Lecture 4: Estimation

Intro Stats with Nathan Favero

American University (Washington, DC)

August 2, 2024

Statistical inference

- Though we often observe only a limited sample of data, we are generally interested in properties/patterns that extend beyond our sample
- In modern social science, we are especially interested in understanding causality, which is often examined through the framework of “counterfactuals”
- Counterfactual: a hypothetical alternative to what actually occurred, where one or more (independent) variables takes on a different value
    - Example: In an experiment, you want to compare treated observations to the counterfactual of what the observations would have looked like if they hadn’t received the treatment

Statistical inference

- Two broad categories in the field of statistics:
  - Descriptive statistics: used to describe or summarize information about data that we've collected
  - Inferential statistics: used to draw (probabilistic) conclusions about (1) a counterfactual in which some variables have different values (for some observations) than those we observe in our sample or (2) a broader population based on a narrower sample of data that we've collected
  - In statistical inference, we can never learn the full truth with 100% precision; thus, statistical inference is all about estimating

Statistical inference

- Both fields (inferential and descriptive) of statistics use formulas to compute statistics that help us learn properties of whatever data we're interested in
- Key difference: with inferential statistics, we want to be able to quantitatively (and probabilistically) describe data that we don't have access to (data for a counterfactual or for an entire population)

Confidence intervals

- When you hear pollsters talk about a “margin of error,” they’re essentially referring to a confidence interval
- With confidence intervals, we can make statements like “we’re 95% confident that the population mean lies between 45 and 49, assuming we’ve made accurate assumptions in our statistical model” (some will quibble with this interpretation, but it’s good enough for me)

Confidence intervals

- We can also use confidence intervals to describe estimates of *associations* between variables
- Example question: Based on estimates from our *sample*, can we conclude that gender is related to government satisfaction levels in the *population*?
  - We need a confidence interval for the difference in average government satisfaction between men and women in the population
  - Suppose our confidence interval indicates that average satisfaction is between 2 and 5 points higher (on a 10-point scale) for women
    - It's therefore unlikely that there's no difference (between men and women), so we'd say the relationship between gender and satisfaction is "statistically significant"
  - If, instead, our confidence interval indicated the the difference is somewhere between -2 and 3, we'd conclude that no difference is a real possibility

Confidence intervals

Example: is job tenure related to wages?

- Sample of women in 1988
- Wages are measured in dollars/hour
- Job tenure is measured in years
- Type into Stata:

```
use https://www.stata-press.com/data/r14/nlsw88.dta
```

- First time, need to install Stata package (select package called `pr0041_2`):

```
findit corrci
```

- Then generate the confidence interval for the correlation:

```
corrci wage tenure
```

Confidence intervals

Example: is job tenure related to wages?

```
. corrci wage tenure

(obs=2,231)

                  correlation and 95% limits
wage   tenure        0.178     0.137    0.218
```

- The first number is the correlation between wage and tenure in the *sample*: 0.178
- We're 95% confident the *population* correlation is between 0.137 and 0.218
- Since the range (0.137, 0.218) doesn't include 0, we conclude that a correlation of 0 isn't likely; thus, there's a statistically significant correlation between wage and tenure

Inference & linear regression

$$Y = A + B \times X + \varepsilon$$

- When we run a regression, we'll only get an estimate of the "true model" (values of A and B) since our data will include random noise (ε) that we can't measure
- We make inferences about the "true model" regression coefficients using confidence intervals or significance tests

Inference & linear regression

$$Y = A + B \times X + \varepsilon$$

- For each coefficient, Stata output will show us a p-value (more on this in future lectures), in order to let us decide if we can reliably conclude there that the coefficient is non-zero
  - If the slope is non-zero ($B \neq 0$), then X and Y are related
- Stata output will also show us confidence intervals for each regression coefficient

Inference & linear regression

Example: does job tenure effect wages?

- Sample of women in 1988
- Wages are measured in dollars/hour
- Job tenure is measured in years
- Type into Stata:

```
use https://www.stata-press.com/data/r14/nlsw88.dta
reg wage tenure
```

Inference & linear regression

Source	SS	df	MS
Model	2339.38077	1	2339.38077
Residual	71762.4469	2,229	32.1949066
Total	74101.8276	2,230	33.2295191

Number of obs	=	2,231
F(1, 2229)	=	72.66
Prob > F	=	0.0000
R-squared	=	0.0316
Adj R-squared	=	0.0311
Root MSE	=	5.6741

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.1858747	.0218054	8.52	0.000	.1431138	.2286357
_cons	6.681316	.1772615	37.69	0.000	6.333702	7.028931

- We're 95% confident that the true slope coefficient lies between .14 and .23
    - We're 95% sure that an additional year on the job typically corresponds to a raise that is between 14 and 23 cents

Inference & linear regression

Source	SS	df	MS
Model	2339.38077	1	2339.38077
Residual	71762.4469	2,229	32.1949066
Total	74101.8276	2,230	33.2295191

Number of obs	=	2,231
F(1, 2229)	=	72.66
Prob > F	=	0.0000
R-squared	=	0.0316
Adj R-squared	=	0.0311
Root MSE	=	5.6741

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tenure	.1858747	.0218054	8.52	0.000	.1431138	.2286357
_cons	6.681316	.1772615	37.69	0.000	6.333702	7.028931

- For the slope coefficient, the p-value is 0.000, which is less that .05, so we can also say that the relationship betwen tenure and wages is statistically significant (and positive)
  - In other words, even after accounting for random noise in the data, we're confident that tenure is truly (positively) related to wages

Sampling

- Statistical inference is usually taught through discussing sampling, so let's go over some foundational sampling concepts
- Population: the universe of units you are interested in learning about
  - Example: all eligible voters in the US
  - Sometimes it is possible to observe or collect data for the entire population of interest (conduct a census)
  - When a census is too difficult or expensive, researchers instead obtain a sample
- Sample: the actual units you observe (have data for)
  - Example: the 1000 eligible voters surveyed by a polling firm

Sampling

- In practice, defining the population is often a bit tricky, particularly if you have a hypothetical population (see textbook) or if you have simultaneous interest in multiple levels of generalizing
- Example: A study of racial attitudes
  - Sample: 1000 US adults who were interviewed
  - You might be particularly interested in the current attitudes of (all) adults living in the US
  - You might also be interested (to some degree) in broader principles of human behavior exibited by individuals around world and in different time periods (past and future)
- Many empirical studies never explicitly identify their population and may imply different populations of interest (some more general than others) in different passages

Sampling

- Probability sampling (such as simple random sampling): describes sampling selection methods that rely on the random selection of units into the sample
    - Use of randomness in selection process allows us to model the properties of samples with probability theory and inferential statistics

Sampling

- Convenience sampling: units are included in the sample because of convenience rather than random selection
- Representative sampling: units are selected such that certain known demographic (or other) characteristics of the sample will resemble the broader population
  - Users of this approach hope that matching on certain known characteristics will lead the sample to resemble the population on other characteristics where the distribution in the population is unknown

Sampling

- No sample is perfect; often some mix of random, representative, and convenience sampling
- Often nonresponse or missing data makes it impossible for all units selected from the sampling frame to be included in the final sample
- The more closely the sampling process resembles the random selection process we assume in our statistical models, the more confidence we can have that the conclusions of our statistical models properly describe our sample