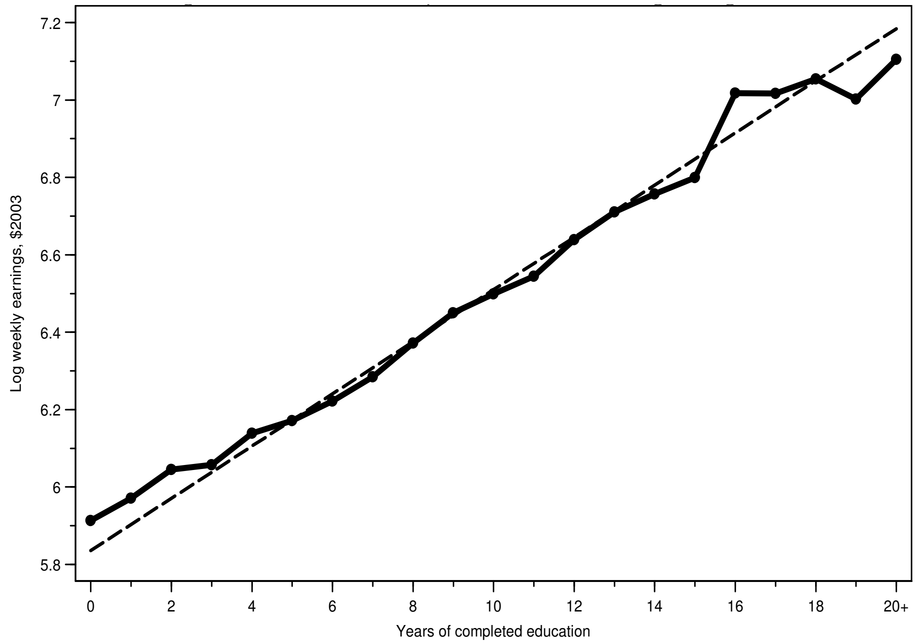
TABLE 6.1
How bad control creates selection bias

Type of worker	Potential occupation: Without college (1)	Potential occupation: With college (2)	Potential earnings: Without college (3)	Potential earnings: With college (4)	Average earnings by occupation: Without college (5)	Average earnings by occupation: With college (6)
Always Blue (AB)	Blue	Blue	1,000	1,500	Blue 1,500	Blue 1,500
Blue White (BW)	Blue	White	2,000	2,500		White 3,000
Always White (AW)	White	White	3,000	3,500	White 3,000	

Lessons Learned

- Regression always makes sense ... in the sense that it provides a best-in-class linear approximation to the CEF
- Regression is a matchmaker; regression is matching
    - MFX from non-linear models are usually indistinguishable from the corresponding regression estimates (MHE 3.4.2)
- We're not always content to run regressions, but that's where we start
    - Our first line of attack on a non-RCT identification problem: it's all about control
- If the regression you've got is not the one you want, that's because the underlying relationship is unsatisfactory
- What's to be done with an unsatisfactory relationship?
    - Move on, grasshopper ... to IV!

Tables and Figures



Sample is limited to white men, age 40-49. Data is from Census IPUMS 1980, 5% sample.

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and