3. Regression Fundamentals
Part 2: Inference

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Describing relationships between variables

- Visualizing relationships: scatterplots
- Covariance and correlation
- Fitting lines using "loess" curves

Simple linear regression

- Line of best fit: minimizing the sum of squared residuals (OLS)
- Interpreting regression coefficients: intercept and slope
- Measuring quality of fit (R^2)

Tonight

Using regression estimates to make *inferences* about a population relationship:

- Hypothesis testing
- Confidence intervals

Tonight's sample datasets

We will refer to two datasets tonight (both found on Github):

1. `caschool.dta`: data on test performance, school characteristics, and student demographics for California school districts, 1998-99 (N=420). Same as Lecture 3 part 1.
2. `wage2.dta`: earnings and other characteristics of male workers in the 1980s - from the Current Population Survey.

The Coleman Report

In 1964, the Civil Rights Act commissioned a large-scale study that came to be known as **the Coleman Report** (1966). The aim was to describe inequalities in educational opportunities for Black and white students in the United States.

They analyzed data from 645,000 students in 4,000 schools and had two main findings:

- Large gaps in tested achievement between Black and white students.
- Differences in school inputs appeared to explain little of these gaps.

The takeaway was that differences in family background, <u>not</u> schools, were key to understanding (and eliminating) achievement gaps.

See https://www.chalkbeat.org/2016/7/13/21103280/50-years-ago-one-report-introduced-americans-to-the-black-white-achievement-gap-here-s-what-we-ve-le/

The Coleman Report

The Coleman Report launched a huge academic literature on the "education production function" examining the relationship between various "inputs" and educational "outputs." For example:

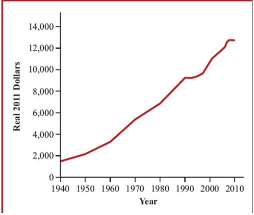
- Class size
- Teacher characteristics (experience, training)
- Technology and equipment
- Facilities
- Out-of-school factors: poverty, neighborhoods, peers

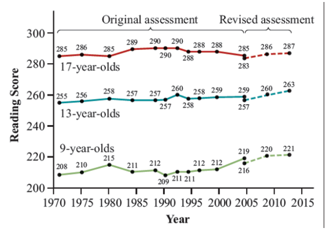
Put another way, these researchers are interested in inputs that have a positive *causal effect* on student outcomes. A large segment of this literature might be called the "does money matter" debate.

The Hanushek critique

In a series of papers beginning in the 1970s, Eric Hanushek has argued that there is not a compelling case that increasing spending on public schools improves student outcomes.

Figure 9.1 Trends in Real Expenditures Per Student, 1940–2010





This argument has been oversimplified as "money doesn't matter."

The Hanushek critique

The "Hanushek critique" is an empirically testable hypothesis. How would you test this using regression?

- What data would you use?
- What would your (population) model be?
- What would your (null) hypothesis be?

Stick to simple (one regressor) linear regression for now.

Inference in linear regression

As noted in part 1, linear regression is a great tool for description and prediction. Sometimes, however, we want to use it to make *inferences* about a population relationship.

Sample statistic	Population parameter
$\bar{x}$	μ
s^2	σ^2
r	ρ
Sample regression	**Population regression**
$y = a + bx$ or $y = \hat{\beta}_0 + \hat{\beta}_1 x$	$y = \beta_0 + \beta_1 x$

We use the OLS intercept and slope $(\hat{\beta}_0, \hat{\beta}_1)$ to estimate the *population* intercept and slope (β_0, β_1).

Inference in linear regression

Since we're using a sample to make inferences about the population, we have to quantify our uncertainty:

- Hypothesis testing
- Confidence intervals

Later: we'll consider what additional meaning can be attached to the regression. For example: is it a causal relationship?

Hypothesis testing

Review: hypothesis tests about μ

Recall how we conducted hypothesis tests about the population mean μ using the sample mean $\bar{x}$:

1. Find the *standard error* of $\bar{x}$: $\sigma/\sqrt{n}$, estimated using $s/\sqrt{n}$
2. Calculate the *test statistic*, which tells you "how far" $\bar{x}$ is from the hypothesized value of μ: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, where μ_0 is your null hypothesis value for μ.
3. Calculate the p-value for your test statistic if the null hypothesis is true. Reject the null if $p < \alpha$ (the significance level).

Assumptions: your data are a random sample from the population. Also, your sample size is large enough to assume that t has a standard normal distribution.

Hypothesis tests about the population slope β_1

The steps are the same for testing hypotheses about the population regression slope β_1 using the sample estimator $\hat{\beta}_1$! We just have to know the *sampling distribution* of $\hat{\beta}_1$ and assure that certain assumptions hold.

