Advanced Quantitative Methods

OLS as Matchmaker

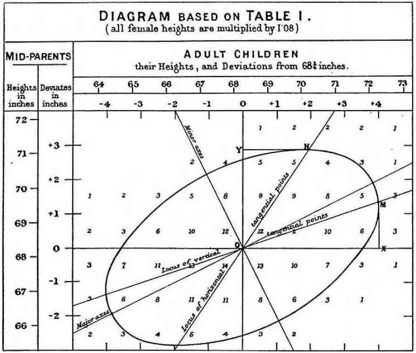
Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Today

- Ordinary Least Squares (OLS)
  - What do controls buy us?
- Implementing OLS in R

Origins

Galton, F. (1886). "Regression towards mediocrity in hereditary stature." *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246-263.



Building intuition with a simple empirical case

- Are there economic returns to attending a private (rather than public) university?

Theory

Could be higher economic returns to private education because:

- **Smaller class sizes**
  - Mechanism: more one-on-one time with instructors
- **More distinguished faculty**
  - Mechanism: Faculty are better teachers; faculty have more connections to employers
- **More intelligent peers**
  - Mechanism: Opportunity to learn from peers; more competitive increases effort

The experimental ideal:

- Random assignment to treatment: randomly assign students to attend a private or public university
- Why? Because we want to compare effectively equivalent people (or districts, institutions, states, time periods etc.)
- And we can't examine the effect of public versus private for each individual because of the *fundamental problem of causal inference*

The problem

- Unless we have control over treatment assignment, we need to account for the fact that those in one group will be different from those in another group for *other* reasons

The problem

- A simple comparison between those who went to public and private university might show a difference in earnings, but for reasons unrelated to which university those in each group attended

The key motivating assumption (although a heroic one)

- "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."

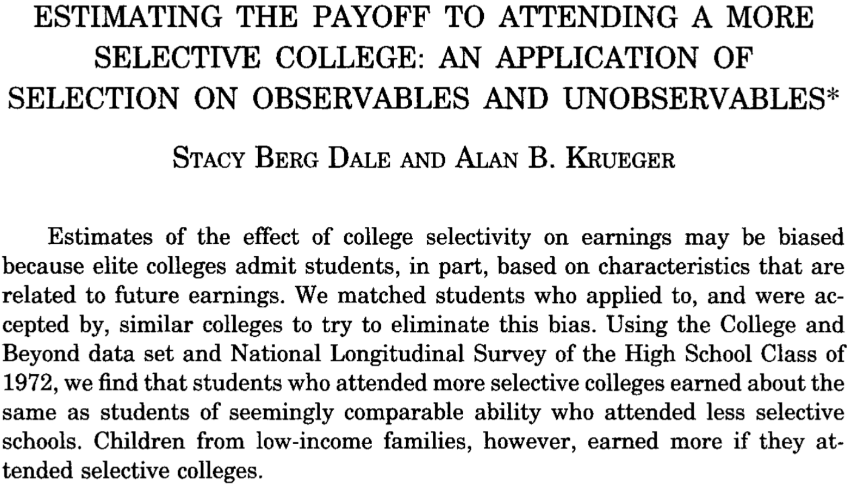
Potential assumption violations

Differences in earnings between those who went to private and public school might arise because of:

- Differences in ambition
- Differences in family income
- Differences in intelligence
- Differences in socio-demographics

Consequences

- A simple difference in means between those who went to private schools and those who went to public school is $10,000
- But is this difference *caused* by getting a private school education?



ESTIMATING THE PAYOFF TO ATTENDING A MORE SELECTIVE COLLEGE: AN APPLICATION OF SELECTION ON OBSERVABLES AND UNOBSERVABLES*

STACY BERG DALE AND ALAN B. KRUEGER

Estimates of the effect of college selectivity on earnings may be biased because elite colleges admit students, in part, based on characteristics that are related to future earnings. We matched students who applied to, and were accepted by, similar colleges to try to eliminate this bias. Using the College and Beyond data set and National Longitudinal Survey of the High School Class of 1972, we find that students who attended more selective colleges earned about the same as students of seemingly comparable ability who attended less selective schools. Children from low-income families, however, earned more if they attended selective colleges.

