Political Science 15
Introduction to Research in Political Science
Lecture 8a: Types of Variables

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Housekeeping

Announcements

- Problem Set 4 is extended to Wednesday 22 July at 11:55pm.
    - Optional reading Chapter 12.1 to help you with extra credit question.
- Problem Set 5 is extended to Wednesday 29 July at 11:55pm.

To read

- From week 4, make sure you have finished Chapter 5 on Multivariate OLS.
- This week, for Lecture 7 on Multivariate OLS in Research, you need to have read Bateson (2012).
- Read Chapter 6 (sections 6.1-6.3 only) for Lecture 8 on Types of Variables and Data Sets (this one).
    - Optional reading Chapters 7-8 for this lecture (more info on logged variables in Ch. 7 and fixed effects in Ch. 8).
- For the end of this week and the start of week 6, where we tackle Lecture 9 on Experiments, read Chapter 10 and Gerber, Green, and Larimer (2008).

Recap

- In Lecture 6, we discussed how multivariate regression works, including:
  - how we use it to control for variables that may otherwise hang out in the error term
  - how to interpret and visualize it
  - how to use it for prediction
  - pitfalls to avoid

- In Lecture 7, we saw how Regina Bateson used multivariate regression to make causal claims, despite using observational data. Her careful approach to demonstrate that her results are robust included:
  - operationalizing the DV in different ways
  - running regressions on different samples
  - controlling for possible confounders
  - placebo test to rule out reverse causation

- Now, we continue our discussion of how to do research in practice, specifically, the types of data and variables that come up in empirical research.

Types of variables and data sets

- **Variables** in a data set come in different forms:
  - Dummy (AKA binary or dichotomous) variables
  - Discrete versus continuous variables
  - Ordinal variables
  - Nominal variables
  - Logged variables (which are one version of a transformed variable, see Chapter 7 of *Real Stats*)
- **Data sets** themselves also come in different forms:
  - Cross-sectional data
  - Time series data
  - Panel (cross-sectional & time series) data

Binary variables

- Binary (dummy) variables are *very* useful.
- For example: they are used in experiments to identify the treated (1) and control (0) units.
- Binary variables make difference in means across two groups, or Average Treatment Effects (ATE), very easy to calculate.
- But they can also be used to calculate the difference in effects (or mean outcomes) between groups, e.g. “Yes” and “No”, group “A” and group “B”.

Discrete vs. continuous data

Discrete data

Comes in “bins” or groups. Example: On a scale of 1 to 5, how much do you like dogs? 1, 2, 3, 4, 5. (5! Obv.) Polity score is another example.

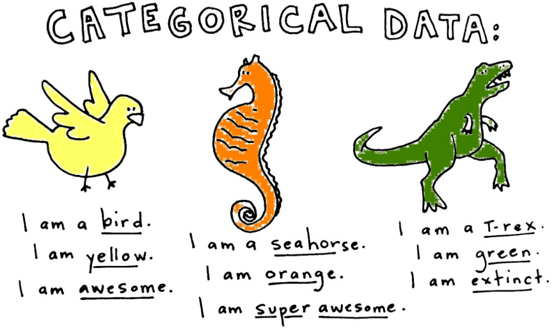
Continuous data

Can take any value in a sequence. Examples: Annual income, votes for each candidate, percentages, neck length of a giraffe.

CONTINUOUS
measured data, can have ∞ values within possible range.
I AM 3.1" TALL
I WEIGH 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I HAVE 8 LEGS
and
4 SPOTS!
@allison_horst

