LS 123 Data, Prediction, and Law

Meeting 5

**Against Prediction, OLS for Causal Inference**

1. **reflection: what did Large N lab tell us?**
   1. classical hypothesis testing
      1. may not be all that interesting but you see it all the time across the social and life sciences (and elsewhere)
      2. rejecting a null hypothesis: the means are the same, the coefficient is zero, etc (usually the null itself is not very informative)
      3. this is a common approach when we talk not just about means of a particular variable but also other central tendencies, like a regression coefficient (so can we safely say that it is not zero)
   2. Central Limit Theorem enables us to move forward
      1. the distribution of a statistic involving sums (the mean is what we are usually talking about) approaches a Normal distribution about the population mean as you keep sampling from that population
      2. so the logic here is that, if you keep sampling, the value of the mean that you are talking about will get closer and closer to the value in the underlying population—we are most often talking about regression coefficients, which are themselves averages
      3. this allows us to make inferences about the population—the name of the game for many social scientists (e.g., voters, in the case of ANES)
   3. what did the lab tell you about the logic of confidence intervals and what “95% confidence level” actually means? (logic of repeated sampling)
   4. the jury example was not particularly apt for talking about classical hypothesis testing
      1. we were talking an expected proportion when we know the population proportion of the variable of interest (so you would likely choose another sort of hypothesis test, e.g. Chi-square)
      2. **Note: t-test is not what you would use for the kind of multinomial distribution in the jury example**
         1. Remember toward the end of class today I said something about how the Student's t-test for significance when we are talking about the mean of a given variable (like "Liberal Feeling Thermometer") did not seem to be the right hypothesis test for the proportion data in Lab 5 that had to do with selecting a jury? It is not the right test, because we are comparing the observed frequency of an event to the expected frequency of an event in each of the categories (Asian, Black, Latino, White, Other). The expected frequency is the known population frequency, and the observed frequency is what we see in the unfair draw from the jury pool. So a Chi-square test would be the test of significance to use. In general, Student's t is used for small samples where we don't know the variance of the error term. Student's curve is like a normal curve that has been flattened out and has fatter tails.
         2. For binomial distributions with a large sample size, you can use a z-test (so a normal distribution), but not when we are comparing multiple categories. For a good explanation without mathematical heavy lifting, see Freedman, Pisani, and Purves (1998) *Statistics*, chs. 26-29. (David Freedman was a statistics professor here at Berkeley.)
         3. So I need to go back and fix this lab, although I think that you can see [my lab with notes](https://github.com/ds-modules/Legalst-123/blob/master/labs/05_Large%20n/05_Large_n_solutions_jon.ipynb) where I used the Chi-square test. If you remember the [jury example from the Data 8 text Computational and Inferential Thinking](https://inferentialthinking.com/chapters/11/2/Multiple_Categories.html), the authors use a statistic called "Total Variation Distance" to measure the overall difference between population proportions and proportions in the jury pool and then show us a distribution of that statistic to demonstrate how unlikely the particular jury draw would be.
         4. I am not sure why I didn't think about this last Spring, but maybe it **is** good to stand up in front of class and profess in order to see things clearly. Proportions for binomially distributed data may be like means (thus you can use a z-test), but the proportions in the jury problem are not.
   5. likewise, the liberal and conservative feeling thermometer question was a bit misleading
      1. what does it mean to ask whether we can reject the hypothesis that the means were drawn from the same underlying groups in the population when we just look at the mean scores from all respondents on the two feeling thermometers?
      2. I reworked the question to make it more interpretable—given these groupings (i.e. partisan identifiers), can we reject the null hypothesis that they have the same feelings towards liberals (or conservatives or scientists, etc)
      3. note that box plots are a very effective way or visualizing this question
2. **Against Prediction ch. 1**
   1. Basic thesis:
      1. Actuarial methods in criminal justice are on the rise
      2. This is bad because actuarialism necessarily:
         1. Assumes rational-actor theory
         2. Assumes equal demand elasticities for committing crime
      3. Instead of probability-based prediction methods, audits should be distributed randomly
      4. Ratchet effect: primary problem
   2. Are there pieces of the argument that I missed that you want to touch on?
   3. What do you find compelling about this argument? What are the potential problems?
   4. Economic Critiques of Harcourt:
      1. Rational-Actor Theory does not require equal demand elasticities, despite Harcourt’s claim
         1. In fact, we are just assuming rationality and assuming elasticity of demand for committing crimes
         2. We don’t really know whether and how much people respond to the negative returns of punishment or how they calculate the probability of apprehension
      2. Use of the term “efficiency” is muddled
         1. Efficiency in economics generally refers to allocative (Pareto) efficiency (i.e. no one can be made better off without making someone else worse off)
         2. Harcourt confuses the dictionary definition of efficiency with the economic one; what he is really talking about is maximizing marginal return to investment (in policing, punishing, etc)
         3. I’m not sure the lack of adherence to language of microeconomics is that much of a problem
   5. Would distributing audits randomly be a wise policy decision?
   6. What are some other issues with the “actuarial turn?” How should we think about them as data scientists?
   7. **ratchet effect” of prediction (Bernard Harcourt)**

