PSC 202

SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

MULTIPLE REGRESSION IN PRACTICE

LOGISTICS

- This week: More on multiple regression, experiments
  - Friday: PS 8 and Section Assignment due
- Next week
  - Monday: More on experiments, exam review
  - Wednesday: Exam 3
- Finals week
  - No in-class final exam
  - Instead: Problem Set 9 due on 12/15 (counts double)

HURDLES TO CAUSALITY

- Is there a credible causal mechanism that connects X to Y?
- Can we rule out the possibility that Y could cause X?
- Is there covariation between X and Y?
- Have we controlled for all confounding variables (Z) that might make the association between X and Y spurious?

WHAT THIS ALLOWS US TO DO

- Multiple regression is a tool that allows us to tackle the fourth hurdle to causality
  - Multiple regression can estimate effect of X on Y controlling for *all* confounders we can think of (Z_1, Z_2, etc.)
- $y = a + b_1*x + b_2*z_1 + b_3*z_2 + b_4*z_3$

WHAT THIS ALLOWS US TO DO

- Example: What determines how students in this class think about Joe Biden?
  - One thing we found looking at bivariate relation: Liberals like him more than conservatives (duh)
  - Does this relationship hold when controlling for other potential independent variables?
  - And what other independent variables can help explain variation in attitudes towards Biden?

R-REGRESSION

	Coefficient	Standard Error	T-Value
Intercept	95.2	61.9	1.54
Liberal-Conservative	-0.33	0.13	-2.46
Age	-1.55	3.13	-0.50

R^2: 0.10

EFFECT OF LIB/CONS

	Coefficient	Standard Error	T-Value
Intercept	95.2	61.9	1.54
Liberal-Conservative	-0.33	0.13	-2.46
Age	-1.55	3.13	-0.50

R^2: 0.10

EFFECT OF LIB/CONS

- Coefficient: -0.33
- Interpretation: For every one point increase on the liberal-conservative scale, the evaluation of J. Biden decreases by 0.33 points, *holding all other variables constant*

TEST STATISTIC

- H_A: -0.33
- H_0: 0
- Standard Error: 0.13

$$t = \frac{H_A - H_0}{\text{Standard Error}}$$

$$t = \frac{-0.33 - 0.00}{0.13} = -2.54$$

- t-value in table slightly different due to rounding
- We reject H_0, so negative effect of liberal-conservative on evaluation is significant at the 5% level

EFFECT OF AGE

	Coefficient	Standard Error	T-Value
Intercept	95.2	61.9	1.54
Liberal-Conservative	-0.33	0.13	-2.46
Age	-1.55	3.13	-0.50

R^2: 0.10

EFFECT OF AGE

- Coefficient: -1.55
- Interpretation: For every one year increase in age, the evaluation of J. Biden decreases by 1.55 points, *holding all other variables constant*

TEST STATISTIC

- **H_A: -1.55**
- **H_0: 0**
- **Standard Error: 3.13**

$$t = \frac{H_A - H_0}{\text{Standard Error}}$$

$$t = \frac{-1.55 - 0.00}{3.13} = -0.50$$

- **We cannot reject H_0, so effect of age on evaluation is *not* significant at the 5% level**

LINEAR REGRESSION

- So far: The independent variables were interval-level
  - Liberal-conservative and age
- What if independent variable is nominal or ordinal-level?
  - e.g. effect of gender

DUMMY VARIABLE REGRESSION

- Nominal or ordinal-level independent variable can easily be incorporated in linear regression
- $y = a + b_1*x_1 + b_2*x_2 + b_3*x_3$
  - x_3=0 if gender=female
  - x_3=1 if gender=male
- Same idea as before, but x_3 can only take values of 0 or 1
- "Dummy variable"
  - 0/1

REGRESSION

	Coefficient	Standard Error	T-Value
Intercept	101.8	60.8	1.68
Liberal-Conservative	-0.44	0.15	-3.07
Age	-1.89	3.08	-0.62
Gender (Male)	11.66	6.29	1.85

R^2: 0.15

EFFECT OF GENDER

- Coefficient: 11.66 (SE 6.29, t-value 1.85)
  - Where female is coded 0 and male coded 1
- Interpretation: If someone is male, their evaluation of J. Biden is expected to be 11.66 points higher than if someone is female, *holding all other variables constant*
- However, we do not reject H_0, so effect of gender on evaluation is not significant at the 5% level