Empirical strategy (Dale & Krueger 2002)

- Compare those who *applied* and *got into* the exact same schools
- Why? Because those students should be extremely similar on many unobserved characteristics

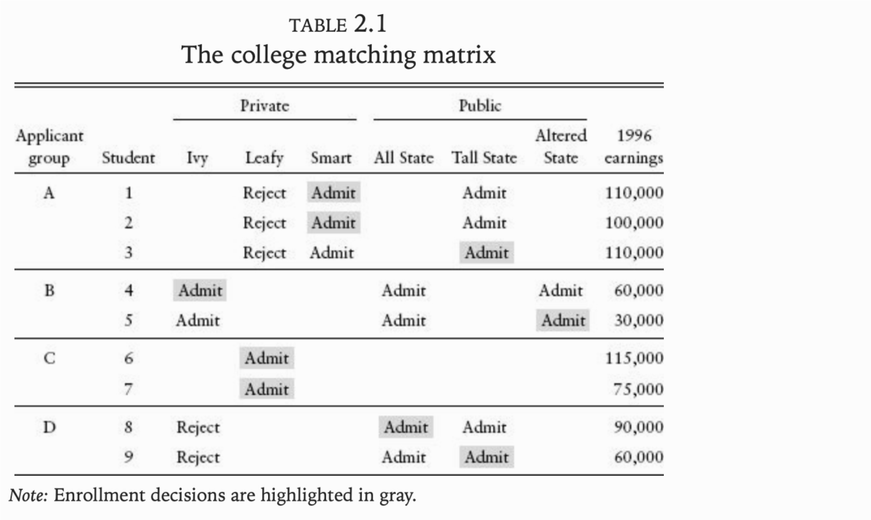


TABLE 2.1
The college matching matrix

		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.