A two-sided hypothesis test about β_1:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0}$$

where $\beta_{1,0}$ is some hypothesized value for the slope, which could be zero.

Hypothesis tests about the population slope β_1

Our test statistic for this hypothesis test is:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

where SE is the standard error of $\hat{\beta}_1$. This t-statistic looks very similar to the one we used with $\bar{x}$.

The sampling distribution of $\hat{\beta}_1$

When the sample size is large $\hat{\beta}_1$ is distributed: $\hat{\beta}_1 \sim N(\beta, \sigma^2_{\hat{\beta}_1})$. What does this mean?

- $\hat{\beta}_1$ has a normal distribution.
- $\hat{\beta}_1$ is **unbiased**. On average, you get the true population slope β_1.
- The variance of $\hat{\beta}_1$ is denoted $\sigma^2_{\hat{\beta}_1}$. The standard error is estimated using $\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ (on the next slide).

This is good news, since normal distributions are easy to work with! In large samples, the t statistic has a standard normal distribution $N(0, 1)$.

The standard error of $\hat{\beta}_1$

The standard error of $\hat{\beta}_1$ is estimated as:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i{}^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Fortunately, this complicated-looking formula is calculated for you in Stata! You will recognize pieces of this, like the residuals $\hat{u}_i$, the sample size n, and something that looks like the variance of x in the denominator.

What happens as the sample size n gets larger?

Example 1

Suppose you estimate a slope of $\hat{\beta}_1 = 5$ and $SE(\hat{\beta}_1) = 2.1$

You are interesting in testing the hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

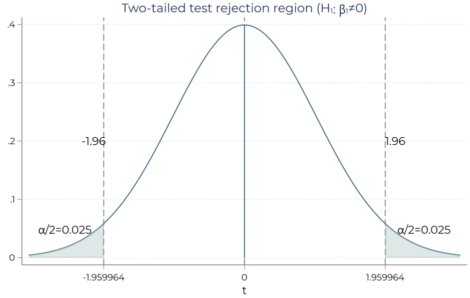
Your test statistic is:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

In large samples this will have a $N(0, 1)$ distribution.

Example 1

We will reject whenever $|t| > 1.96$:



Example 1

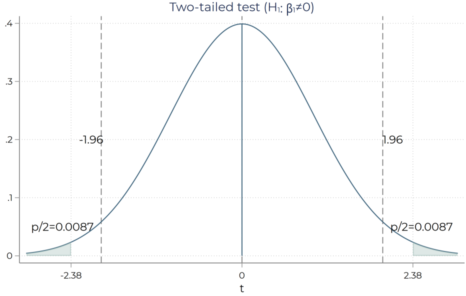
In this case:

$$t = \frac{5-0}{2.1} = 2.38$$

What is the probability of getting a t-statistic of 2.38 or larger (in either direction) if H_0 is true?

Example 1

The p-value is $0.0087 \times 2 = 0.0174$. If H_0 were true, there would only be a 1.74% chance of drawing a sample with this estimated slope of 5. We can **reject** H_0 at the 5% significance level.



Exercise: California schools data

Open the *caschool* data and do the following:

1. Calculate the least squares slope and intercept for the relationship between overall test scores (*testscr*) and expenditures per pupil (*expn_stu*). Interpret these.
2. Interpret the *standard error* for the slope.
3. Test the null hypothesis of **no** relationship between district test scores and spending (the "Hanushek critique").
4. Is the estimated slope an *educationally meaningful* relationship? How would you determine this?

Writing prediction equation with standard errors

Sometimes you will see estimated prediction equations written out with the standard errors in parentheses. For example:

$$testscr = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\, str$$

This facilitates easy calculation of a t-statistic, although getting the p value would require a lookup.

Confidence intervals

Review: confidence interval for μ

Recall how we constructed a confidence interval for μ using $\bar{x}$:

1. Find the *standard error* of $\bar{x}$: $\sigma/\sqrt{n}$, estimated using $s/\sqrt{n}$
2. The 95% confidence interval is:

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right]$$

The multiplying factor varies for other confidence levels.

- 1.64 for 90% confidence interval
- 2.58 for 99% confidence interval

Confidence intervals for the population slope β_1

The process is the same for confidence intervals about the population regression slope β. We again rely on the *sampling distribution* of $\hat{\beta}_1$ and its underlying assumptions.

The 95% confidence interval for β_1 is:

$$\left[\hat{\beta}_1 - 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \times SE(\hat{\beta}_1)\right]$$

Again, the multiplying factor varies for other confidence levels.