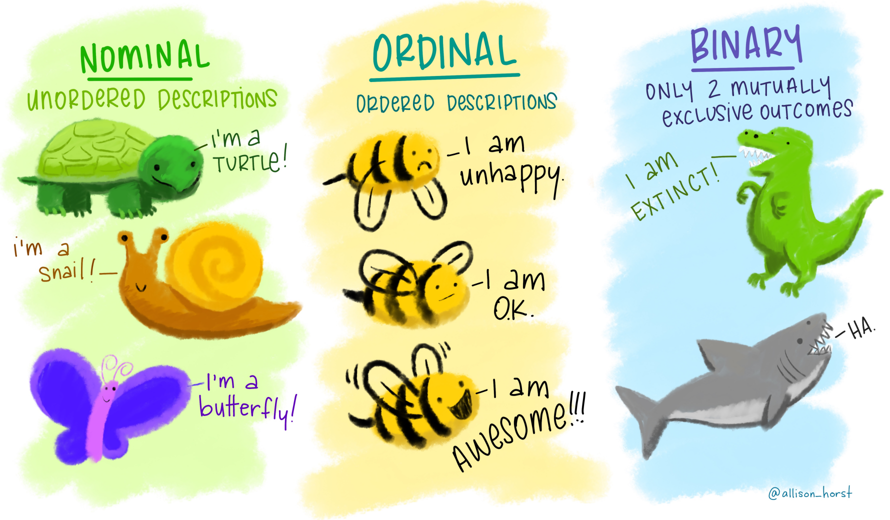
Categorical data

- It is **descriptive**: it describes how the world is. Often, categorical data comes from qualitative research.



- Categorical data can be ordinal or nominal. What does this mean? Ordinal can be ordered (low, medium, high) versus nominal, which cannot be ordered (majors: political science, economics, sociology).

Categorical data and discrete data



Transforming categorical data into discrete data in R

- We might want to turn a category into discrete or binary data, like treated versus control.
- In R, categorical data is called a “factor”.
- Factors come in different *levels*: e.g. types of cell phone bans (texting ban, outright ban, hands-free, etc.).
- Here are some commands in R that allow you to transform a categorical variable into a binary (dummy) variable:

```
class(data$variable)
```

to find out if something is a factor

```
lev <- levels(data$variable)
```

to extract the levels of that factor and store them as an object

```
ifelse(data$variable == lev[1] |
data$variable == lev[2], 1, 0)
```

if a variable is at *either* the first *or* second level, code it as a 1 (treated), otherwise code it as a 0 (control)

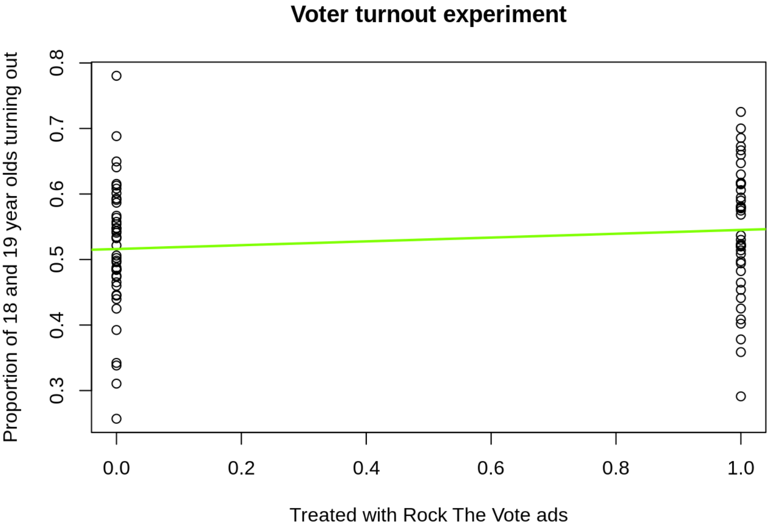
Political Science 15
Introduction to Research in Political Science
Lecture 8b: Using Dummy and Log Variables in Regression

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Using dummy variables in regression

Recall the “Rock the Vote” example from Section 3.



Using dummy variables in regression

```
Call:
lm(formula = turnout ~ treated, data = RockTheVote)

Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  0.51606      0.01510    34.170     <2e-16 ***
treated      0.02909      0.02149     1.354      0.179
```

- Check. What is the mean value of turnout in the control group?
  $\hat{\beta}_0 = 0.51606$, i.e. 51.6%.
- Check. What is the mean value of turnout in the treated group?
  $\hat{\beta}_0 + \hat{\beta}_1 = 0.51606 + 0.02909 = 0.54515$, i.e. 54.5%.

```
> mean(RockTheVote$turnout[RockTheVote$treated == 1])
[1] 0.54515
```

- So the coefficient $\hat{\beta}_1$ on the dummy variable `treated` represents the difference in mean turnout in the two groups!