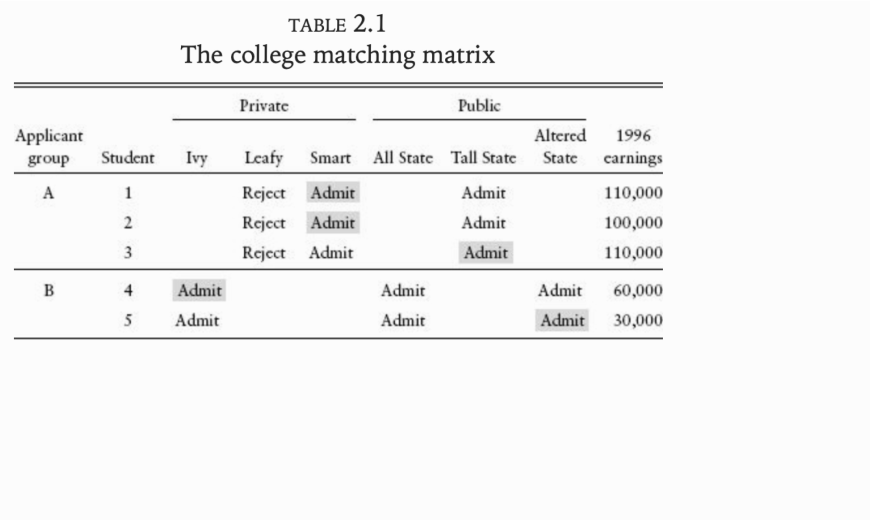


TABLE 2.1
The college matching matrix

		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

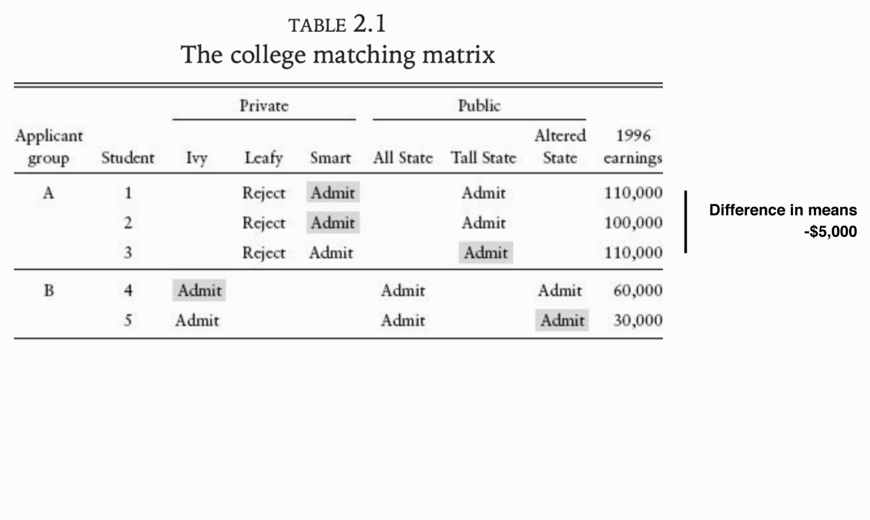


TABLE 2.1
The college matching matrix

		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

Difference in means
-$5,000

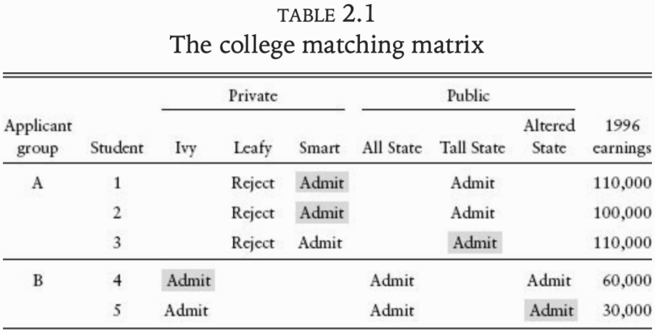


TABLE 2.1
The college matching matrix

		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

Difference in means
-$5,000

Difference in means
$30,000

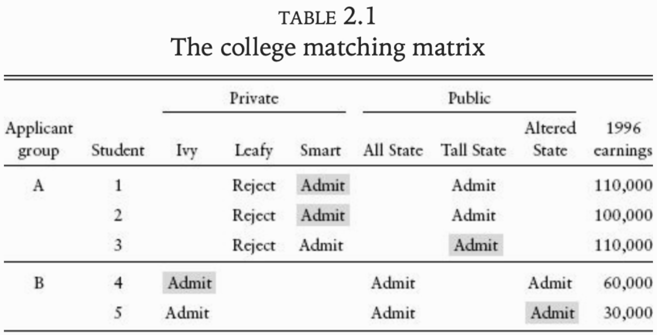


TABLE 2.1
The college matching matrix

		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

Difference in means
-$5,000

Difference in means
$30,000

Controlled difference
(-$5,000 + $30,000) / 2
= $12,500

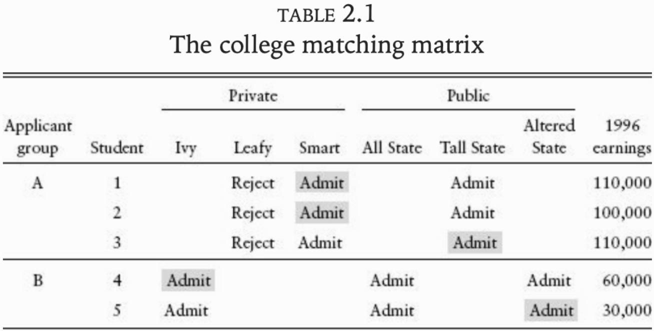


TABLE 2.1
The college matching matrix

		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

Difference in means
-$5,000

Difference in means
$30,000

Uncontrolled difference
Private: ($110,000 + $100,000 + $60,000) / 3 = $90,000
Public: ($110,000 + $30,000) / 2 = $70,000
$20,000

Controlled difference
(-$5,000 + $30,000) / 2
= $12,500

But:

- We often want to adjust for multiple variables
- Data are often sparse (or continuous) such that perfect matches on all control variables is effectively impossible

Regression as an approximate match maker

1. Y_i as the earnings Y for student i
   - Often called a "dependent variable", "outcome variable", or "response variable"

Regression as an approximate match maker

1. Y_i as the earnings Y for student i
   - Often called a “dependent variable”, “outcome variable”, or “response variable”

2. $T_i \in \{0, 1\}$ as whether or not student i went to public ($T_i = 0$) or private ($T_i = 1$) school.
   - Often called the “treatment variable” (often denoted T_i in political science, D_i in economics, and W_i in statistics)

Regression as an approximate match maker

1. Y_i as the earnings Y for student i
   - Often called a "dependent variable", "outcome variable", or "response variable"
2. $T_i \in \{0, 1\}$ as whether or not student i went to public ($T_i = 0$) or private ($T_i = 1$) school.
   - Often called the "treatment variable" (often denoted T_i in political science, D_i in economics, and W_i in statistics)
3. X_i as values of the control variable(s) for student i
   - "The controls"
   - Also thought of as the "independent variables" (conceptually, the treatment is too)

Regression as an approximate match maker

$$Y_i = \alpha + \tau T_i + \beta X_i + \epsilon_i,$$

Regression as an approximate match maker

$$Y_i = \alpha + \tau T_i + \beta X_i + \epsilon_i,$$

Data (Roman letters: what we have)	
Y_i	earnings of students *i*
T_i	treatment indicator for students *i*
X_i	control variable(s) for students *i*

Regression as an approximate match maker

$$Y_i = \alpha + \tau T_i + \beta X_i + \epsilon_i,$$

Data (Roman letters: what we have)	
Y_i	earnings of students i
T_i	treatment indicator for students i
X_i	control variable(s) for students i

Parameters (Greek letters: what we want to estimate)	
α	a constant
τ	the treatment effect
β	the relationship between the controls and the outcome

What is OLS actually doing?

