



Lecture 2: Carrying Out an Empirical Project

[Research questions]

You will come to understand statistical approaches to answering questions like these:

- Is a particular rehabilitation program effective in reducing recidivism?
- Does gang membership increase crime?
- Does juvenile arrest affect high school dropout?
- Does inequality increase crime rates?

What do these questions have in common?

[Theory]

- Barring data restrictions, the way you approach research questions is guided by criminological theory.
- E.g. Social control, strain, differential association, social disorganization
- These theories point to constructs that account for crime.
 - For statistical analysis, we create variables that are *supposed to* represent theoretical constructs.

[Types of Data]

Your approach to answering research questions is constricted by the data to which you have access.

- Nonexperimental data: naturally occurring, preferably collected in a systematic manner
- Experimental data: random assignment of cases to two or more conditions.

[Posing a Question]

- Wooldredge focuses on the economic literature, some of which may be relevant to your topic. You should primarily focus on criminological theory and literature.
- Top criminology journals:
 - *Criminology, Criminology & Public Policy, Justice Quarterly, Journal of Quantitative Criminology, Journal of Research in Crime and Delinquency*

[Literature search]

- Google Scholar is a good start, although it tends to be biased towards older articles since it ranks articles by number of citations.
 - Follow the “cited by” link for important articles to find newer articles on the same topic
 - The “related articles” link can be useful as well.
 - You can set your preferences to link straight to the ASU library from Google Scholar.
- Library databases can be useful as well: Criminal Justice Abstracts, etc.
- Don't forget books!

[Data sources]

- National Time Series
 - UCR, NCVS, Census, GSS
 - Easy to acquire, limited range of information
- Large panel datasets
 - NLSY79, children of NLSY79, NLSY97, Add Health, NELLS, NYS, RYDS, PTD
 - Varying number and difficulty of hoops to jump through in order to acquire data.
 - Rich data, some are nationally representative
 - Varying levels of access, can merge with national time series

[Data sources]

■ ICPSR

- <http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/index.jsp>
- Thousands of original datasets of varying quality with varying levels of documentation.
- Can search by topic and quickly download data.

[Data format]

- You will be doing analysis within Stata.
- Using the “import” command in the File menu, Stata can open the following formats:
 - CSV, SAS, XML, possibly others.
- Other stat packages can often save in the Stata format. SPSS can save in Stata format.

[Spend time with your data]

- Look at it.
 - Use data editor or browser in Stata's "Data" menu
 - Use the following commands: list, tab, scatter, summarize, histogram, lowess
- How is missing information handled? Make sure it's a non-numeric code.

[Spend time with your data]

- What is each variable's level of measurement?
 - Binary (0/1)
 - Nominal/categorical
 - Don't enter directly into regression!
 - Transform into dummy variables
 - Ordinal
 - Consider transforming into dummy variables
 - Interval: seriousness scale used for sentencing
 - Doubling the value doesn't necessarily mean that seriousness is doubled.
 - Ratio
 - All statistics and transformations are permitted

[Spend time with your data]

- Look out for mistakes in the data.
 - Min, max, scatter plot
 - Nonsensical combinations of responses
 - If extreme outliers are mistakes, recode to correct values (if possible) or delete.
 - What should you do with outliers you suspect to be untrue?
 - Ex: In NLSY97, several teens report having sex 999 times with 99 different partners in the past year.
 - You can censor the data. Set maximum to 100, for example. You can also run the analysis with and without those cases.

[Hypothesis Testing]

- “Null hypothesis testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.” - Rozeboom (1997)

[Bone-headed? What are the critiques?]

- 1) Flawed statistical properties (Type 1 vs. Type 2 error, false positives vs. false negatives)
- 2) Over-reliance on statistics, need more qualitative studies and theoretical development.
- 3) Too much emphasis on p-values, not enough on effect sizes. Statistical vs. substantive significance.

