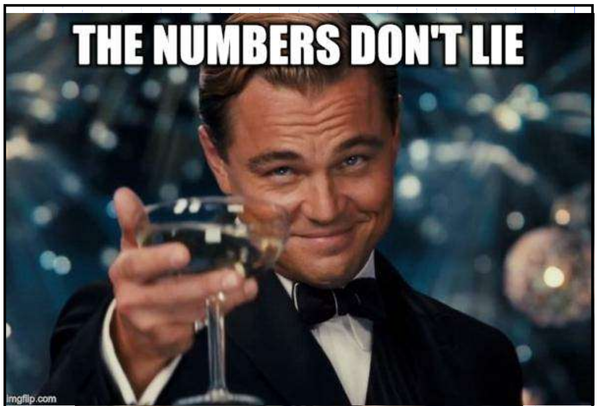


Q1: How many basic statistical tests are there with an interval DV?

- A. 1 — 0%
- B. 3 — 32%
- C. 5 — 58%
- D. 7 — 8%
- E. 9 — 3%

0

What grade do you expect to get?

- A. A — 44%
- B. B — 41%
- C. C — 15%
- D. D — 0%
- E. F — 0%

22

Oral Presentations in 26 days!

- **Content**
  - Should present your research project
  - Hypo (DV, IV; others?), background, data analysis, results, concs
- **Grading**
  - Students observing will grade each presentation
  - Five 1-5 scales: Clear? Prepared? Organize? Creative? Professional?
- **Order**
  - First, voluntary order until no volunteers
  - Then, repeated systematic sampling (k=?) down the list
  - I'll read each name twice; if you don't come up, you get a zero
  - If all names are called b4 time's up, we're done ~~(no class 5/11)~~
  - ~~If there are still volunteers or names then, we'll continue 5/10~~

SOC497/L: SOCIOLOGY RESEARCH METHODS

Multivariate Analysis

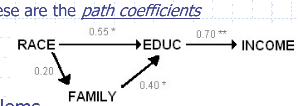
Ellis Godard

Outline for Today

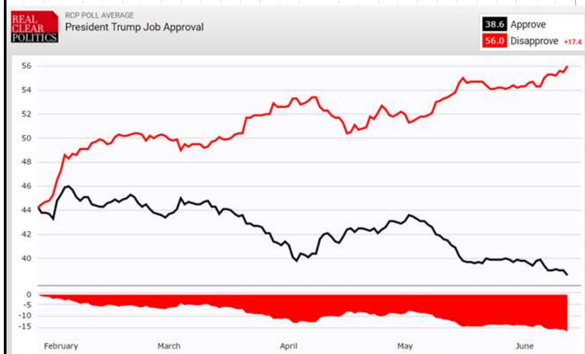
- Primer on Oral Presentations *(slide above)*
- Not quite Multivariate (could be 2)
  - Path Analysis & SEM
  - Time-Series Analysis
- Beyond Bivariate (3+)
  - Factor Analysis & Cluster Analysis
  - Multiple Regression
    - With a dummy variable
    - With an interaction term
- SPSS Demo *(time permitting or in lab)*
- Lab Exercise

Path Analysis

- Causal model about relationships among many variables
- Graphically represents components of a cohesive model
  - See 17-3 on page 447
- Involves correlation statistic for each pair of variables
  - Between each two is a causal "link" in a chain (or path)
  - These are the *path coefficients*

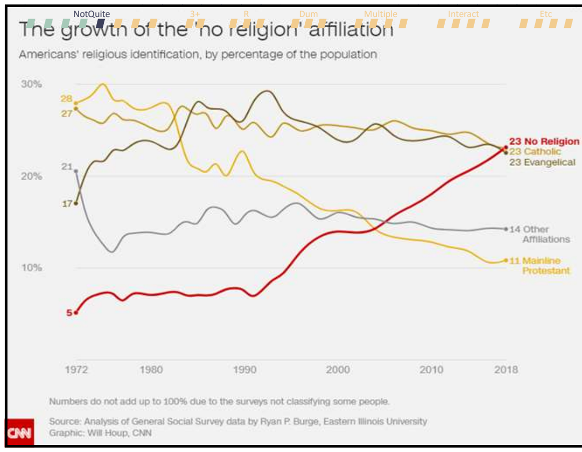


- Problems
  - Does not show or tell what the causal order *should* be
  - Can't confirm or falsify the order posited/theorized by researcher