general problem with prediction is that it is self fulfilling

e.g., if your risk score determines that you need more supervision, there will be more ways for you to fail and become a recidivist (justifying the high risk classification)

similarly, if policy focuses surveillance on a targeted population (for example, African Americans, or Latinos), then you will detect more offenders in the targeted population, and soon the targeted population accounts for a disproportionate number of people under supervision

law enforcement, wishing to be efficient in its allocation of resources, then focuses even more heavily on the targeted population, since it obtains a higher “hit rate” there

this is what Bernard Harcourt calls a ratchet effect—the initially targeted population winds up with a disproportion of criminal justice contacts and criminal records, all as a result of an initial difference in the amount of effort to detect and punish crime in the target population

1. **correlation and regression** 
   1. note that we are talking about ratio or interval measures—that is what we need for both correlation and for OLS regression; the outcome of interest needs to be a continuous variable (so like an integer or a real number, not a category)
   2. you will see regression models all the time, and we will learn more about regression as we move forward
   3. today we’re going to talk about regression for making arguments about cause and effect—that means that they are embedded in some kind of theory
      1. a different activity from what we do for most of this course, which is trying to predict (although the two ideas are connected)
      2. causal models are important in social sciences and in understanding a lot of the readings (and so is prediction—e.g., the PCRA and COMPAS recidivism prediction instruments)
   4. later on we will focus on regression as a predictive tool (and we are going to do lots with it, in various forms—see class meeting 11)
   5. I want to say a few words about
      1. regression assumptions
      2. regression as a tool of causal inference (which is where you often see it in the social sciences)
2. **regression analysis: applications**
   1. scientific method: develop an hypothesis from existing information, test it, and be ready to modify the theory that guided you in hypothesis making
   2. means also falsification: there can be evidence that invalidates your underlying theory
   3. the gold standard for technique is the experiment: you change only the variable of interest and compare it to a known (control) condition
   4. problem: experiments are not always possible
      1. ethical reasons: what if we want to know the effects of lead exposure on children—can we set up an experiment with a treatment and a control group?
      2. practical reasons: does consumption of red wine reduce the risk of heart disease? could do a controlled trial but the results could take decades
      3. sheer impossibility of creating a treatment and control: what keeps two countries from going to war? trade links? democratic regimes? relative size of economies? you cannot create countries and test them, yet empirically it is an important question
   5. regression analysis is attempt to do mathematically what you can’t do experimentally: all else equal, what is the effect of this one causal variable?
   6. interpreting regression results
      1. intercept term—often not something that we worry about too much (what is the effect when all the coefficients are zero, but that’s not something that really happens), since sometimes it is meaningless (e.g., the negative weight in the height-weight example)
      2. regression coefficient—effect of that one independent variable on the dependent variable
      3. it is another form of classical hypothesis test, in that you are trying to reject the null hypothesis that the coefficient may be zero at some confidence level (e.g. 95%)for
      4. standard error of regression coefficient—confidence interval of your estimate, just as for estimates of the mean (so anything 2 standard errors away from zero is statistically significant at the 0.05 level; again using logic of repeated sampling) (formula for std error of regression coefficient is relatively complicated so let’s keep it at concept for now)
      5. there are different kinds of regression models—the idea is that the functional form needs to be linear, or is something that can be transformed into a linear equation (equation of degree 1)
      6. regression is a relatively forgiving modeling technique (“robust” in that changing the model specification often does not change your result) but you need to be self-conscious