It's minimizing the error in what the regression equation *predicts* earnings to be ($\hat{Y}_i$), and what it actually is (Y_i).

$$\hat{Y}_i = \alpha + \tau T_i + \beta X_i,$$

The error, e_i, of that prediction is simply:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \tau T_i + \beta X_i)$$

What is the error?

- Omitted variables
- Measurement error
- Random variation

What is OLS actually doing?

A prediction error in a OLS regression model is called a ‘residual’, and in OLS we square the error (hence Ordinary Least *Squares*):

$$e_i^2 = [Y_i - (\alpha + \tau T_i + \beta X_i)]^2$$

What is OLS actually doing?

Now imagine that we calculate the prediction errors for every person in our dataset and add up all the errors. This is called the residual *sum* of squares (RSS):

$$RSS = \sum_{i=1}^{n} [Y_i - (\alpha + \tau T_i + \beta X_i)]^2 = \sum_{i=1}^{n} e_i^2$$

What is OLS actually doing?

OLS regression is designed to find the values of the parameters (e.g., α, τ, and β) that *minimize* the sum of the squared errors (i.e. minimize the "residual sum of squares").

Example

$$Y_i = \alpha + \tau T_i + \epsilon_i$$

$$\alpha = 5000, \tau = 10000$$

$$\hat{Y}_i = 5000 + 10000\, T_i$$

i	Earnings (Y_i)	Private school (T_i)	Prediction ($\hat{Y}_i$)	Error (e_i)	Squared residual
1	27000	1	15000	12000	12000^2
2	28000	1	15000	13000	13000^2
3	21000	1	15000	6000	6000^2
4	22000	1	15000	7000	7000^2
5	19000	0	5000	14000	14000^2
6	21000	0	5000	16000	16000^2
7	23000	0	5000	18000	18000^2
			RSS $= 12000^2 + 13000^2 + \ldots\ 18000^2$		

Example

$$Y_i = \alpha + \tau T_i + \epsilon_i$$

$$\alpha = 20000, \tau = 5000$$

$$\hat{Y}_i = 20000 + 5000 T_i$$

i	Earnings (Y_i)	Private school (T_i)	Prediction ($\hat{Y}_i$)	Error (e_i)	Squared residual
1	27000	1	25000	2000	2000^2
2	28000	1	25000	3000	3000^2
3	21000	1	25000	-4000	-4000^2
4	22000	1	25000	-3000	-3000^2
5	19000	0	20000	-1000	-1000^2
6	21000	0	20000	1000	1000^2
7	23000	0	20000	3000	3000^2
					$\text{RSS} = 2000^2 + 3000^2 + \ldots\ 3000^2$

So what?

The upshot, is that if we have chosen our controls (X_i) correctly (a big *if*!), we can give a *causal* interpretation to the parameter we care about, τ, i.e. the effect on earnings of attending a private university rather than public university.

What happens to our estimate when a control is not included?

Short regression:

$$Y_i = \alpha^S + \tau^S T_i + \epsilon_i$$

Long regression (i.e. includes control):

$$Y_i = \alpha^L + \tau^L T_i + \beta X_i + \epsilon_i$$

What happens to our estimate when a control is not included?

$$\tau^S - \tau^L = \beta \times \pi_1,$$

where π_1 captures the relationship between the treatment and the control:

$$X_i = \pi_0 + \pi_1 T_i + \epsilon_i$$

Why should I care? Because a control only matters if it is correlated with both the treatment *and* the outcome.

Alternative notation

Conditional expectation:

$$E[Y_i|T_i, X_i],$$

Read as: “The expected value of Y conditional on the treatment and controls.”

$$E[Y_i|T_i = 1, X_i] - E[Y_i|T_i = 0, X_i] = \tau$$

How did Dale & Krueger (2002) do this?