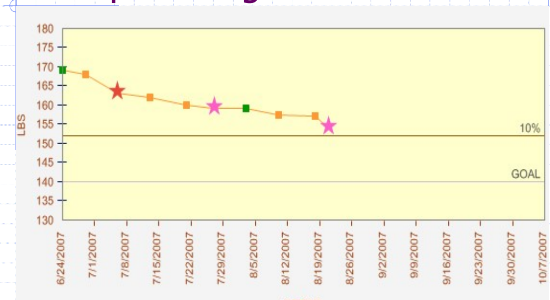


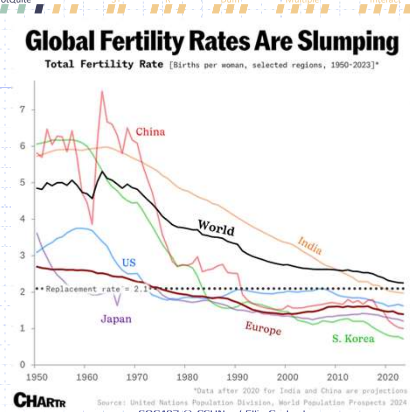
Time Series Analysis

- Represent changes in one or more variables over time.
  - Population and crime rates change over time for a city
  - Time "just" becomes one of the variables
- Time as a control variable
  - See multiple regression
- Time-lagged regression analysis
  - e.g. relate larceny rate to *previous* year's unemployment rate
  - Compare across years to see if that relation changes over time



Example: Weight Loss Chart

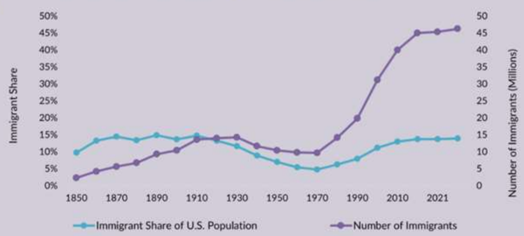




Chartr, 10/13/24

Immigrants*: Raw & Proportional

Figure 1. Size and Share of the Foreign-Born Population in the United States, 1850-2022



Sources: Migration Policy Institute (MPI) tabulation of data from U.S. Census Bureau 2010-22 American Community Surveys (ACS), and 1970, 1990, and 2000 decennial census; and Campbell J. Gibson and Emily Lennon, "Historical Census Statistics on the Foreign-Born Population of the United States: 1850 to 1990" (Working Paper no. 29, U.S. Census Bureau, Washington, DC, 1999).

https://www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states

* Note that this is *immigrants*, not annual *immigration* .

Q2: Which addresses the correlations for separate relationships in an overall model?

A. Dummy variables
0%

B. Factor Analysis
33%

C. Interaction Terms
3%

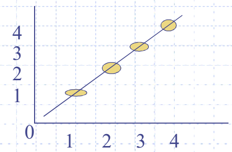
D. Path Analysis
65%

E. Time Series Analysis
0%

0

Linear Regression

- General formula to describe assoc. between 2 variables
  - Y = f(X)
  - Y = a + bx (not mx + b)
  - INCOME = 3.4 + 863*EDUC



- Can be used to predict values of Y, given values of X
- There is always an implicit error term (y = a + bx + i)

Factor Analysis

- Used to discover patterns among several variables
  - Like simultaneous correlations
    - Complex algebra; *not* by hand
  - But about grouping variables (like indices), not cause/effects
  - Generates eigen values suggest which variables "go together"
- Problems:
  - Generates factors w/o substantive meaning
    - Combining variables is a *measurement* decision
    - Factor "loadings" assess statistical relationship among variables, rather than correlative dimensions of the same concept
  - Facilitates *ad hoc* & *post hoc* analyses
    - Grouping variables ought to be for theoretical reasons
    - Grouping based on 1 dataset risks patterns due to sampling bias
  - Result aren't falsifiable
    - *Something* will *always* hold together as loaded factors

