Flinta* R-Tutorium

Unit 4 - Econometrics

Hannah Massenbauer

WU Wien

May 11, 2024

Econometrics

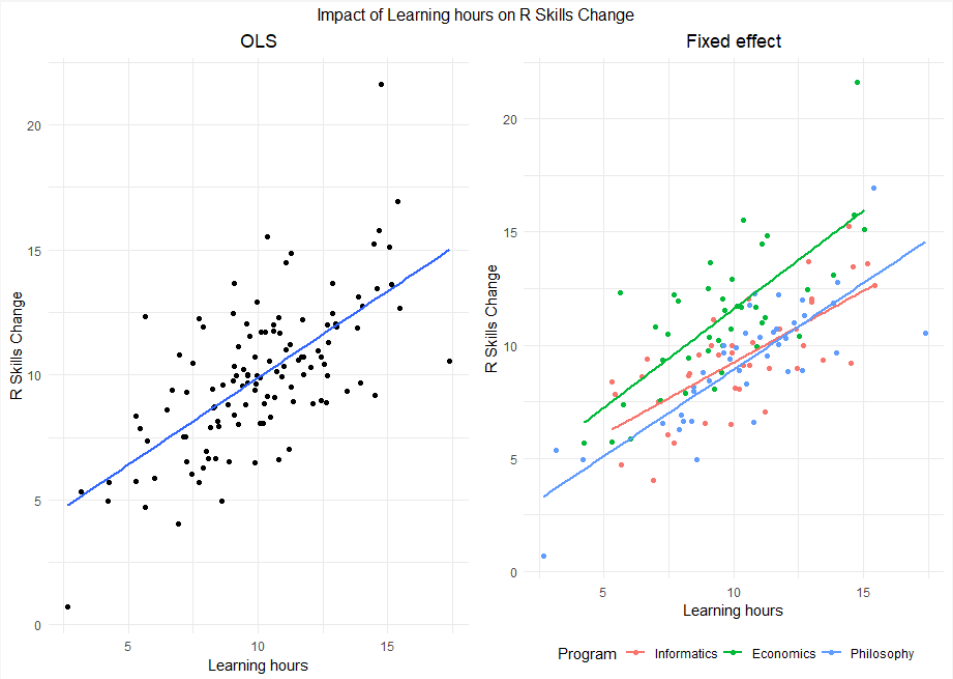
We have assumptions about the world that we would like to test. Econometrics offers tools to analyze data and optimally draw inferences.

Pose a hypothesis

H_0: Participants in this R-Tutorium will understand R better.
In 1 year your R-skills will be better than your colleagues', who have not participated.

Outlook for today

1. Types of data, measurement
2. Estimation method. Advantages and disadvantages.
    - OLS & fixed effects
3. Regression table what now?
4. Observing outcomes: Test statistics + graphical visualisations



Data

Using applied data poses many pitfalls. Here are two things to consider:

1. We only observe a limited sample

e.g. We observe only a handful of students in this course not the whole cohort

We observe only a sample, there remains uncertainty about the population! If we use a random, representative sample where the outcome variable is normally distributed we have large sample justification.

2. How to measure things?

e.g. How to measure the performance in R? In which unit? Is it measured by a test outcome, survey?

Measure matters. Think about what the variable captures and what it may not.

Estimation - What does it mean

Example I

How do your R skills change depending on the number of units you visit (0,1,2,3,4)

Example II

How does your grade change depending on the hours you learn?

$\rightarrow$ Most basic case: we estimate a linear relationship between two variables

$$f(x) = \hat{y} = \alpha + \beta \cdot x + \epsilon \tag{1}$$

$$\text{Change in R skills} = \alpha + \beta \cdot \text{Tutorium attendance} + \epsilon \tag{2}$$

Ordinary Least Squares

Most used estimation method, as it is easy to interpret and under circumstances the most efficient.

In this course we care about:

- Intercept: Value of y for $x = 0$.
- Interpreting coefficient: captures the relation between x and y
- Error term: Captures everything we are unable to explain with the included variables

Relationship between Residuals, Standard Errors, and Variance

- **Residuals**: Differences between observed and predicted values.

$$e_i = y_i - \hat{y}_i$$

- **Variance**: Reflects the variability of residuals across all observations. $Var(e_i)$
- **Standard Errors**: Measure the uncertainty in coefficient estimates.

$$se(\hat{\beta}_j) = \sqrt{\frac{Var(e)}{(n-k-1) \cdot Var(x_j)}}$$

- n: Number of observations.
- k: Number of predictors (how many variables you use to explain y).
- $Var(x_j)$: Variance of predictor variable x_j.