Confidence interval interpretation

A 95% confidence interval will contain the true population parameter in 95% of random samples.

It is also the set of values that cannot be rejected as null hypotheses in a two-tailed test.

Exercise: California schools data

Return to the *caschool* data and:

1. Find and interpret the 95% confidence interval for β_1.
2. Find and interpret the 90% confidence interval for β_1. Is it wider or narrower, and why?
3. Find and interpret the 99% confidence interval for β_1. Is it wider or narrower, and why?

Hypothesis tests and confidence intervals for β_0

One can also conduct hypothesis tests and construct confidence intervals for β_0, the population regression intercept. All that is needed is the standard error for β_0.

Since we are rarely interested in the intercept itself, I will omit these from the lecture notes. However, you can find them in the Stock & Watson text, and Stata will report these values.

Regression on a binary regressor

Regression on a dummy variable

When the explanatory variable is binary (0-1, a "dummy" variable), x only takes on two values. For example, in the *caschool* data, create an indicator variable for small ($str \leq 20$) versus large ($str > 20$) classes:

- $d = 0$: classes are large (> 20)
- $d = 1$: classes are small (≤ 20)

Regression on a dummy variable

```
. reg testscr d

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     15.05
       Model |  5286.87866         1  5286.87866   Prob > F        =    0.0001
    Residual |  146822.715       418  351.250514   R-squared       =    0.0348
-------------+----------------------------------   Adj R-squared   =    0.0324
       Total |  152109.594       419  363.030056   Root MSE        =    18.742

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           d |   7.185129    1.85201     3.88   0.000     3.544715    10.82554
       _cons |   649.9994   1.408711   461.41   0.000     647.2304    652.7685
------------------------------------------------------------------------------
```

There are only two possible predictions:

- When $d = 0$: $\hat{y} = a = 649.99$
- When $d = 1$: $\hat{y} = a + b = 649.99 + 7.19 = 657.18$

Note a is the mean of y for $d = 0$ (large classes), and b is the *difference in means* between those with $d = 0$ and $d = 1$ (small vs. large classes).

Regression on a dummy variable

Be cautious when interpreting the estimated slope $\hat{\beta}_1$ here, keeping in mind that d only takes on two values.

However, there is no change in our interpretation of the standard error, 95% confidence interval, t statistic, p value, etc. Note that the confidence interval is effectively that for the difference in two means ($d = 1$ vs. $d = 0$).

Exercise: *wage2* data

Open the *wage2* data and do the following:

1. Create a dummy variable that equals 1 for men with a college education or higher (16 or more years of education).
2. Estimate the least squares intercept and slope for the relationship between monthly earnings (*wage*) and this new college variable. Interpret the results.
3. Test the hypothesis that male workers with a college degree earn more than those without a college degree.
4. Report and interpret the 95% confidence interval for the slope.

Heteroskedasticity

Standard error under homoskedasticity

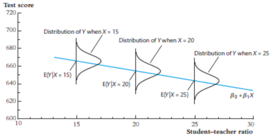
The standard error that Stata reports is actually based on a simpler calculation than the one shown earlier. It uses:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\beta_1}} = \sqrt{\frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

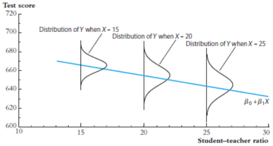
where the numerator is the standard error of the regression (SER). This formula assumes **homoskedasticity** which means the variance in errors are constant for different levels of x.

Homoskedasticity vs. heteroskedasticity

Figure 5.2 Homoskedasticity and Heteroskedasticity



(a) The errors are homoscedastic



(b) The errors are heteroskedastic

Implications of heteroskedasticity

There is rarely any theoretical reason to expect homoskedasticity. You can calculate robust standard errors in Stata using the `robust` option with `regress`. These are "robust" to the presence of heteroskedasticity.

If you don't do this, you run the risk of incorrect inference—i.e., your standard errors, confidence intervals, and hypothesis tests may be wrong!

Exercise: *wage2* data

Continue with the *wage2* data and do the following:

1. Create a scatterplot showing the relationship between monthly earnings and years of education.
2. Does this plot support the assumption of homoskedasticity? Why or why not?
3. Estimate the OLS intercept and slope between monthly earnings and years of education, and note the standard error on the slope.
4. Repeat the above, but use the `robust` option. How does this change things? Do your estimates of the intercept and slope change?

Next time

Linear regression with multiple regressors

- Read: Stock & Watson chapter 6 and 7
- Familiarize yourself with the three sample articles: Magnuson et al. (2004), Gershenson & Holt (2015), and Reber & Smith (2023).