Benefits of that "Assumption"

- **Describes the relationship between 2 variables**
- **Inferential values: Allows us to predict…**
  - …the *average* value of the DV for each value of the IV
  - … the avg *increase* in DV, *for each increase of 1* in IV

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	223.190	1.393		160.193	.000
	Index of online media use	-5.102	.366	-.201	-13.932	.000

a. Dependent Variable: weight

  - …how much of the variance in the DV is explained by variance in the IV

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.201[a]	.041	.040	79.830

a. Predictors: (Constant), Index of online media use

  - **Can write equation:**
    - **Weight = 223.190 – (5.102*OLUSE)**
    - **Note p values**

Note that equations include variable names!

Q3: A regression equation…

A. summarizes evidence against line of best fit
0%

B. always includes an interaction term
3%

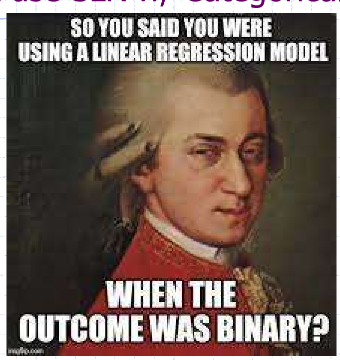
C. always includes an implicit error term
97%

D. connects as many points as possible
0%

E. Always involves two nominal variables
0%

0

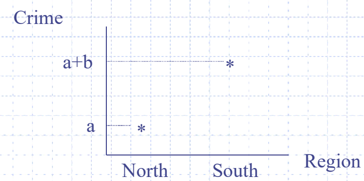
Don't use SLR w/ Categorical DVs



Dummy Variables

- Consider a model that uses Region to predict crime rates
  - Region has 5 nominal categories
  - Slope doesn't make sense if include all 5; assumes settled order
- Instead, create a new variable for each region
  - SOUTH has two values, 0 and 1
    1 means the state is in the South; 0 means it isn't
- Using a dummy as the IV
  - The equation using only that one IV would be:
    Crime=a + b(South)
  - The crime rate in the South is:
    Crime = a+b(1) = a+b
  - The crime rate in the North is:
    Crime=a+b(0)=a.

Regional Differences in Crime



Multiple Regression

- More than one independent variable (X)
- A different slope (b) for each one, e.g.
  - INCOME = 2.4 + 535*EDUC - 158*KIDS
  - Each IV has its own coefficient (e.g. 535, 158)
  - Each slope is the *average predicted* change in Y for each increase of 1 in *that* X
- ONE equation for ALL of the variables
  - NOT a separate equation for each variable
  - No limit to the number or type of variables
  - Can plug in values of IV, to predict value of DV

Multiple Regression w/ 1 Dummy: Region, Urbanization and Crime

- We hypothesize that crime is a function of urbanity and southerness – that is, that it will be higher in cities and in the south.
- Assuming the relationship is linear, the regression equation would be:
  Crime=a + b_1(Urban) + b_2(South).

Multiple Regression Equations

- For the non-south cases, one part of the equation "falls out", leaving 2 equations:

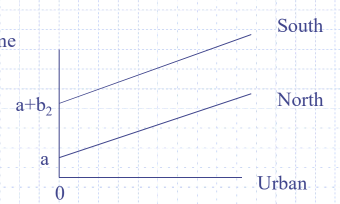
$$Crime_{South} = a + b1(Urban) + b2(1)$$
$$= a + b1(Urban) + b2$$

$$Crime_{North} = a + b1(Urban) + b2(0)$$
$$= a + b1(Urban)$$

SOC497 @ CSUN w/ Ellis Godard 33

Multiple Regression Plots

- Since we have 2 equations, we have 2 lines:



Assumes same slope across regions

SOC497 @ CSUN w/ Ellis Godard 34

Regression w/ Interaction Term: Region, Urbanization and Crime

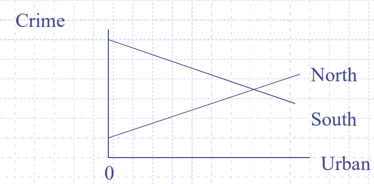
Crime = f(Urban, Region, Urban*Region)

Create "interaction" term btwn Urban and Region:
COMPUTE=SOUTH*URBAN

Crime=a + b_1(Urban) + b_2(South) +b_3(Urban*South)

SOC497 @ CSUN w/ Ellis Godard 35

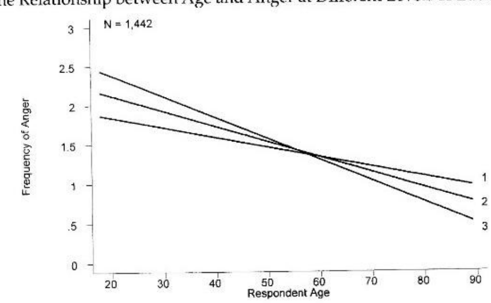
Plot with Interaction Term Included: Region, Urbanization and Crime



Assumes different slopes across regions

SOC497 @ CSUN w/ Ellis Godard 36

The Relationship between Age and Anger at Different Levels of Education



Note: Lines correspond to equation 2 of Table 1 and reflect predicted anger frequency scores for white men. Intercepts will vary slightly with other combinations of sex and race. Lines 1, 2, and 3 reflect predictions for individuals with 17, 13, and 10 years of education, respectively. These education values are the 90th, 50th, and 10th percentiles.

Scott Schieman, "Socioeconomic Status and the Frequency of Anger Across the Life Course". In *Sociological Perspectives* (journal of the PSA), Vol 46, No. 2 (Summer 2003), pp. 207-222.

Q4: Which is most like an index?

A. Dummy variable 66%
B. Interaction term 13%
C. Path coefficient 16%
D. Pinky finger 5%
E. Do you really want a 5th option to struggle with? No, you don't. So, don't pick this one 0%

20

SOC497 @ CSUN w/ Ellis Godard 38

Multiple Reg. Demo? In lab?

- Using regmulti-demo.sav
  - Labs done, Days absent, and Midterm grade for Summer 2020
- Two *separate* simple linear regressions:
  - Labsdid predicting MTGrade & Absent predicting MTGrade
  - Compare r-squared values
  - Compare coefficient directions (one is negative!)
  - Can't compare coefficient sizes – different scale, intercept, etc.
  - Compare p values for coefficients
- One multiple regression
  - Look at change in r-squared
  - *Can* compare coefficients, because each takes other in consideration
  - *Do* compare p values for coefficients

Not in lab, but might demo?

- Compute interaction term
  - LABATT = labsdid * absent
- New multiple regression
  - Test interaction on its own
  - New regression w/ everything (but not if p>.05)
  - Write new equation

Multiple Regression Lab

- Use regmulti-lab.sav, find parameters for an equation which predicts a respondent's level of education with five possible independent variables:
  - Their father's level of education
  - Their mother's level of education
  - The number of siblings the respondent has
  - The age of the respondent
  - The respondent's age when they first married
- Write the equation, using all of the intercepts and slopes.
- With which variable(s) is education inversely proportional (it goes down when they go up)?
- For which variable(s) is the probability too high that the slope differs from zero only due to sampling error? (that is, which p values is too risky?)

- Rerun using only those variables w/ significant coefficients (slopes)
- Rewrite the equation using only those (*new!*) coefficients.
- Using this second equation, what level of education do we predict for a respondent with 2 siblings whose mother got 16 years, whose father got 20, who married at 23, and who is 29 years old?

Q5: Each of today's backgrounds had…

A. Boats
5%

B. A baseball
0%

C. Water
0%

D. A wheel
92%

E. Wombats
3%

0

Team Scores

Points	Team	Points	Team
3.15	A		
2.92	B		
2.8	C		