Recall the steps for hypothesis testing:

- 1) State null and research hypotheses
- 2) Select significance level
- 3) Determine critical value for test statistic (decision rule for rejecting null hypothesis)
- 4) Calculate test statistic
- 5) Either reject or fail to reject (not “accept” null hypothesis)

[Standards for appropriate null hypothesis significance testing (NHST)]

- 1) Report descriptive statistics for all variables used in analysis
- 2) Report effect size in an easily interpretable way (elasticity, standardized betas)
- 3) Report standard errors, t-stats or p-values
- 4) Report confidence intervals for coefficients of interest
- 5) Discuss size of coefficients

[Standards for appropriate null hypothesis significance testing (NHST)]

- 6) Contextualize effect size. Discuss beforehand what a small, medium or large effect would be.
- 7) Do not use statistical significance as only criterion of importance
- 8) Same as above.
- 9) Distinguish between descriptions of statistical and substantive significance
- 10) Consider statistical power

[Standards for appropriate null hypothesis significance testing (NHST)]

- 11) If you fail to reject the null, make use of confidence intervals.
- 12) Don't accord substantive significance to your non-statistically significant estimates
- 13) Don't "accept" the null hypothesis
- 14) Specify the correct null hypotheses
- 15) Include/exclude variables for theoretical, not just statistical, reasons

#2: Report effect size in an easily interpretable way

- Consider the units of analysis for your independent and dependent variables. Are they meaningful?
- Does the coefficient have a real-world application that would make sense for a policy maker or practitioner?
- Examples:
 - Arrest → legal earnings
 - Religiosity → self-control
 - SAT score → college admission

#2: Report effect size in an easily interpretable way

- Several options for reporting effect size:
 - Original coefficient (if units are meaningful)
 - Logarithmic transformation (Wooldredge pp. 43-46)
 - Elasticity
 - Standardized beta

#2: Report effect size in an easily interpretable way, logarithmic transforms

- It may make more sense to think of the effect of X on Y in terms of constant *percent* increases. To transform the regression in this way, log the dependent variable.

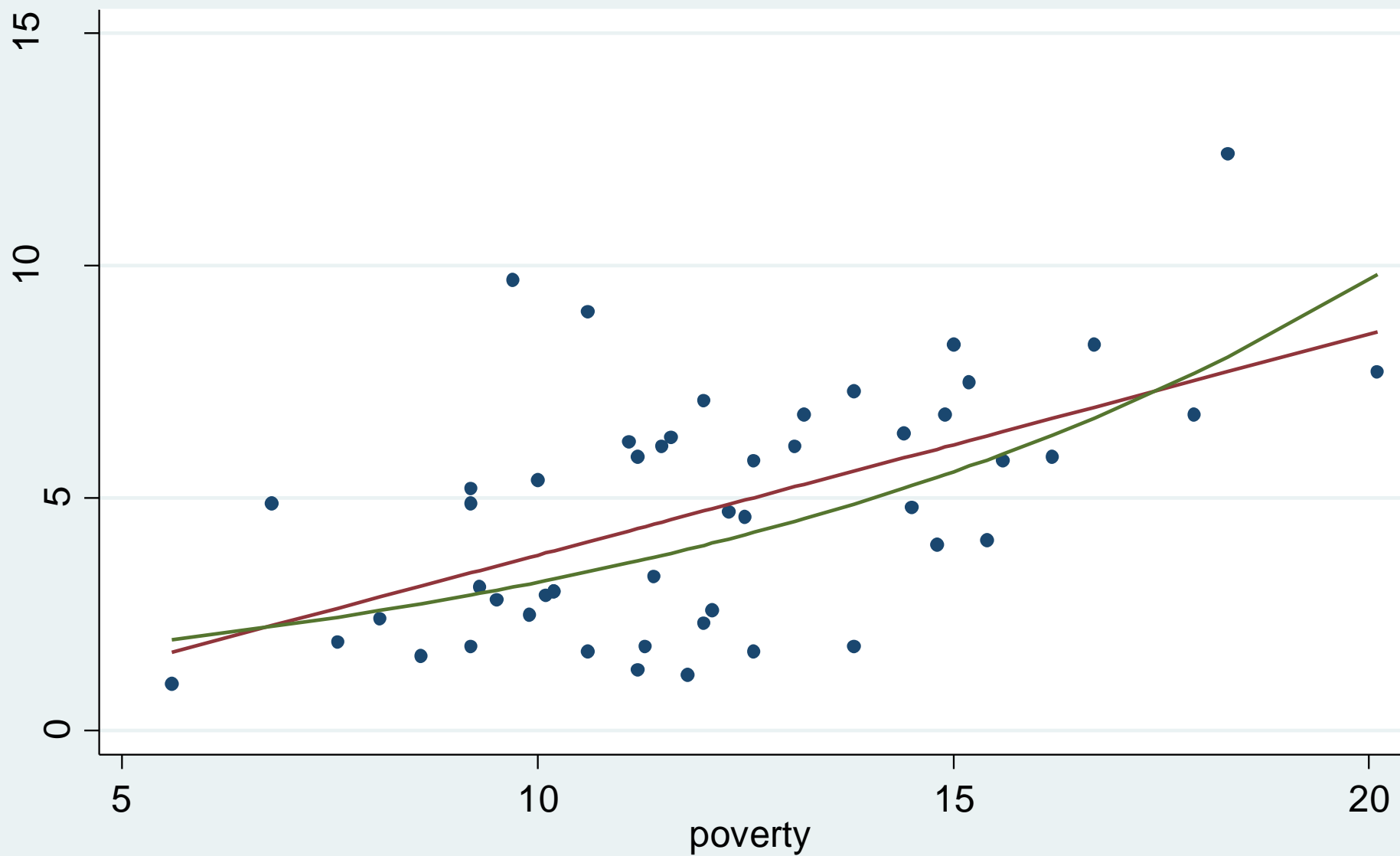
$$\log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