   7. OLS regression
      1. common technique, and what we will talk about for now
      2. “ordinary least squares”
         1. fit an equation so that you minimize the sum of the squared residuals
         2. what is a residual? difference between the estimate and the measured value
         3. this was much more difficult to do before electronic computers, clearly
      3. OLS regression makes a number of assumptions about data though, that you need to be aware of
      4. this finds the Best Linear Unbiased Estimator (BLUE); that is, the tradeoff btw bias and variance is made on the side of variance
      5. advantage: OLS models are easy to interpret—what is the effect of a one-unit change in the independent variable on the dependent variable
3. **regression analysis: assumptions and pitfalls**
   1. OLS regression: assume that **dependent variable** is an interval measure (and in fact it is not supposed to be censored at a particular value like zero, and you can use techniques for when it is)
   2. sometimes this is more possible than others: something is clearly a scale, or it is a ratio measure like money or time (which are censored at zero, but let’s not worry about that now)
   3. OLS regression also assumes that the error term for each estimate is independent (i.e., is uncorrelated with the error term of other estimates) and has a mean (ie. “expected value”) of zero
   4. clearly you are assuming that the underlying relationship between the IV and DV is linear, rather than a higher order
   5. regression assumptions (although regression is pretty robust, and sometimes you can fudge these a bit)
      1. independent variables are continuous or dichotomous; dependent variable is quantitative, continuous, and unbounded (plus all variables are measured without error)
      2. all IVs have variance not equal to 0 (ie. there is some variation in their value)(duh, otherwise they’d be constants)
      3. no perfect multicollinearity between any two IVs
      4. for each set of values for the independent variables, the mean value of the error term is zero
      5. each IV is uncorrelated with the error term
      6. the variance of the error term for each set of values for the IVs is constant (homoskedasticity assumption)
      7. error terms for any two observations of the values of IVs are uncorrelated
      8. error terms for each set of values for the IVs are normally distributed
   6. so keep those in mind, even if we are not going to test you on them
   7. why is regression so common? because it can be a tool of inferring causal relationships
      1. yeah, correlation (even partial correlation) is not causation
      2. but regression expresses mathematically the idea of “all else equal, what is the effect of this particular variable on the outcome I care about?”
      3. e.g. all else equal, what is the effect of education on income?
         1. for causal inference you need to have a **theory**
         2. for this example, the theory is provided by economics theories about human capital, productivity, signaling effects, and so on
         3. so you can isolate the effect of education among other variables (region, age, gender, race, occupational category, risk…)
      4. in the social sciences, we often care about making these arguments with data
         1. example: why did so many voters seem to turn against the Affordable Care Act?
         2. Michael Tesler looked at ANES data from 1994 and again from 2010 and found that racial animus was the key difference; it had no relationship to health insurance attitudes in the 1990s but a strong relationship after Obama became president
         3. based on a theory going way back to 1964 from Phil Converse: most people use shorthand to decide their policy stances and base them on group affiliation
      5. note that in the lab we should have done a number of things that we did not (and that would be much easier in a statistical package like Stata or SPSS, and is probably easier with R)
         1. use case weights, as ANES suggests
         2. estimate standard errors that are clustered because of the sampling strategy for ANES
         3. still, regression is pretty robust and even if you are quick and dirty you can tell whether your theory can be disconfirmed by the data
         4. also quick and dirty: interval assumption on the dependent variable, favor/oppose vaccines in schools
      6. note on measures of fit: R-squared is typical measure in OLS
         1. how much of the total variance is accounted for by the variables in the model
         2. another way to think of it is—how much of the variation that we see in the dependent variable