SLIDERS AND SWITCHES

EFFECT OF GENDER

- Evaluation = 101.8 - 0.44*Lib/Cons - 1.89*Age + 11.66*Gender
  - female: gender=0
  - male: gender=1

EFFECT OF GENDER

- Evaluation = 101.8 - 0.44*Lib/Cons - 1.89*Age + 11.66*Gender
  - female: gender=0
  - male: gender=1
- What is the predicted evaluation for a person with a lib/cons score of 50, an age of 20, and who is male?

EFFECT OF GENDER

- Evaluation = 101.8 - 0.44*Lib/Cons - 1.89*Age + 11.66*Gender
  - female: gender=0
  - male: gender=1
- What is the predicted evaluation for a person with a lib/cons score of 50, an age of 20, and who is male?
- Evaluation = 101.8 - 0.44*50 - 1.89*20 + 11.66*1 = 53.66

EFFECT OF GENDER

- Evaluation = 101.8 - 0.44*Lib/Cons - 1.89*Age + 11.66*Gender
  - female: gender=0
  - male: gender=1
- What is the predicted evaluation for a person with a lib/cons score of 50, an age of 20, and who is female?

EFFECT OF GENDER

- Evaluation = 101.8 - 0.44*Lib/Cons - 1.89*Age + 11.66*Gender
  - female: gender=0
  - male: gender=1
- What is the predicted evaluation for a person with a lib/cons score of 50, an age of 20, and who is female?
- Evaluation = 101.8 - 0.44*50 - 1.89*20 + 11.66*0 = 42.0

WHAT THIS ALLOWS US TO DO

- Multiple regression is a tool that allows us to tackle the fourth hurdle to causality
  - Have we controlled for *all* confounding variables (Z) that might make the association between X and Y spurious?
  - We can now estimate effect of X on Y controlling for all confounders we can think of (Z_1, Z_2, etc.)

WHAT THIS ALLOWS US TO DO

- If we have not one theory about what influences Y, but many theories, we can test which one's have an effect on Y and which don't

TODAY

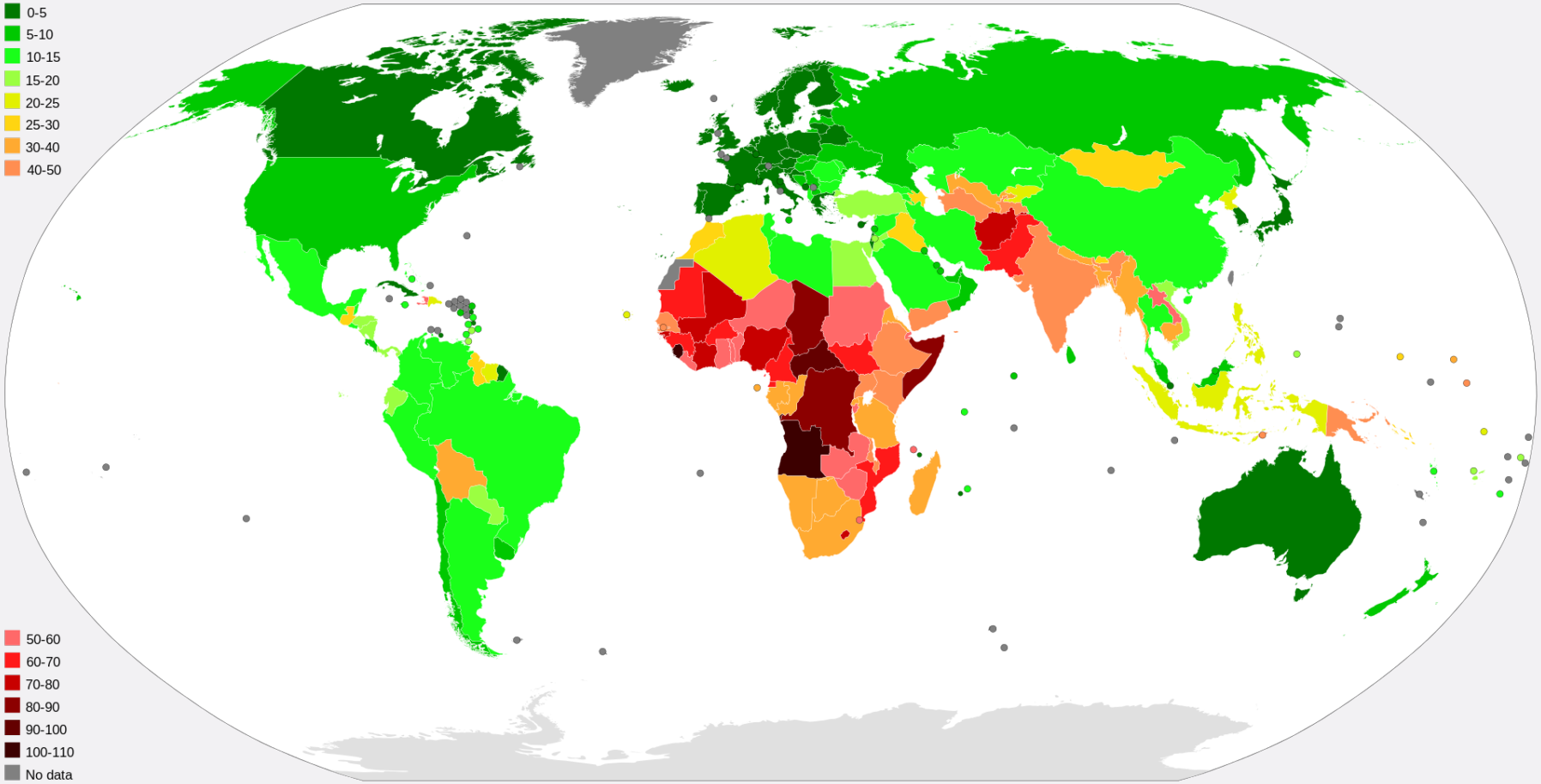
- Linear regression in research
- Linear regression in the private sector