- While this assumes a constant effect of X on $\log(Y)$, in an increasing function, it translates to an increasing effect of X on Y as X increases.

$$y_i = e^{\beta_0 + \beta_1 x_i + u_i}$$

#2: Report effect size in an easily interpretable way, logarithmic transforms

- In the poverty and homicide example, the coefficient for poverty on logged homicide is .11. This means that a 1 percentage point increase in the poverty rate is associated with an 11% increase in the homicide rate.
- The following slide shows the scatter plot for poverty and homicide, the linear regression line, and the transformation of the regression line when homicide rates are logged.
- This shows that logging the dependent variable introduces a non-linear relationship.



#2: Report effect size in an easily interpretable way, elasticity

- A common kind of elasticity reports the effect of a 1% change in X in terms of percent change in Y (at the mean for both).

$$\beta_x^{el} = \frac{\beta_x \cdot \bar{x}}{\bar{y}}$$

- In the homicide rate and poverty example, we would have $(.475 \cdot 12.09) / 4.77 = 1.20$
 - This means that a 1% increase in the poverty rate results in a 1.2% increase in the homicide rate.
 - Is this consistent with the earlier result? Yes. Know difference between **percent** and **percentage point** increase.
- In Stata, immediately after running the regression:
 - `margins, eyex(poverty) atmeans`

#2: Report effect size in an easily interpretable way, elasticity

- Another way to obtain elasticity is to log both the dependent and independent variables:

$$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + u_i$$

- In the homicide rate and poverty example, we get a slightly different answer: 1.31, meaning that a 1% increase in the poverty rate results in a 1.31% increase in the homicide rate.
- Why the difference?
 - Margins evaluates the elasticity at the mean
 - The regression estimates a *constant elasticity* across all values of X

#2: Report effect size in an easily interpretable way

- Standardized betas report the expected effect of a 1 standard deviation change in X in terms of standard deviations in Y

$$\beta_x^{st} = \frac{\beta_x \cdot \sigma_x}{\sigma_y}$$

- In order to transform a regression coefficient into a standardized form, multiply by the standard deviation of x, and divide by the standard deviation of Y

#2: Report effect size in an easily interpretable way

- In the homicide rate and poverty example, we would have $(.475 \times 3.01) / 2.58 = .55$
 - This means that a 1 standard deviation increase in the poverty rate results in a .55 standard deviation increase in the homicide rate.
- You could also get this easily in Stata:
 - `reg homrate poverty, beta`
 - Or:
 - `egen homst=std(homrate)`
 - `egen povst=std(poverty)`
 - `reg homst povst`

[#2: Report effect size in an easily interpretable way]

- There is no bright line for what constitutes a large effect using standardized coefficients. But most researchers would consider .2 small, .5 medium, and .8 or above large.
- Wooldridge's chart on page 46 is a good reference for logarithmic transformations and how to interpret them.