Using dummy variables in regression

Works with multivariate too. Once we have turned a categorical variable into a **binary** variable, we can put it in our regression as an IV. This will calculate the difference in means between the 2 groups (controlling for our other covariates).

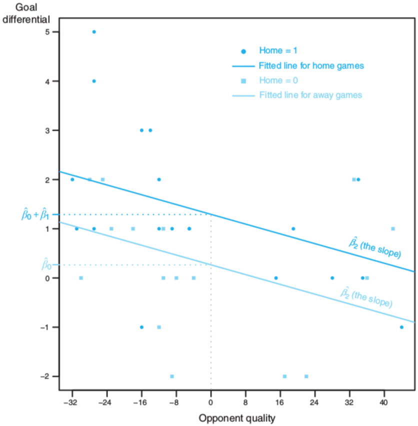
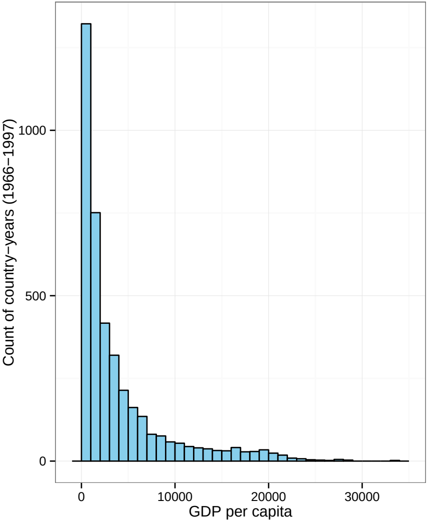
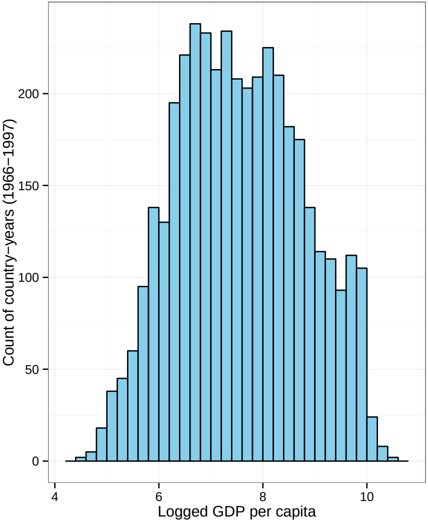


FIGURE 6.7: Fitted Values for Model with Dummy Variable and Control Variable: Manchester City Example

Log transformation (see Chapter 7 - optional)

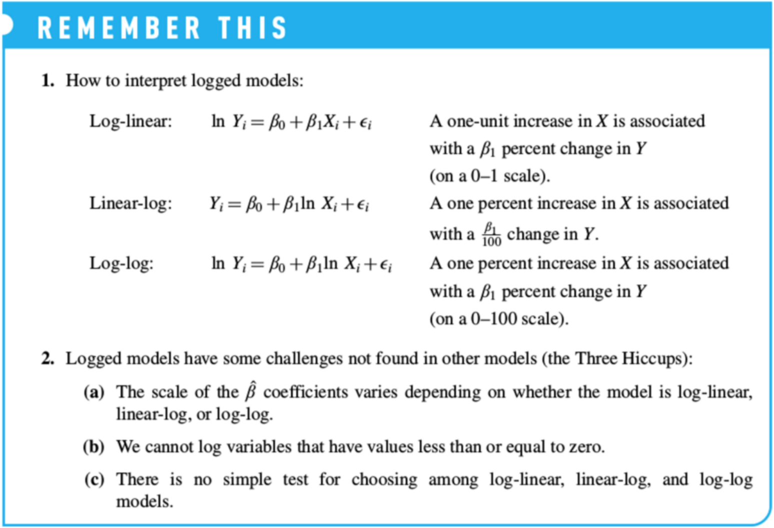
Sometimes you might want to take the log of a variable. This is particularly the case when the variable is skewed, like income.





Log interpretations

Read more p. 218 of Chapter 7 in *Real Stats*, p. 329 of digital version.



REMEMBER THIS

1. How to interpret logged models:

Log-linear:	$\ln Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	A one-unit increase in X is associated with a β_1 percent change in Y (on a 0–1 scale).
Linear-log:	$Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	A one percent increase in X is associated with a $\frac{\beta_1}{100}$ change in Y.
Log-log:	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	A one percent increase in X is associated with a β_1 percent change in Y (on a 0–100 scale).