Without controls:

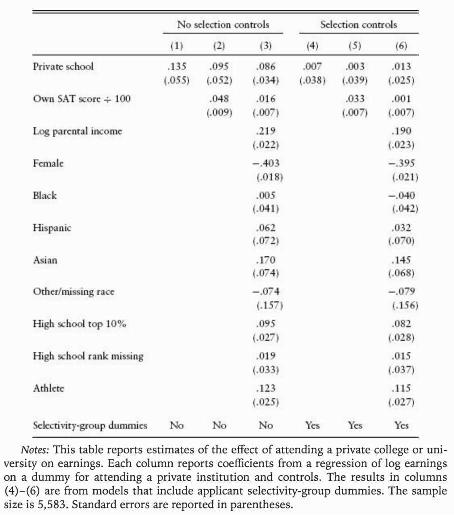
$$\ln Y_i = \alpha + \tau T_i + \epsilon_i$$

With controls (see p. 61):

$$\ln Y_i = \alpha + \tau T_i + \underbrace{\sum_{j=1}^{150} \gamma_j GROUP_{ij} + \delta_1 \ln SAT_i + \delta_2 \ln PI_i}_{\textbf{Control variables}} + \epsilon_i,$$

where $GROUP_{ij}$ is the matched school admission group, SAT_i is a standardized test score, and PI_i is parental income.

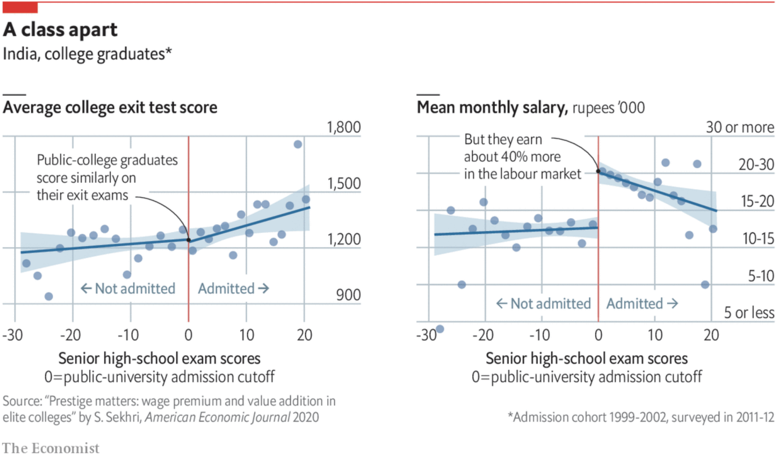
Dale & Krueger (2002)



	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135 (.055)	.095 (.052)	.086 (.034)	.007 (.038)	.003 (.039)	.013 (.025)
Own SAT score ÷ 100		.048 (.009)	.016 (.007)		.033 (.007)	.001 (.007)
Log parental income			.219 (.022)			.190 (.023)
Female			−.403 (.018)			−.395 (.021)
Black			.005 (.041)			−.040 (.042)
Hispanic			.062 (.072)			.032 (.070)
Asian			.170 (.074)			.145 (.068)
Other/missing race			−.074 (.157)			−.079 (.156)
High school top 10%			.095 (.027)			.082 (.028)
High school rank missing			.019 (.033)			.015 (.037)
Athlete			.123 (.025)			.115 (.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.

A quick aside about Indian test scores



The value is about more than just learning

A class apart

India, college graduates*

Average college exit test score

Public-college graduates score similarly on their exit exams

Mean monthly salary, rupees '000

But they earn about 40% more in the labour market

Source: "Prestige matters: wage premium and value addition in elite colleges" by S. Sekhri, *American Economic Journal* 2020

*Admission cohort 1999-2002, surveyed in 2011-12

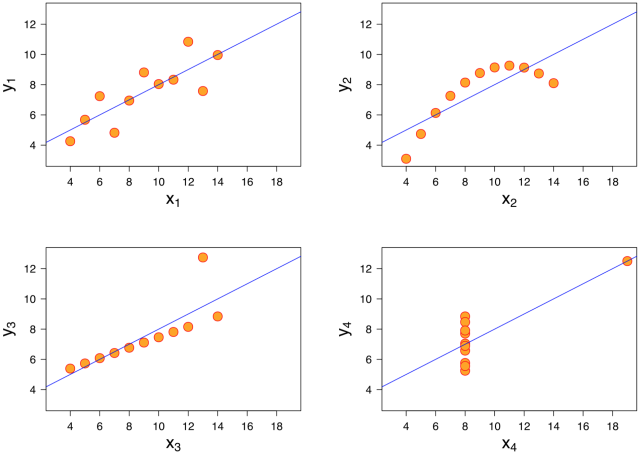
The Economist

1,800
1,500
1,200
900
← Not admitted
Admitted →
-30
-20
-10
0
10
20
Senior high-school exam scores
0=public-university admission cutoff
30 or more
20-30
15-20
10-15
5-10
5 or less
← Not admitted
Admitted →
-30
-20
-10
0
10
20
Senior high-school exam scores
0=public-university admission cutoff