#4, #11: Report confidence intervals around effect sizes

- Rather than focus on a single point estimate, report confidence intervals, especially when your key coefficient is not statistically significant.
- If you report the 95% confidence interval, for example, you can say with 95% certainty that the “true effect” lies within the interval (provided your model’s assumptions hold)
- In the homicide and poverty example, we have the following*:
 - 95% ci: [.268, .682]
 - 95% ci: standardized: [.314, .796]
 - 95% ci: elasticity: [.67, 1.74]

#4, #11: Report confidence intervals around effect sizes

- If I see that the key variable in a research article is not statistically significant, I always want to see a confidence interval. It will contain zero, but different confidence intervals communicate very different things
- Width of the confidence interval matters:
 - $[-.05, .08]$ – we know the true standardized effect is not large (high power)
 - $[-.82, 1.18]$ – we know nothing about the true standardized effect (low power)
- Whether the confidence interval contains large negative or positive effects matters:
 - $[-.67, .02]$ – we're confident that the true effect is not strongly positive

[#4, #11: Report confidence intervals around effect sizes]

- Example: Loeffler (2013), “Does Imprisonment Alter the Life Course?”
- Loeffler used random assignment to judges as a source of exogenous variation in imprisonment decisions to determine if imprisonment affects recidivism and employment.
- He first showed the results using observational data, and then the quasi-experimental results that used variation from judge assignment to identify the effect.
- The idea is that the conventional results have unobserved variables bias whereas the judge-effect results will be unbiased as judge assignment is random.

#4, #11: Report confidence intervals around effect sizes

Table 2. Estimated Effects of Imprisonment on Recidivism within 5 Years

Variables	(1) Unconditional		(2) OLS		(3) 2SLS		(4) 2SLS	
	<i>b</i>	(SE)	<i>b</i>	(SE)	<i>b</i>	(SE)	<i>b</i>	(SE)
Prison	.088***	(.007)	.031***	(.008)	.075	(.104)	-.035	(.113)

Table 3. Estimated Effects of Imprisonment on Employment at 5 Years

Variables	(1) Unconditional		(2) OLS		(3) 2SLS		(4) 2SLS	
	<i>b</i>	(SE)	<i>b</i>	(SE)	<i>b</i>	(SE)	<i>b</i>	(SE)
Prison	-.069***	(.006)	-.036***	(.006)	.031	(.083)	.055	(.091)

- What are the 95% confidence intervals for OLS and 2SLS (judge effects) results?
- Do the 2SLS results contradict the OLS results?
- See Loeffler 2013:156-157 for interesting discussion



[#5, #6: Discuss size of coefficients. Tall, grande, venti?]

- Ideally, in the discussion of theory and the prior literature, one would consider benchmarks for a small, medium or large effect.
 - Unfortunately, theory usually provides little guidance.
 - Were Nagin & Tremblay (2005:918) correct? “Theories are generally little more than simpleminded human brain products offered for falsification.”
 - If a statistically significant and positive/negative coefficient is enough to confirm a theory, maybe the theory is too simple.



[#5, #6: Discuss size of coefficients. Tall, grande, venti?]

- In well-trod areas, prior research can provide some guidance about what to expect.
- Consider your results without looking at the standard errors, T-scores or p-values.
 - Is it a good test of theory?
 - Are the results meaningful?




[#7, #8: Statistically significant ≠ important, big or meaningful!]

- Don't assess the importance of your findings solely on statistical significance (or lack thereof).
- When comparing the effects of multiple control variables, the T-score does not indicate which is the most important or most influential.
- Small p-scores don't confirm theories, nor signal agreement with previous studies. Small p-scores may indicate statistical power.
- #9: Be careful with the word "significant." Do you mean statistically significant or substantively/analytically significant (i.e. important)?

[#7, #8: Statistically significant ≠ important, big or meaningful! (good example)]

An Incident-Based Comparison of Female and Male Sexual Offenders

Sexual Abuse: A Journal of
Research and Treatment
1-23
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1079063214544333
sax.sagepub.com


Katria S. Williams¹ and David M. Bierie¹

Abstract

Identifying the ways in which male and female sex offenders differ is an important but understudied topic. Studies that do exist have been challenged by a reliance on small and select samples. Improving on these limitations, we use the National Incident-Based Reporting System (NIBRS) to compare male and female sex offenders among all 802,150 incidents of sexual assault reported to police across 37 states between 1991 and 2011. Findings indicated some broad similarities between groups, including the

- “Differences are discussed in terms of substantive significance rather than statistical significance. This is particularly important because of the large sample size in the study; even substantively small differences will demonstrate statistical significance.”



[#7, #8: Statistically significant ≠ important, big or meaningful!]

- It is difficult to accumulate knowledge without comparing effect sizes across studies.
 - Only 6% of *Criminology* articles and 3% of *Justice Quarterly* articles (2001-2002) discussed the size of their effects.
 - Statistical significance alone does not signal agreement between different studies, nor does it add much to our knowledge.
- In addition, if researchers are dissuaded from pursuing publication when their p-values are $>.05$, then the body of published research is a biased sample of all research conducted.

