Lecture 3: Linear Regression (Part II)

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

This Lecture

- In Lecture 1, we discussed three steps to empirical analysis

1. Identification
   - What assumptions are needed to get the causal parameter you want (in the population) given the data you observe?
2. Estimation
   - What is the estimation method by which you will calculate the parameter?
3. Inference
   - How can you quantify the uncertainty in the estimate (given that you do not have the population but only a sample)?

- This lecture focuses on the last step

Plan for Today

1. Variance of OLS estimators
   - Variance of OLS under homoskedasticity
   - Multivariate regression
2. Statistical inference
   - Sampling distribution of OLS estimators
   - The t-test
   - p-values and confidence intervals
   - Testing multiple linear restrictions: the F-test
3. Examining OLS coefficients in "real" examples

Variance of the OLS Estimator

- Unbiasedness is about the centre of the sampling distribution of our estimators $\widehat{\beta}_j$
- But what about the spread of this distribution?
  - If we draw repeated samples and estimate the beta's each time, we get the correct answer *on average*
  - But we would also like to know, on average, how far we expect the $\widehat{\beta}$ in a particular sample to be away from this true population value
  - For that we need to look at the variance
- The measure of spread we use is the variance
- Now we will calculate the variance of the OLS estimator

Variance of the OLS Estimator

- Recall that we had assumed **homoskedasticity**

$$Var(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I_n}$$

- Conditional on X, the variance of u is constant
- Remember: this assumption is not needed for OLS unbiasedness
- The reason to make it right now is that it makes calculating the variance quite easy and has some attractive efficiency properties

Variance of the OLS estimator

- For simplicity, let us think of the bivariate case: $y = \beta_0 + \beta_1 x + u$

$$Var\left(u|x\right) = E\left(u^2|x\right) - \left[E\left(u|x\right)\right]^2 = E\left(u^2|x\right) = \sigma^2$$

- This implies that $E\left(u^2|x\right)$ does not depend on x, so:

$$E\left(u^2|x\right) = E\left(u^2\right) = Var\left(u\right)$$

- Hence, σ^2 is often called the error variance or disturbance variance

- σ is the standard deviation of the error

- Sometimes useful to write this in terms of y:

$$E\left(y|x\right) = \beta_0 + \beta_1 x$$
$$Var\left(y|x\right) = \sigma^2$$

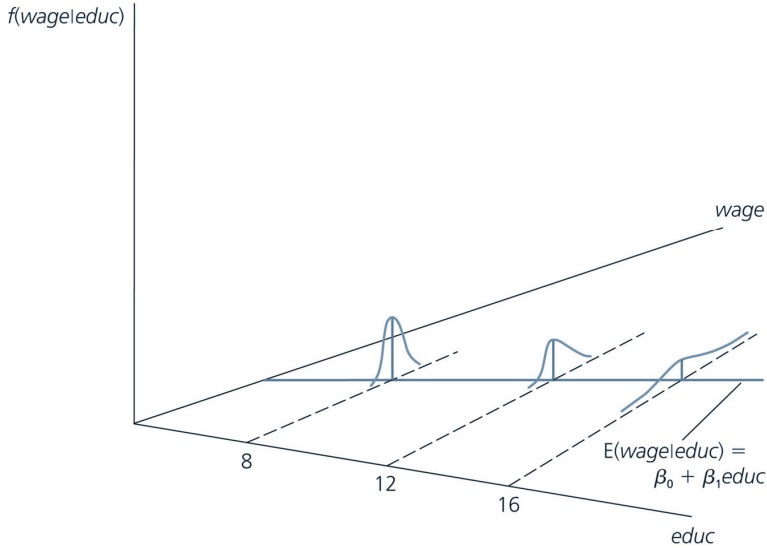
Failure of This Assumption

- If $Var\,(u|x)$ depends on x, we have something called **heteroskedasticity**
- Example:

$$wage = \beta_0 + \beta_1 education + u$$

  - Homoskedasticity implies $Var\,(wage|educ) = \sigma^2$
  - Average wage is allowed to increase with education (assuming $\beta_1 > 0$), but the **variability** of the wage around its mean is assumed to be constant for all education levels
  - Is this realistic?