LINEAR REGRESSION EXAMPLE

- Linear regression widely used in social science articles
  - Will show up in articles you read in your other political science classes

HOW IS THIS USEFUL?

- What causes high infant mortality rates?



- Infant mortality rates (Death under 1 year of age per 1,000 live births)

INFANT MORTALITY

- DV:
  - Death under 1 year of age per 1,000 live births
- IVs:
  - GDP per capita (logged)
  - Poverty: % of population living on less than $1.90 per day
  - Health expenditure: % of GDP
  - Clean water: % of population with access
  - Democracy: Index from -10 (least democratic) to 10 (most democratic)
  - Civil War: 0 if no, 1 if yes

REGRESSION

	Coefficient	Standard Error	T-Value
Intercept	88.25	34.31	2.57
Log Gdp Per Capita	-2.09	3.67	-0.57
Poverty	0.46	0.14	3.45
Health Expenditure	-0.21	1.05	-0.20
Clean Water	-0.57	0.23	-2.51
Democracy	-0.64	0.47	-1.37
Civil War	3.17	4.90	0.65

R^2: 0.79

EFFECT OF POVERTY

- Coefficient: 0.46 (SE 0.13, t-value 3.45)
- Interpretation: For every one percentage point increase of the population living in poverty, infant mortality increases by 0.46 deaths, holding all other variables constant

REGRESSION

	Coefficient	Standard Error	T-Value
Intercept	88.25	34.31	2.57
Log Gdp Per Capita	-2.09	3.67	-0.57
Poverty	0.46	0.14	3.45
Health Expenditure	-0.21	1.05	-0.20
Clean Water	-0.57	0.23	-2.51
Democracy	-0.64	0.47	-1.37
Civil War	3.17	4.90	0.65

R^2: 0.79

EFFECT OF CLEAN WATER

- Coefficient: -0.57 (SE 0.23, t-value -2.51)
- Interpretation: For every one percentage point increase of the population having access to clean water, infant mortality decreases by 0.57 deaths, holding all other variables constant

REGRESSION

	Coefficient	Standard Error	T-Value
Intercept	88.25	34.31	2.57
Log Gdp Per Capita	-2.09	3.67	-0.57
Poverty	0.46	0.14	3.45
Health Expenditure	-0.21	1.05	-0.20
Clean Water	-0.57	0.23	-2.51
Democracy	-0.64	0.47	-1.37
Civil War	3.17	4.90	0.65