Typical problems with regression

- Omitted variable bias
- Control for post-treatment variables (more later)
- Outliers
- Multi-collinearity
- Non-linear “functional form”

Example of functional form problems



Implementing OLS regression in R

```
# Simple model with a treatment variable and controls
model_1 <- lm(y ~ t + x1 + x2 + x3, data = D)
summary(model_1)
```

We won’t need to use non-linear models in this class

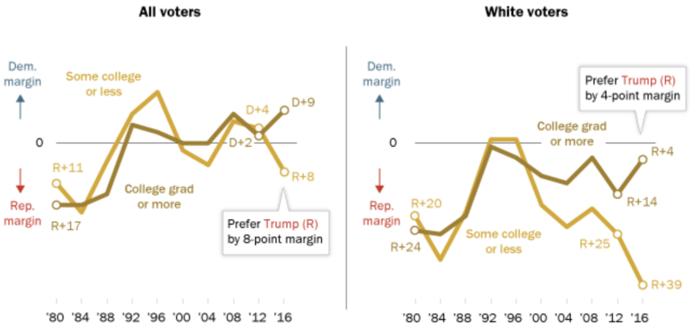
- e.g. Logistic regression

  -
    ```
    model_1 <- glm(y ~ x + z,
                    data = D, family = binomial)
    ```
- e.g. Poisson

  -
    ```
    model_2 <- glm(y ~ x + z,
                    data = D, family = poisson)}
    ```

Explaining the vote for Trump in the 2016 US Presidential election

Did low education affect the vote for Trump?



Wide education gaps in 2016 preferences, among all voters and among whites

Presidential candidate preference, by educational attainment

Source: Based on exit polls conducted by Edison Research for the National Election Pool, as reported by CNN. Data from prior years from national exit polls. In 1980, race was coded by the interviewer instead of being asked of the respondent.

PEW RESEARCH CENTER

Mutz (2018) on education & status threat

“[R]egardless of which outcome measures I examined, including indicators of economic status did not eliminate the impact of education. … However, after the relationship between Trump support and perceived status threat is taken into account, even lack of a college education no longer predicts Trump support for any of the measures. These findings strongly suggest that group-based status threat was the main reason that those without college educations were more supportive of Trump.”

Mutz (2018) on education & status threat

"[T]hese results speak to the importance of group status in the formation of political preferences. Political uprisings are often about downtrodden groups rising up to assert their right to better treatment ... The 2016 election, in contrast, was an effort by members of already dominant groups to assure their continued dominance."

Mutz (2018) on education & status threat

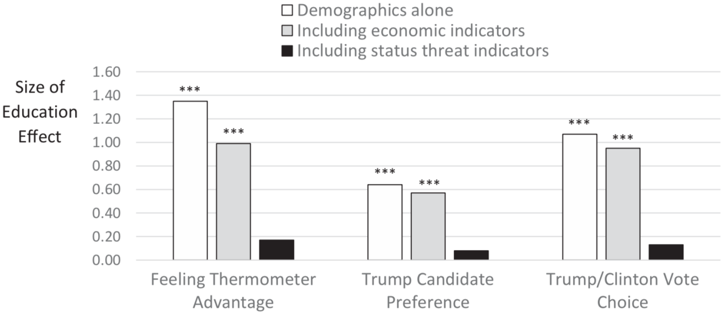


Fig. 3. Status threat accounts for the impact of education on the 2016 presidential election. Note that bars represent the predictive strength of education on each of three different outcome measures after taking into account (*i*) demographics alone, (*ii*) demographics and economic predictors only, and (*iii*) demographics and threat indicators only. Details are in Table S5. ***$P < 0.001$.