Var(*wage*|*educ*) increasing with *educ*.



Sampling Variance of the OLS Estimators

- Under the assumptions we have made, we have that

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$
$$Var\left(\widehat{\beta}_0\right) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

- See Wooldridge (2013) for proofs

- Note these are invalid under heteroskedasticity, so in that case we would need a different calculation!

Sampling Variance of OLS Estimators

- Problem: σ^2 is unknown $\Rightarrow$ we need to estimate it
- It turns out that an unbiased estimator for σ^2 is (see Wooldridge 2013):

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2 = SSR/(n-2)$$

- SSR = Sum of Squared Residuals

The Standard Error

- For hypothesis testing we will use the standard deviation of our estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$
- These are usually called standard errors
- $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$ is the **standard error of the regression**
- We can use $\widehat{\sigma}$ to obtain the **standard error of the parameter** $\widehat{\beta}_1$ which tells us how precise our OLS estimator for β_1 is

$$SE\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}}{\sqrt{SST_x}} = \frac{\widehat{\sigma}}{\left(\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{1/2}}$$

$$\Rightarrow SE\left(\widehat{\beta}_1\right) = f(\overset{+}{\sigma^2}, \overset{-}{SST_x}), \text{ where } SST_x \text{ depends on } n$$

Sampling Variance of OLS Estimators in Multivariate Settings

- What if you have a multi-variate setting? This is also quite straightforward!
- Under the assumptions we have made, conditional on sample values of the independent variables:

$$Var(\widehat{\beta_j}) = \frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}$$

where $SST_j = \sum(x_{ij} - \bar{x_j})^2$ and R_j^2 is the R-squared from regressing x_j on all other independent variables and including a constant

Hypothesis Testing

- Consider the following model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u, \;\; t = 1, 2, ..., n$$

- The population parameters $\beta_0, \beta_1, \ldots, \beta_k$ are unknown
- We can hypothesize about β_j and use statistical inference to test whether the data support this hypothesis
- To perform statistical inference and test hypotheses, we need to know the full sampling distribution of $\widehat{\beta}_j$

How Do We Derive the Distribution of $\widehat{\beta}_j$?

- The Gauss-Markov assumptions allowed us to identify the first two moments of this distribution...
- ...but they are not sufficient to pin down the distribution of $\widehat{\beta}_j$
- Conditional on the sample values of the independent variables, the distribution of $\widehat{\beta}_j$ depends on the distribution of the errors
- Hence, we need to make an additional assumption on the distribution of the population error

The Normality Assumption

- Assumption: the population error u is **independent** of the explanatory variables $x_1, x_2, \ldots, x_k$ and is **normally distributed** with zero mean and variance σ^2

$$u \sim Normal\left(0, \sigma^2\right)$$

$\Rightarrow$ This is the strongest assumption so far!

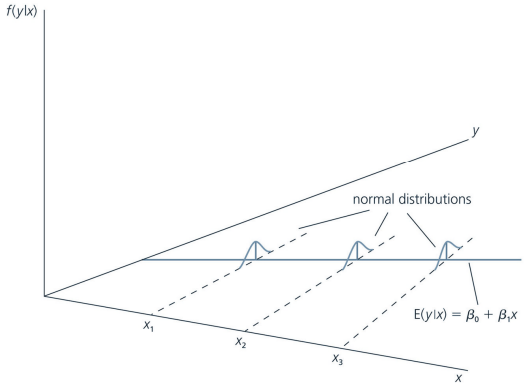
The Normality Assumption

- Independence implies $E(u|x_1, \ldots, x_k) = E(u) = 0$, as well as $Var(u|x_1, \ldots, x_k) = Var(u) = \sigma^2$
- In a cross-sectional context, assumptions MLR.1-MLR.6 are known as the classical linear model (CLM) assumptions
- Under these assumptions, the OLS estimators are minimum variance unbiased estimators

CLM Assumptions (Wooldridge 2013, p. 111)