2. Logged models have some challenges not found in other models (the Three Hiccups):

(a) The scale of the $\hat{\beta}$ coefficients varies depending on whether the model is log-linear, linear-log, or log-log.

(b) We cannot log variables that have values less than or equal to zero.

(c) There is no simple test for choosing among log-linear, linear-log, and log-log models.

Political Science 15
Introduction to Research in Political Science
Lecture 8c: Types of Data Sets

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Types of data sets

Cross-sectional data

- A sample of a population in a given period of time. You observe a bunch of units at one time period. Example: representative public opinion poll before an election.

Repeated cross-sectional data

- Taking different samples of a population over time. You observe different units over time. Example: multiple waves of a representative public opinion poll, where different people respond (pick up the phone) in each wave (sample).

Panel data

- Seeing the same population repeatedly over time. You see the same units over given time periods. Example: country GDP and wars by year from 1980-2015, turnout in each CA precinct between 2000-2018.

The benefits of panel data

- Seeing the same population again and again over time gives us a lot more leverage in identifying causal relationships.

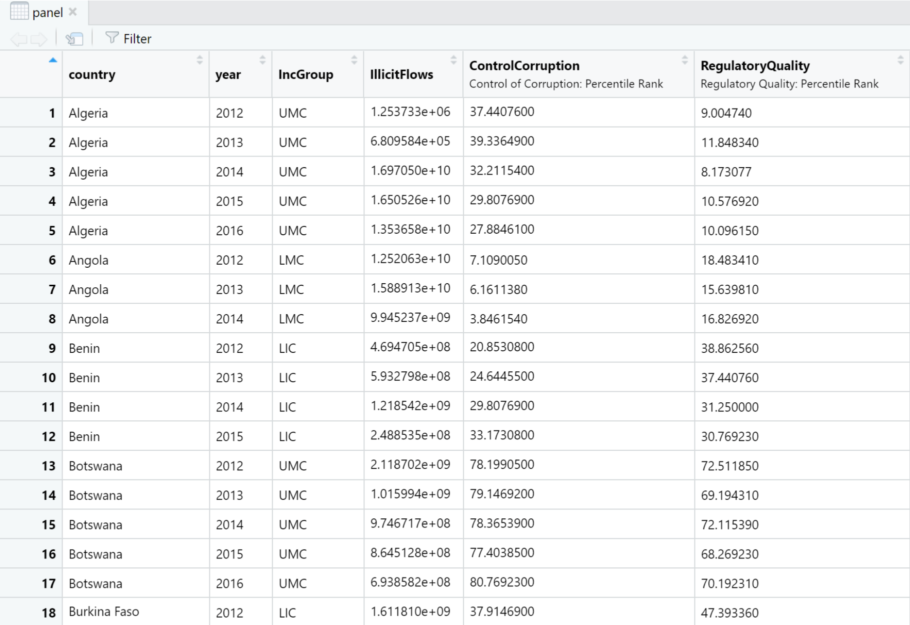
- Rather than comparing across units (as we do with a cross-section), we can now look at a single unit and examine how it changes over time.

- To do this, we use **fixed effects estimators** (read more in Chapter 8).

- Fixed effects models control for unit specific effects – it nets out the average way that unit behaves over time. You could think of this as the "culture" of a unit. Example: average turnout in a given district (unit) over time.

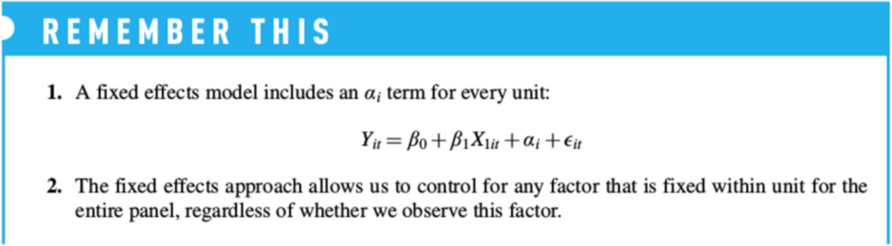
- Fixed effects models also control for time period effects – they control for the average behavior in a given year. Example: average income during recession years, average turnout in each presidential election year.