R^2: 0.79

EFFECT OF CIVIL WAR

- Coefficient: 3.17 (SE 4.90, t-value 0.65)
- Interpretation: If a country has a civil war, its infant mortality increases by 3.17 deaths, holding all other variables constant
- However, we cannot reject H_0

ANOTHER EXAMPLE

ELSEVIER

Journal of Public Economics 87 (2003) 1801–1824

JOURNAL OF PUBLIC ECONOMICS

www.elsevier.com/locate/econbase

A free press is bad news for corruption

Aymo Brunetti[a], Beatrice Weder[b,*]

[a]*State Secretariat for Economic Affairs, Bern, Basel, Switzerland*
[b]*University of Mainz, Mainz, Germany*

Received 2 June 1999; received in revised form 13 June 2001; accepted 25 June 2001

- What is the effect of freedom of the press on corruption?

LINEAR REGRESSION EXAMPLE

- Unit of analysis: Countries
- Dependent variable: Corruption
- Independent variable: Press Freedom

LINEAR REGRESSION EXAMPLE

- H_A: In a comparison of countries, those with more press freedom will have lower levels of corruption than those with less press freedom
- H_0: There is no relationship between press freedom and levels of corruption

LINEAR REGRESSION EXAMPLE

- Data:
  - Corruption: Indicator by International Country Risk Guide, from 0 to 6
    - 0: a lot of corruption
    - 6: little corruption
  - Press freedom: Indicator by Freedom House, from 0 to 15
    - 0: no violations of press freedom
    - 15: highest degree of violations of press freedom