- The CLM assumptions can be summarized as

$$y|\mathbf{x} \sim \textit{Normal}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2\right)$$



The Distribution of $\widehat{\beta}_j$

- Under the CLM assumptions, $\widehat{\beta}_j$ has the following distribution (conditional on the sample values of the independent variables):

$$\widehat{\beta}_j \sim Normal\left[\beta_j, Var\left(\widehat{\beta}_j\right)\right]$$

- How can we establish this result?
  - Conditional on $\mathbf{x}$, $\widehat{\beta}_j$ is a linear combination of the errors in the sample, which are independently and normally distributed
  - The sum of independent normal random variables is normal
  - Hence, $\widehat{\beta}_j$ is normally distributed

- We can also write:

$$\frac{\widehat{\beta}_j - \beta_j}{sd\left(\widehat{\beta}_j\right)} \sim Normal\,(0,1)$$

The Distribution of $\widehat{\beta}_j$

- This result turns out to be crucial:

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} \sim t_{n-k-1}$$

- Notice this differs from

$$\frac{\widehat{\beta}_j - \beta_j}{sd\left(\widehat{\beta}_j\right)} \sim Normal\,(0,1)$$

because $se\left(\widehat{\beta}_j\right)$ replaces the constant σ in $sd\left(\widehat{\beta}_j\right)$ by the random variable $\widehat{\sigma}$

Hypothesis Testing

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} = t_{\widehat{\beta}_j} \sim t_{n-k-1}$$

- $t_{\widehat{\beta}_j}$ measures how many estimated standard deviations $\widehat{\beta}_j$ is away from the null
- The null is less likely to be true if $\widehat{\beta}_j$ is far away from it, given the standard error $\Rightarrow$ Iff $t_{\widehat{\beta}_j}$ is large, we should reject the null
- But how large is large?

The Null Is Often Zero!

- In practice, we are often interested in testing the null hypothesis that x_j has no partial effect on y:

$$H_0 : \beta_j = 0$$

- The t-statistic then is:

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se\left(\widehat{\beta}_j\right)}$$

- This is a special case of the more general t-statistic that will allow us to test broader hypotheses

Step 1: Specify an **Alternative Hypothesis**

- The first step is to decide on the relevant alternative hypothesis (H_1)
- The following is an example of a **one-sided** H_1

$$H_1 : \beta_j > 0$$

- With this alternative hypothesis, we could also write the null as follows:

$$H_0 : \beta_j \leqslant 0$$

Step 2: Set a **Significance Level**

- Now, we need to pick a **significance level**
- This is the probability of rejecting the null when it is in fact true
- We call this type of error a **type I error** or a **false positive**
- Standard choices are 1, 5 and 10%—let us say we pick 5%

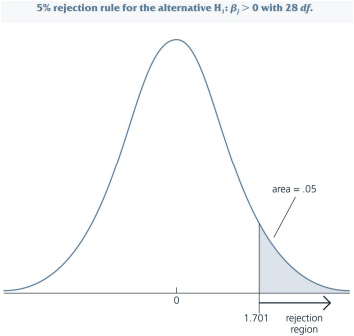
Step 3: Calculate t-statistic and Critical Value

- Calculate the t-**statistic**
- Under the null, we know $\beta_j = 0$, so:

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} = \frac{\widehat{\beta}_j}{se\left(\widehat{\beta}_j\right)} = t_{\widehat{\beta}_j} \sim t_{n-k-1}$$

- Calculate the **critical value** c
- For a one-sided test at the 5pct significance level, c is the 95th percentile in a t distribution with $n - k - 1$ degrees of freedom
- Reject H_0 in favor of H_1 if $t_{\widehat{\beta}_j} > c$

Testing Against One-Sided Alternatives (Wooldridge 2013)



- Increasing the significance level reduces c, making it easier to reject
- As $n - k - 1$ increases, the t distribution approaches the normal distribution

An Example

- Suppose we estimate

$$\widehat{\ln(wage)} = \underset{(0.104)}{0.284} + \underset{(0.007)}{0.092}\,education + \underset{(0.0017)}{0.0041}\,experience + \underset{(0.003)}{0.022}\,tenure$$

  where the numbers in parentheses are the standard errors of the estimated coefficients, and $n = 526$