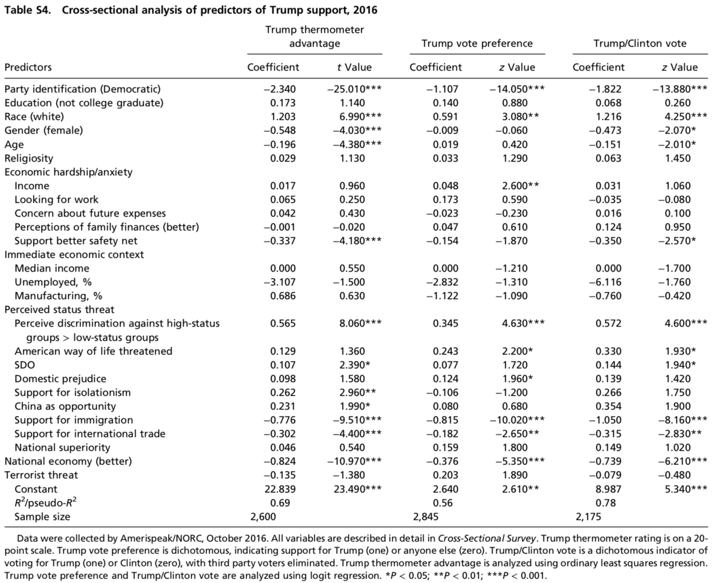


Table S4. Cross-sectional analysis of predictors of Trump support, 2016

	Trump thermometer advantage		Trump vote preference		Trump/Clinton vote	
Predictors	Coefficient	*t* Value	Coefficient	*z* Value	Coefficient	*z* Value
Party identification (Democratic)	−2.340	−25.010***	−1.107	−14.050***	−1.822	−13.880***
Education (not college graduate)	0.173	1.140	0.140	0.880	0.068	0.260
Race (white)	1.203	6.990***	0.591	3.080**	1.216	4.250***
Gender (female)	−0.548	−4.030***	−0.009	−0.060	−0.473	−2.070*
Age	−0.196	−4.380***	0.019	0.420	−0.151	−2.010*
Religiosity	0.029	1.130	0.033	1.290	0.063	1.450
Economic hardship/anxiety						
Income	0.017	0.960	0.048	2.600**	0.031	1.060
Looking for work	0.065	0.250	0.173	0.590	−0.035	−0.080
Concern about future expenses	0.042	0.430	−0.023	−0.230	0.016	0.100
Perceptions of family finances (better)	−0.001	−0.020	0.047	0.610	0.124	0.950
Support better safety net	−0.337	−4.180***	−0.154	−1.870	−0.350	−2.570*
Immediate economic context						
Median income	0.000	0.550	0.000	−1.210	0.000	−1.700
Unemployed, %	−3.107	−1.500	−2.832	−1.310	−6.116	−1.760
Manufacturing, %	0.686	0.630	−1.122	−1.090	−0.760	−0.420
Perceived status threat						
Perceive discrimination against high-status groups > low-status groups	0.565	8.060***	0.345	4.630***	0.572	4.600***
American way of life threatened	0.129	1.360	0.243	2.200*	0.330	1.930*
SDO	0.107	2.390*	0.077	1.720	0.144	1.940*
Domestic prejudice	0.098	1.580	0.124	1.960*	0.139	1.420
Support for isolationism	0.262	2.960**	−0.106	−1.200	0.266	1.750
China as opportunity	0.231	1.990*	0.080	0.680	0.354	1.900
Support for immigration	−0.776	−9.510***	−0.815	−10.020***	−1.050	−8.160***
Support for international trade	−0.302	−4.400***	−0.182	−2.650**	−0.315	−2.830**
National superiority	0.046	0.540	0.159	1.800	0.149	1.020
National economy (better)	−0.824	−10.970***	−0.376	−5.350***	−0.739	−6.210***
Terrorist threat	−0.135	−1.380	0.203	1.890	−0.079	−0.480
Constant	22.839	23.490***	2.640	2.610**	8.987	5.340***
R^2/pseudo-R^2	0.69		0.56		0.78	
Sample size	2,600		2,845		2,175	

Data were collected by Amerispeak/NORC, October 2016. All variables are described in detail in *Cross-Sectional Survey*. Trump thermometer rating is on a 20-point scale. Trump vote preference is dichotomous, indicating support for Trump (one) or anyone else (zero). Trump/Clinton vote is a dichotomous indicator of voting for Trump (one) or Clinton (zero), with third party voters eliminated. Trump thermometer advantage is analyzed using ordinary least squares regression. Trump vote preference and Trump/Clinton vote are analyzed using logit regression. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

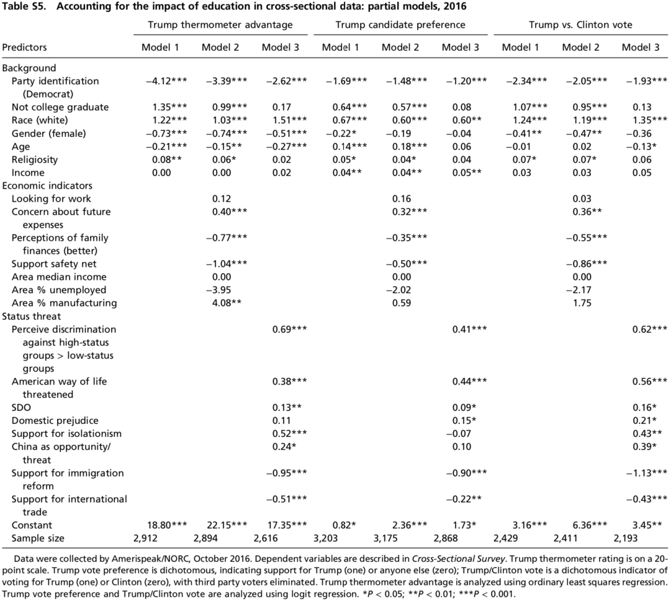


Table S5. Accounting for the impact of education in cross-sectional data: partial models, 2016