LINEAR REGRESSION EXAMPLE

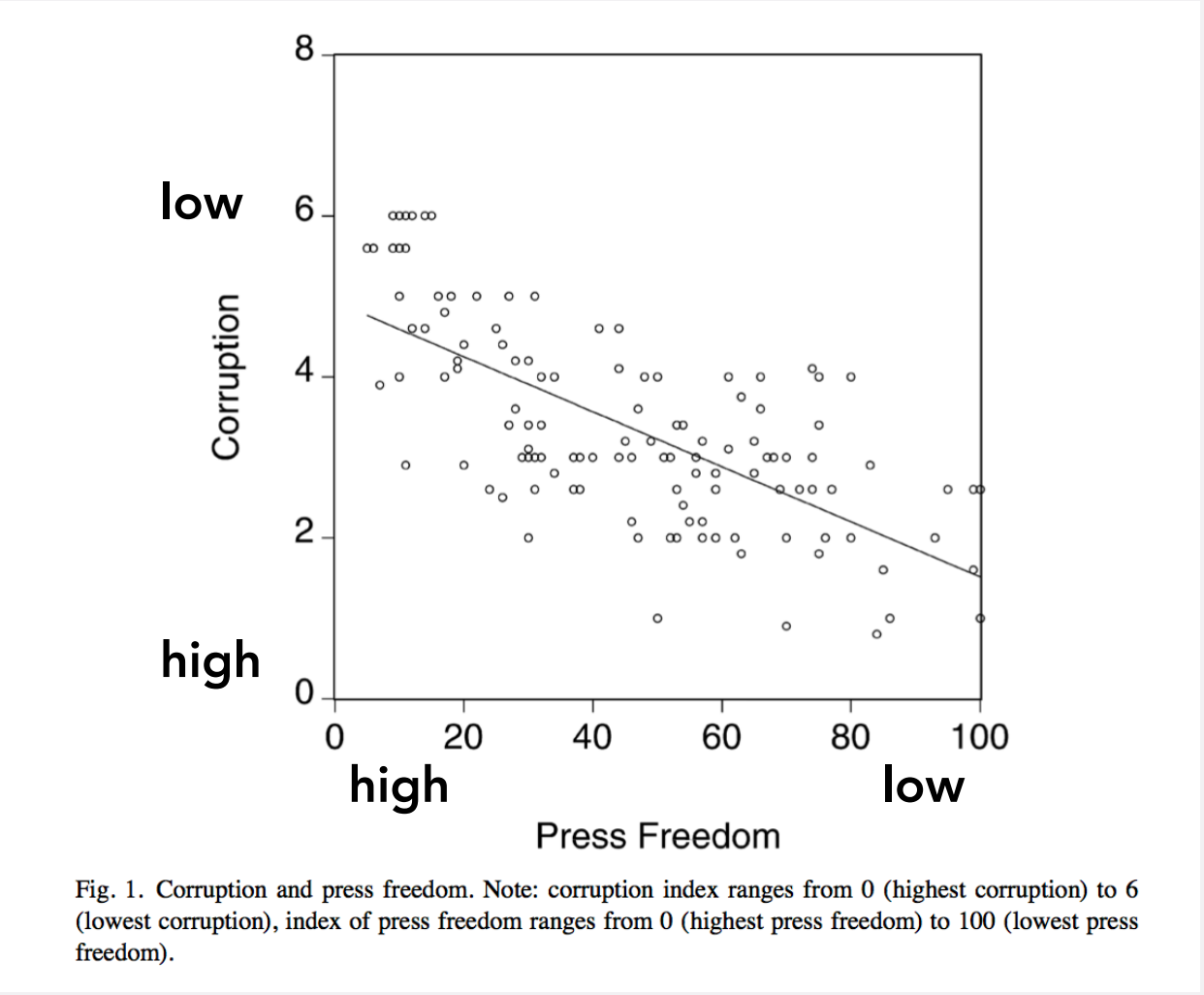


Fig. 1. Corruption and press freedom. Note: corruption index ranges from 0 (highest corruption) to 6 (lowest corruption), index of press freedom ranges from 0 (highest press freedom) to 100 (lowest press freedom).

- Does this correlation hold up when controlling for other variables that could affect corruption?

LINEAR REGRESSION EXAMPLE

3.3. Specification

As noted above, the theoretical and empirical literature have identified a number of determinants of corruption. On the one hand there are direct internal and external control mechanisms. On the other hand there are more indirect determinants such as distortions and sociological determinants of higher corruption. This suggests that estimates of corruption should at least include proxies for the direct control mechanisms which leads to our following preferred specification:

$$\mathrm{CORR}_i = \beta_0 + \beta_1 \mathrm{PRESS}_i + \beta_2 \mathrm{BUREAU}_i + \beta_3 \mathrm{RULE}_i + \varepsilon_i. \qquad (1)$$

- CORR: Corruption variable
- PRESS: Press freedom variable
- BUREAU: Quality of bureaucracy measure
- RULE: Measure of rule of law

LINEAR REGRESSION EXAMPLE

Table 1
Dependent variable: average corruption in 1994–1998

	(1) OLS	(2) OLS (LDCs only)	(3) TSLS	(4) OLS	(5) OLS (LDCs only)	(6) OLS	(7) TSLS
Constant	2.560	2.614	3.392	1.945	1.506	2.946	4.139
	(10.508)	(10.516)	(5.003)	(1.721)	(1.260)	(2.180)	(1.867)
PRESS	−0.017	−0.015	−0.028	−0.017	−0.015	−0.020	−0.037
	(−6.350)	(−4.789)	(−3.266)	(−4.023)	(−3.501)	(−4.439)	(−1.926)
BUREAU	0.220	0.254	0.221	0.200	0.128	0.089	0.073
	(2.893)	(2.708)	(2.310)	(2.058)	(1.220)	(0.942)	(0.663)
RULE	0.265	0.146	0.143	0.259	0.068	0.154	0.044
	(3.482)	(1.624)	(1.527)	(2.583)	(0.607)	(1.530)	(0.251)
log(GDP)				0.104	0.226	0.107	0.127
				(0.681)	(1.358)	(0.538)	(0.523)
HUMCAP				−0.043	−0.085	−0.052	−0.064
				(−1.007)	(−1.562)	(−1.058)	(−1.088)
TRADE				0.002	0.004	0.003	0.003
				(1.103)	(2.091)	(1.358)	(1.367)
BLACK				0.001	0.001	0.001	0.001
				(1.882)	(1.288)	(1.350)	(0.730)
ETHNIC				−0.246	−0.053	−0.457	−0.410
				(−0.690)	(−0.154)	(−1.170)	(−1.021)
AFRICA						−0.142	−0.102
						(−0.521)	(−0.252)
LATIN						−0.563	−0.857
						(−2.298)	(−2.530)
OECD						0.419	0.075
						(0.983)	(0.150)
Observations	125	93	104	68	47	68	68
Adjusted R^2	0.67	0.38	0.67	0.74	0.38	0.77	0.72

t Statistics in parentheses; White-corrected standard errors; political rights as instrument in Columns (3) and (7).