- We want to test $H_0 : \beta_{exper} = 0$ against $H_1 : \beta_{exper} > 0$

An Example

- The t-statistic is $t_{experience} = 0.0041/0.0017 \approx 2.41$
- For a one-sided test, with 522 degrees of freedom, the 1% critical value is 2.326
- $t_{experience} > 2.326 \Rightarrow$ we reject the null that $\beta_{experience} = 0$ at the 1% significance level
- We can also say: "$\widehat{\beta}_{experience}$ is statistically greater than zero at the 1% significance level"

Two-Sided Alternatives

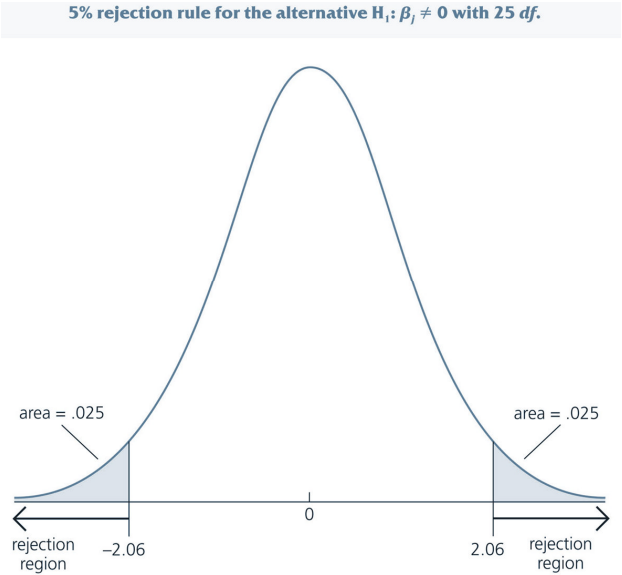
- In practice, we often want to test $H_0 : \beta_j = 0$ against the two-sided alternative
- This is the prudent choice
- With a two-sided alternative, we reject $H_0 : \beta_j = 0$ if

$$|t_{\widehat{\beta}_j}| > c$$

Two-Sided Alternatives

- We have to pick a **critical value** that takes into account the two-sided nature of H_1
- $\Rightarrow$ For a two-tailed test at the x% significance level, c is chosen so that the area in each tail of the t-distribution equals $\frac{x}{2}\%$
- For a test at the 5% significance level, c is chosen so that the area in each tail equals 2.5%
- If we reject $H_0 : \beta_j = 0$, we say that x_j is **(statistically) significant** or **significantly different from zero**

Two-Sided Alternatives (Wooldridge 2013)



Testing Other Hypotheses About β_j

- So far, we have focused on testing whether β_j is zero or not
- A more general null is

$$H_0 : \beta_j = a_j$$

  where a_j is our hypothesized value for β_j

- The t-statistic is

$$t = \frac{\widehat{\beta}_j - a_j}{se\left(\widehat{\beta}_j\right)}$$

- Proceed as before—it is only the calculation of the t-statistic that changes

p-values

- For a two-sided test, the p-value is

$$Prob\left(|T| > |t|\right)$$

  where t is the numerical value of the test statistic and $T \sim t_{n-k-1}$

- The p-value is the probability that a t-distributed random variable exceeds the calculated t-statistic

- In other words, the p-value tells us: if the null hypothesis is true, what is the probability of observing a t-statistic as extreme as we did?

p-values

- Another, perhaps more intuitive definition is: the p-value is the smallest significance level at which the null hypothesis would be rejected, given the observed value of the t-statistic

- If the p-value is low, this means that...
  1. We got a value of the t statistic that is unlikely when the null is true
  2. We can reject the null at a small significance level

- If the p-value is low, this is evidence against the null!