	Trump thermometer advantage			Trump candidate preference			Trump vs. Clinton vote		
Predictors	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Background									
Party identification (Democrat)	−4.12***	−3.39***	−2.62***	−1.69***	−1.48***	−1.20***	−2.34***	−2.05***	−1.93***
Not college graduate	1.35***	0.99***	0.17	0.64***	0.57***	0.08	1.07***	0.95***	0.13
Race (white)	1.22***	1.03***	1.51***	0.67***	0.60***	0.60**	1.24***	1.19***	1.35***
Gender (female)	−0.73***	−0.74***	−0.51***	−0.22*	−0.19	−0.04	−0.41**	−0.47**	−0.36
Age	−0.21***	−0.15**	−0.27***	0.14***	0.18***	0.06	−0.01	0.02	−0.13*
Religiosity	0.08**	0.06*	0.02	0.05*	0.04*	0.04	0.07*	0.07*	0.06
Income	0.00	0.00	0.02	0.04**	0.04**	0.05**	0.03	0.03	0.05
Economic indicators									
Looking for work		0.12			0.16			0.03	
Concern about future expenses		0.40***			0.32***			0.36**	
Perceptions of family finances (better)		−0.77***			−0.35***			−0.55***	
Support safety net		−1.04***			−0.50***			−0.86***	
Area median income		0.00			0.00			0.00	
Area % unemployed		−3.95			−2.02			−2.17	
Area % manufacturing		4.08**			0.59			1.75	
Status threat									
Perceive discrimination against high-status groups > low-status groups			0.69***			0.41***			0.62***
American way of life threatened			0.38***			0.44***			0.56***
SDO			0.13**			0.09*			0.16*
Domestic prejudice			0.11			0.15*			0.21*
Support for isolationism			0.52***			−0.07			0.43**
China as opportunity/ threat			0.24*			0.10			0.39*
Support for immigration reform			−0.95***			−0.90***			−1.13***
Support for international trade			−0.51***			−0.22**			−0.43***
Constant	18.80***	22.15***	17.35***	0.82*	2.36***	1.73*	3.16***	6.36***	3.45**
Sample size	2,912	2,894	2,616	3,203	3,175	2,868	2,429	2,411	2,193

Data were collected by Amerispeak/NORC, October 2016. Dependent variables are described in *Cross-Sectional Survey*. Trump thermometer rating is on a 20-point scale. Trump vote preference is dichotomous, indicating support for Trump (one) or anyone else (zero); Trump/Clinton vote is a dichotomous indicator of voting for Trump (one) or Clinton (zero), with third party voters eliminated. Trump thermometer advantage is analyzed using ordinary least squares regression. Trump vote preference and Trump/Clinton vote are analyzed using logit regression. **P* < 0.05; ***P* < 0.01; ****P* < 0.001.

Download the code and data

1. Download sample R code from the website's Weekly Readings
2. Download Mutz's (2018) replication data

Modify the sample R code to:

1. Reproduce Model 2 in Table S5
2. Reproduce Model 3 in Table S5
3. Output a regression table showing Models 1, 2, and 3 simultaneously (as Mutz does in Table S5) using the R library `modelsummary`

Exercise solution

```
# TABLE S5 MODEL 2
table_s5_m2 <- lm(cutdifftherm ~ party3 + noncollegegrad + white + GENDER +
                                 AGE7 + religion + INCOME +
                                 # Economic variables:
                                 lookwork + ecoworry + perecoperc +
                                 safetynet + medianincome,
                                 data = D)
```

```
# TABLE S5 MODEL 3
table_s5_m3 <- lm(cutdifftherm ~ party3 + noncollegegrad + white + GENDER +
                                 AGE7 + religion + INCOME +
                                 majorindex + pt4r + sdoindex + prejudice +
                                 isoindex + china + immigindex + tradeindex,
                                 data = D)
```

```
# Display all three models with modelsummary()
modelsummary(list(table_s5_m1, table_s5_m2, table_s5_m3),
             estimate = "{estimate} (t={statistic}{stars})", fmt = 2)
```