LINEAR REGRESSION EXAMPLE

Table 1
Dependent variable: average corruption in 1994–1998

	(1) OLS	(2) OLS (LDCs only)	(3) TSLS	(4) OLS	(5) OLS (LDCs only)	(6) OLS	(7) TSLS
Constant	2.560	2.614	3.392	1.945	1.506	2.946	4.139
	(10.508)	(10.516)	(5.003)	(1.721)	(1.260)	(2.180)	(1.867)
PRESS	−0.017	−0.015	−0.028	−0.017	−0.015	−0.020	−0.037
	(−6.350)	(−4.789)	(−3.266)	(−4.023)	(−3.501)	(−4.439)	(−1.926)
BUREAU	0.220	0.254	0.221	0.200	0.128	0.089	0.073
	(2.893)	(2.708)	(2.310)	(2.058)	(1.220)	(0.942)	(0.663)
RULE	0.265	0.146	0.143	0.259	0.068	0.154	0.044
	(3.482)	(1.624)	(1.527)	(2.583)	(0.607)	(1.530)	(0.251)
log(GDP)				0.104	0.226	0.107	0.127
				(0.681)	(1.358)	(0.538)	(0.523)
HUMCAP				−0.043	−0.085	−0.052	−0.064
				(−1.007)	(−1.562)	(−1.058)	(−1.088)
TRADE				0.002	0.004	0.003	0.003
				(1.103)	(2.091)	(1.358)	(1.367)
BLACK				0.001	0.001	0.001	0.001
				(1.882)	(1.288)	(1.350)	(0.730)
ETHNIC				−0.246	−0.053	−0.457	−0.410
				(−0.690)	(−0.154)	(−1.170)	(−1.021)
AFRICA						−0.142	−0.102
						(−0.521)	(−0.252)
LATIN						−0.563	−0.857
						(−2.298)	(−2.530)
OECD						0.419	0.075
						(0.983)	(0.150)
Observations	125	83	104	68	47	68	68
Adjusted R^2	0.67	0.38	0.67	0.74	0.38	0.77	0.72

t Statistics in parentheses; White-corrected standard errors; political rights as instrument in Columns (3) and (7).

LINEAR REGRESSION EXAMPLE

	(1) OLS
Constant	2.560
	(10.508)
PRESS	−0.017
	(−6.350)
BUREAU	0.220
	(2.893)
RULE	0.265
	(3.482)

- PRESS: Press freedom variable
- A one unit increase in the press freedom index (0-15, higher=less freedom) is associated with a 0.017 unit decrease in the corruption index (0-6, lower=more corruption), holding all other variables constant
- t-value is -6.35, so we can reject H_0

LINEAR REGRESSION EXAMPLE

Table 1
Dependent variable: average corruption in 1994–1998

	(1) OLS	(2) OLS (LDCs only)	(3) TSLS	(4) OLS	(5) OLS (LDCs only)	(6) OLS	(7) TSLS
Constant	2.560	2.614	3.392	1.945	1.506	2.946	4.139
	(10.508)	(10.516)	(5.003)	(1.721)	(1.260)	(2.180)	(1.867)
PRESS	−0.017	−0.015	−0.028	−0.017	−0.015	−0.020	−0.037
	(−6.350)	(−4.789)	(−3.266)	(−4.023)	(−3.501)	(−4.439)	(−1.926)
BUREAU	0.220	0.254	0.221	0.200	0.128	0.089	0.073
	(2.893)	(2.708)	(2.310)	(2.058)	(1.220)	(0.942)	(0.663)
RULE	0.265	0.146	0.143	0.259	0.068	0.154	0.044
	(3.482)	(1.624)	(1.527)	(2.583)	(0.607)	(1.530)	(0.251)
log(GDP)				0.104	0.226	0.107	0.127
				(0.681)	(1.358)	(0.538)	(0.523)
HUMCAP				−0.043	−0.085	−0.052	−0.064
				(−1.007)	(−1.562)	(−1.058)	(−1.088)
TRADE				0.002	0.004	0.003	0.003
				(1.103)	(2.091)	(1.358)	(1.367)
BLACK				0.001	0.001	0.001	0.001
				(1.882)	(1.288)	(1.350)	(0.730)
ETHNIC				−0.246	−0.053	−0.457	−0.410
				(−0.690)	(−0.154)	(−1.170)	(−1.021)
AFRICA						−0.142	−0.102
						(−0.521)	(−0.252)
LATIN						−0.563	−0.857
						(−2.298)	(−2.530)
OECD						0.419	0.075
						(0.983)	(0.150)
Observation	125	93	104	68	47	68	68
Adjusted R^2	0.67	0.38	0.67	0.74	0.38	0.77	0.72