Example

- Suppose $df = 40$ and $t = 1.85$, then the p-value is

$$p\text{-value} = Prob\left(|T| > 1.85\right) = 2Prob\left(T > 1.85\right) = 2\left(0.0359\right) = 0.0718$$

- If the null is true we would observe an absolute value of the t-statistic as large as or larger than 1.85, 7.18% of the time

- With $p = 0.07$, we can reject H_0 at the 10% significance level, but not at the 5% level

- When you fail to reject H_0, **do not say that you "accept H_0"**

Confidence Intervals

- Since

$$\frac{\widehat{\beta}_j - \beta_j}{se\left(\widehat{\beta}_j\right)} \sim t_{n-k-1}$$

  a 95% CI for the unknown β_j can be constructed as

$$\left[\widehat{\beta}_j - c \cdot se\left(\widehat{\beta}_j\right), \widehat{\beta}_j + c \cdot se\left(\widehat{\beta}_j\right)\right]$$

  where c is the 97.5th percentile in a t_{n-k-1} distribution

- Interpretation: obtain repeated random samples, then β_j is covered by the calculated CI 95% of the time

Testing Multiple Hypotheses

- So far, we have focused on testing a single hypothesis
- In some cases, we want to test multiple hypotheses about the population parameters at once
- It is inappropriate to use q t-tests to conclude that multiple coefficients would be significant jointly
- We need a single test to check all these hypotheses jointly
- Suppose we are working with the following multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

An Example

- We want to test

$$H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0$$

- This joint hypothesis test entails q separate hypotheses

- H_1 : H_0 is not true $\Rightarrow$ At least one of the q variables is different from zero

The F-Test

- Impose the q exclusion restrictions to arrive at the restricted model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-q} x_{k-q} + u$$

- To test H_0, essentially we compare the SSR in the original ('unrestricted') model and in the restricted model

- Is the relative increase in SSR in the restricted model large enough to warrant rejecting H_0?

The F-test

- The F statistic is defined as

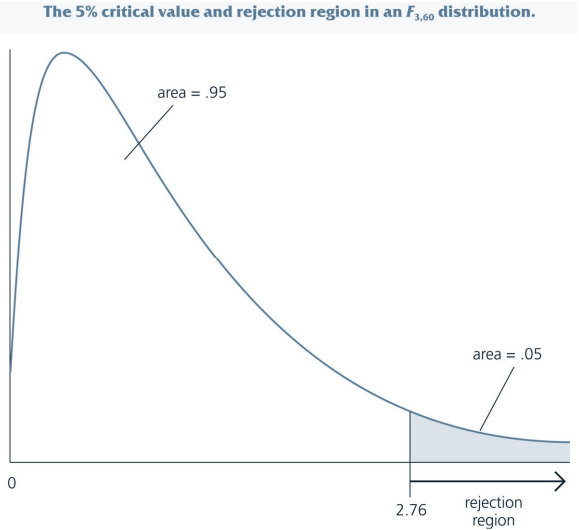
$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

- Under the null (and under CLM assumptions), F is distributed as an F random variable with $(q, n-k-1)$ degrees of freedom:

$$F \sim F_{q,n-k-1}$$

- Reject if F exceeds the critical value for your chosen significance level

The F-Test (Wooldridge 2013)



The F-Test and the Individual t-Tests

- Multicollinearity might lead to imprecise estimates and individually insignificant but jointly significant variables
- The reverse may also be true!
- An individual variable may be significant, but when it is mixed in with a bunch of insignificant variables, the resulting F-test may be insignificant

The F-Statistic for the Overall Significance of a Regression

- Here, H_0 is

$$H_0: \ x_1, x_2, \ldots, x_k \text{ do not help to explain } y$$
$$H_0: \ \beta = \beta_2 = \ldots = \beta_k = 0$$

- The restricted model is

$$y = \beta_0 + u$$

which always has a zero R^2

- This means the F-statistic can be written as

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where R^2 is the R^2 from the model that regresses y on $x_1, \ldots, x_k$

A Real-World Example: Gender Gaps in Match Achievement

- In many countries, boys score higher on math tests than girls
- To the extent this comes from different household investments or different investments in school, it is a matter of concern for educational policy
- Carneiro et al. (2017) look at gaps in math ability at early stages of schooling (from kindergarten till the 2nd grade) in Ecuador
- Specifically, they look at whether gender gaps differ by parental (esp. maternal) education and household wealth
  - Perhaps highly-educated mothers (fathers) do not invest differently into girls vs. boys?
  - Perhaps richer households do not invest differently into girls vs. boys?