Panel data example (Lépissier, Davis, and Ibrahim 2019)



	country	year	IncGroup	IllicitFlows	ControlCorruption Control of Corruption: Percentile Rank	RegulatoryQuality Regulatory Quality: Percentile Rank
1	Algeria	2012	UMC	1.253733e+06	37.4407600	9.004740
2	Algeria	2013	UMC	6.809584e+05	39.3364900	11.848340
3	Algeria	2014	UMC	1.697050e+10	32.2115400	8.173077
4	Algeria	2015	UMC	1.650526e+10	29.8076900	10.576920
5	Algeria	2016	UMC	1.353658e+10	27.8846100	10.096150
6	Angola	2012	LMC	1.252063e+10	7.1090050	18.483410
7	Angola	2013	LMC	1.588913e+10	6.1611380	15.639810
8	Angola	2014	LMC	9.945237e+09	3.8461540	16.826920
9	Benin	2012	LIC	4.694705e+08	20.8530800	38.862560
10	Benin	2013	LIC	5.932798e+08	24.6445500	37.440760
11	Benin	2014	LIC	1.218542e+09	29.8076900	31.250000
12	Benin	2015	LIC	2.488535e+08	33.1730800	30.769230
13	Botswana	2012	UMC	2.118702e+09	78.1990500	72.511850
14	Botswana	2013	UMC	1.015994e+09	79.1469200	69.194310
15	Botswana	2014	UMC	9.746717e+08	78.3653900	72.115390
16	Botswana	2015	UMC	8.645128e+08	77.4038500	68.269230
17	Botswana	2016	UMC	6.938582e+08	80.7692300	70.192310
18	Burkina Faso	2012	LIC	1.611810e+09	37.9146900	47.393360

Fixed effects estimator



REMEMBER THIS

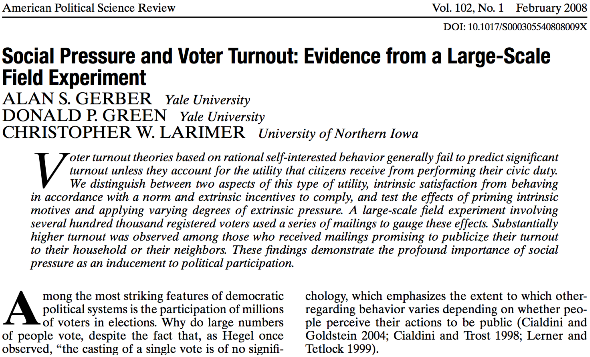
1. A fixed effects model includes an α_i term for every unit:

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \alpha_i + \epsilon_{it}$$

2. The fixed effects approach allows us to control for any factor that is fixed within unit for the entire panel, regardless of whether we observe this factor.

Note: You don't have to memorize or use this model for the problem sets or final exam, but you do need to know the benefits of a fixed effects model and what kind of data it needs to work.

Next lecture: Experiments



American Political Science Review Vol. 102, No. 1 February 2008

DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*

DONALD P. GREEN *Yale University*

CHRISTOPHER W. LARIMER *University of Northern Iowa*

Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty. We distinguish between two aspects of this type of utility, intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply, and test the effects of priming intrinsic motives and applying varying degrees of extrinsic pressure. A large-scale field experiment involving several hundred thousand registered voters used a series of mailings to gauge these effects. Substantially higher turnout was observed among those who received mailings promising to publicize their turnout to their household or their neighbors. These findings demonstrate the profound importance of social pressure as an inducement to political participation.

Among the most striking features of democratic political systems is the participation of millions of voters in elections. Why do large numbers of people vote, despite the fact that, as Hegel once observed, "the casting of a single vote is of no signifi-

chology, which emphasizes the extent to which other-regarding behavior varies depending on whether people perceive their actions to be public (Cialdini and Goldstein 2004; Cialdini and Trost 1998; Lerner and Tetlock 1999).

- Prepare for the modules by having read **Gerber, Green and Larimer** (posted on Gauchospace). Pay attention to the abstract, introduction, conclusion, figures and tables. You will also need to read this for Problem Set 5.
- Also read **Chapter 10** in the *Real Stats* textbook.