t Statistics in parentheses; White-corrected standard errors; political rights as instrument in Columns (3) and (7).

LINEAR REGRESSION EXAMPLE

In addition, we test a second, broad specification which includes a number of the other potentially relevant determinants of corruption discussed above:

$$\begin{aligned} \text{CORR}_i = {} & \beta_0 + \beta_1 \text{PRESS}_i + \beta_2 \text{BUREAU}_i + \beta_3 \text{RULE}_i + \beta_4 \text{GDP}_i \\ & + \beta_5 \text{HUMCAP}_i + \beta_6 \text{TRADE}_i + \beta_7 \text{BLACK}_i + \beta_8 \text{ETHNIC}_i + \varepsilon_i. \end{aligned} \quad (2)$$

LINEAR REGRESSION EXAMPLE

Table 1
Dependent variable: average corruption in 1994–1998

	(1) OLS	(2) OLS (LDCs only)	(3) TSLS	(4) OLS	(5) OLS (LDCs only)	(6) OLS	(7) TSLS
Constant	2.560	2.614	3.392	1.945	1.506	2.946	4.139
	(10.508)	(10.516)	(5.003)	(1.721)	(1.260)	(2.180)	(1.867)
PRESS	−0.017	−0.015	−0.028	−0.017	−0.015	−0.020	−0.037
	(−6.350)	(−4.789)	(−3.266)	(−4.023)	(−3.501)	(−4.439)	(−1.926)
BUREAU	0.220	0.254	0.221	0.200	0.128	0.089	0.073
	(2.893)	(2.708)	(2.310)	(2.058)	(1.220)	(0.942)	(0.663)
RULE	0.265	0.146	0.143	0.259	0.068	0.154	0.044
	(3.482)	(1.624)	(1.527)	(2.583)	(0.607)	(1.530)	(0.251)
log(GDP)				0.104	0.226	0.107	0.127
				(0.681)	(1.358)	(0.538)	(0.523)
HUMCAP				−0.043	−0.085	−0.052	−0.064
				(−1.007)	(−1.562)	(−1.058)	(−1.088)
TRADE				0.002	0.004	0.003	0.003
				(1.103)	(2.091)	(1.358)	(1.367)
BLACK				0.001	0.001	0.001	0.001
				(1.882)	(1.288)	(1.350)	(0.730)
ETHNIC				−0.246	−0.053	−0.457	−0.410
				(−0.690)	(−0.154)	(−1.170)	(−1.021)
AFRICA						−0.142	−0.102
						(−0.521)	(−0.252)
LATIN						−0.563	−0.857
						(−2.298)	(−2.530)
OECD						0.419	0.075
						(0.983)	(0.150)
Observations	125	93	104	68	47	68	68
Adjusted R^2	0.67	0.38	0.67	0.74	0.38	0.77	0.72

t Statistics in parentheses; White-corrected standard errors; political rights as instrument in Columns (3) and (7).