Gender Gaps in Math Achievement (Carneiro et al. 2017)

- Carneiro et al. run the following specification:

$$\begin{aligned} Y_{ihgs} &= \alpha + \beta_1 E2_{ihgs} + \beta_2 E3_{ihgs} + \beta_3 Female_{ihgs} \\ &\quad + \beta_4 (E3 \times Female)_{ihgs} + \beta_5 Age_{ihgs} + \epsilon_{ihgs} \end{aligned}$$

where the subscripts denote child i, in household h, in grade g, and school s

- Y_{ihgs} is the test score
- E2: Mother has at least some secondary education
- E3: Mother has at least some university education
- Female: Dummy $= 1$ if child is a girl
- Age: vector of dummies for age in single months

Gender Gaps in Math Achievement (Carneiro et al. 2017)

Table 2: Mother's education, gender, and math achievement

	All grades	Kindergarten	First grade	Second grade
Dummy: girls	-.131***	-.072***	-.134***	-.165***
	(.019)	(.021)	(.022)	(.021)
Dummy: Secondary school	.264***	.288***	.271***	.278***
	(.023)	(.027)	(.028)	(.028)
Dummy: University	.465***	.479***	.448***	.552***
	(.044)	(.054)	(.051)	(.054)
Interaction: University*girls	.106**	.071	.131**	.093**
	(.052)	(.069)	(.056)	(.061)
F-test (p-value)	0.60	0.99	0.96	0.20

Note: Sample size is 31,398 in the all grades regression, and 10,466 in each of the grade-specific regressions. All regressions include age in months and its square. F-test is the p-value on an F-test that the sum of the coefficients on the dummy for girls and the interaction between girls and the dummy for mothers with university education is zero. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Gender Gaps in Math Achievement (Carneiro et al. 2017)

Table 3: Mother's education, father's education, wealth, and math achievement

	(1)	(2)	(3)	(4)
Dummy: girls	-.131***	-.130***	-.121***	-.131***
	(.019)	(.020)	(.018)	(.021)
Dummy: Mother secondary school	.264***			.177***
	(.023)			(.025)
Dummy: Mother university	.465***			.217***
	(.044)			(.060)
Interaction: Mother university*girls	.106**			.097
	(.052)			(.072)
Dummy: Father secondary school		.263***		.150***
		(.023)		(.022)
Dummy: Father university		.495***		.254***
		(.055)		(.058)
Interaction: Father university*girls		.040		0.015
		(.076)		(.082)
Wealth-Middle			.221***	.122***
			(.023)	(.027)
Wealth-Top			.510***	.327***
			(.045)	(.055)
Interaction: Wealth-Top*girls			.030	-.032
			(.056)	(.065)
F-test 1 (p-value)	0.60			0.63
F-test 2 (p-value)		0.22		0.16
F-test 3 (p-value)			0.10	0.02
Number of observations	31,398	24,504	31,398	24,504

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for fathers with university education is zero. F-test 3 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for the top wealth decile is zero. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

What We Have Covered

1. Variance of OLS estimators
   - Variance of OLS under homoskedasticity
   - Multivariate regression
2. Statistical inference
   - Sampling distribution of OLS estimators
   - The t-test
   - p-values and confidence intervals
   - Testing multiple linear restrictions: the F-test
3. Examining OLS coefficients in "real" examples

Readings

- For this lecture:
  - Review Chapter 4 and Chapter 5 (inference) of Wooldridge's book
  - Also review Section 2.25 of Cunningham's Mixtape (Variance of OLS estimators)

Optional

- Schady, N., Carneiro, P., & Cruz-Aguayo, Y. (2017). Where the Girls Are Not: Households, Teachers, and the Gender Gap in Early Math Achievement. Inter-American Development Bank.