#10, #13: Statistical power, Type I vs. Type II error

	H_0 is true	H_0 is false
Fail to reject	Correct	Type II error
Reject	Type I error	Correct

- The probability of type I error (false positive) is alpha level, chosen by researcher (preferably before running regression model!)
- The probability of type II error (false negative) is generally unknown, determined by power of the test

[Type I vs. Type II error, cont.]

- Type I and Type II errors are inversely related.
 - Criminal case: “beyond a reasonable doubt” minimizes false convictions (Type I), increases chance of letting guilty free (Type II)
 - Civil case: “preponderance of evidence” increases chance of false finding of culpability, decreases chance of letting culpable off
 - Not used, but promoted by some as appropriate burden of proof for capital cases: “beyond the shadow of a doubt”
- In statistical tests, the smaller our chosen alpha level, the higher the possibility that we’ll fail to reject a false hypothesis (type II error)

[Type II error & statistical power]

- The dangers associated with Type I and Type II errors are not constant across applications, yet criminologists typically only pay attention to Type I errors.
- The chances of Type II error decrease as our “statistical power” increases.
- “Statistical power” refers to our ability to reject the null hypothesis.
- In order to reject a null hypothesis when there is a very small effect size, one needs a lot of statistical power.
- Bigger samples = more power.

Statistical power

- Two types of power analysis: *a priori* and post-hoc
- When designing a study, one might conduct an *a priori* power test to determine the appropriate sample size to achieve enough statistical power to detect effects of a certain expected magnitude
 - Desired statistical power is set, and necessary sample size is calculated
- Post-hoc power analysis assesses statistical power of existing studies
 - Sample size and other characteristics are used to determine power
 - Only important if you don't have statistically significant findings. (E.g. Weisburd et al., 2003)

Statistical power: for more info

- <http://www.ats.ucla.edu/stat/stata/dae/powerreg.htm>
 - Stata command
- <http://www.stat.uiowa.edu/~rlenth/Power/index.html>
 - There is no single way to calculate power across all types of statistical analysis. This page contains several online applets for computing statistical power
- To summarize:
 - Power is the inverse of Type II errors (failing to reject false null)
 - Sample size \uparrow , power \uparrow
 - Alpha \uparrow , power \uparrow

#14: Specify proper null hypotheses

- The typical procedure is to test the null hypothesis that $\beta=0$. But sometimes this is not the appropriate test.
- Theoretically-expected nonzero effects
 - Sentencing guidelines: the effect of the recommended sentence on the actual sentence should be 1 ($\beta=1$)
 - Racial threat. As the minority population increases, whites feel more threatened and spend more on criminal justice. But Blalock (1967) suggested that there's a tipping point, and after a certain proportion of the population is minority, there is no more effect of increasing the minority population.
 - So the relevant hypothesis is that the relationship is non-linear. The coefficient on the squared minority population should be nonzero..

#14: Specify proper null hypotheses

- Testing previous findings
 - Across time / across studies: Has there been a change in the effect? Is the relationship different in the NLSY79 vs. the NLSY97?
 - Within a study. In Model 1, β_1 is statistically significant, but in Model 2, β_1 is not statistically significant. This may not mean that the two estimates are statistically significantly different. But you can explicitly test whether the two are equal.

#14: Specify proper null hypotheses

- Other possibilities
 - Equality of coefficients within a model
 - Joint significance of multiple coefficients
 - four types of high school dropout
 - age and age-squared
 - Interaction effects: e.g. effect of high school work on educational attainment at different levels of identity as a good student/college-bound

[Concluding Remarks]

- Think first of developing a meaningful model without regard for statistical significance.
- Think about: appropriate null hypothesis, alpha level, statistical power.
- Report: descriptive statistics, effect sizes, standards for small/large effect based on theory and prior studies, confidence intervals

[Concluding Remarks]

- Distinguish theoretical hypotheses from statistical hypotheses.
 - e.g. statistically significant peer effects do not necessarily confirm social learning theory
- Separate substantive from statistical significance.
 - Statistical significance is more about precision and sample size than importance
 - Effect sizes are what matter for theory and policy.

[Next time:

Homework: Problems 2.6, C2.2, C2.4iii

Read: Wooldridge Chapter 3