WHAT YOU UNDERSTAND NOW

Table 1
Dependent variable: average corruption in 1994–1998

	(1) OLS	(2) OLS (LDCs only)	(3) TSLS	(4) OLS	(5) OLS (LDCs only)	(6) OLS	(7) TSLS
Constant	2.560	2.614	3.392	1.945	1.506	2.946	4.139
	(10.508)	(10.516)	(5.003)	(1.721)	(1.260)	(2.180)	(1.867)
PRESS	−0.017	−0.015	−0.028	−0.017	−0.015	−0.020	−0.037
	(−6.350)	(−4.789)	(−3.266)	(−4.023)	(−3.501)	(−4.439)	(−1.926)
BUREAU	0.220	0.254	0.221	0.200	0.128	0.089	0.073
	(2.893)	(2.708)	(2.310)	(2.058)	(1.220)	(0.942)	(0.663)
RULE	0.265	0.146	0.143	0.259	0.068	0.154	0.044
	(3.482)	(1.624)	(1.527)	(2.583)	(0.607)	(1.530)	(0.251)
log(GDP)				0.104	0.226	0.107	0.127
				(0.681)	(1.358)	(0.538)	(0.523)
HUMCAP				−0.043	−0.085	−0.052	−0.064
				(−1.007)	(−1.562)	(−1.058)	(−1.088)
TRADE				0.002	0.004	0.003	0.003
				(1.103)	(2.091)	(1.358)	(1.367)
BLACK				0.001	0.001	0.001	0.001
				(1.882)	(1.288)	(1.350)	(0.730)
ETHNIC				−0.246	−0.053	−0.457	−0.410
				(−0.690)	(−0.154)	(−1.170)	(−1.021)
AFRICA						−0.142	−0.102
						(−0.521)	(−0.252)
LATIN						−0.563	−0.857
						(−2.298)	(−2.530)
OECD						0.419	0.075
						(0.983)	(0.150)
Observations	125	93	104	68	47	68	68
Adjusted R^2	0.67	0.38	0.67	0.74	0.38	0.77	0.72

t Statistics in parentheses; White-corrected standard errors; political rights as instrument in Columns (3) and (7).

TODAY

- Linear regression in research
- Linear regression in the private sector

BIG DATA

The Age of Big Data

Big Data Comes to Dieting

50 Best Jobs in America for 2021

#2	Data Scientist	$113,736	4.1/5	5,971

How Big Data Became So Big

DATA

Data Scientist: The Sexiest Job of the 21st Century

LINEAR REGRESSION EXAMPLE

- Linear regression models also widely used by data analysts in private sector

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: “If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that? ”

LINEAR REGRESSION EXAMPLE

The desire to collect information on customers is not new for Target or any other large retailer, of course. For decades, Target has collected vast amounts of data on every person who regularly walks into one of its stores. Whenever possible, Target assigns each shopper a unique code — known internally as the Guest ID number — that keeps tabs on everything they buy. "If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we've sent you or visit our Web site, we'll record it and link it to your Guest ID," Pole said. "We want to know everything we can."

Also linked to your Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you've moved recently, what credit cards you carry in your wallet and what Web sites you visit. Target can buy data about your ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own. (In a statement, Target

LINEAR REGRESSION EXAMPLE

All that information is meaningless, however, without someone to analyze and make sense of it. That's where Andrew Pole and the dozens of other members of Target's Guest Marketing Analytics department come in.

Almost every major retailer, from grocery chains to investmentbanks to the U.S. Postal Service, has a "predictive analytics" department devoted to understanding not just consumers' shopping habits but also their personal habits, so as to more efficiently market to them. "But Target has always been one of the smartest at this," says Eric Siegel, a consultant and the chairman of a conference called Predictive Analytics World. "We're living through a golden age of behavioral research. It's amazing how much we can figure out about how people think now."

LINEAR REGRESSION EXAMPLE

- $y = a + b_1*x_1 + b_2*x_2 + b_3*x_3 + \ldots$
  - y: Pregnant or not?
  - x_1: $ spent on milk
  - x_2: $ spent on clothes
  - x_3: $ spent on vitamin supplements
  - …

LINEAR REGRESSION EXAMPLE

analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date.

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

LINEAR REGRESSION EXAMPLE

About a year after Pole created his pregnancy-prediction model, a man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

LINEAR REGRESSION EXAMPLE

On the phone, though, the father was somewhat abashed. “I had a talk with my daughter,” he said. “It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology.”

LINEAR REGRESSION RECAP

- Observational analysis
  - Takes data as we find it in the world
  - Regression tries to find the "data-generating process"
  - In the real world, what factors cause corruption to be higher or lower, and by how much?
  - Problem: We never know if we have controlled for *all* potential variables
- Next time: How can we get around this problem?