Test statistics

Shows how closely your observed data match the distribution expected under the null hypothesis

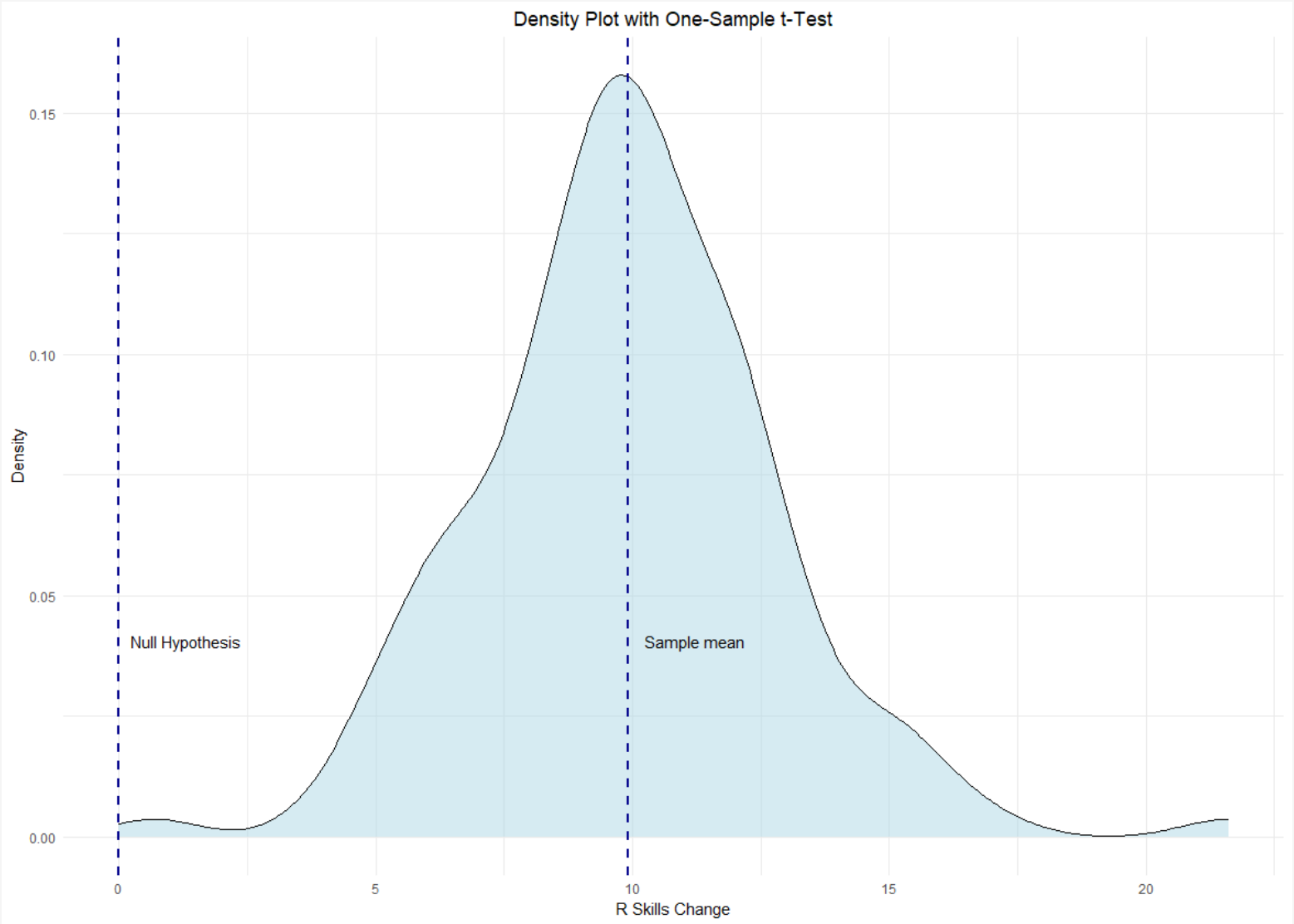
1. **T-test**: Is the observed mean likely to be representative of the population's mean from which the samples were drawn, or if they could have occurred by random chance.

$H_0 : \beta = 0$

In our context H_0 would imply attendance has no effect. We are able to reject that hypothesis or not.

2. **P-value**: Probability that obtained test statistic (t-test) is more extreme than the threshold. Thus, the p-value, depends on the chosen threshold (relates to significance levels e.g. 5%).

T-test. An Illustration



Glossary

- **Coefficient:** Slope of the linear regression line. When x changes how does y change?

$$\beta = \frac{\sum^{n} x_i y_i}{(\sum^{n} x_i x_i')} \tag{3}$$

- **Standard error:** Measures the precision of an estimate β. The smaller the standard error, the more precise the estimate.

$$se = \frac{sd}{\sqrt{N}} \tag{4}$$

- **T-statistics:** Significance of an estimated parameter (β) in a regression model. A larger absolute t-statistic suggests that our sample mean is more different from the hypothetical value ($\beta = 0$)

$$\text{T-stat} = \frac{\beta}{se} \tag{5}$$

- **Statistics** = inductive reasoning which is based on a sample
- **Parameter** = describes an aspect of the population (like a mean does under normality)

Helpful ressources

- Source for (almost) everything: `https://bookdown.org/mike/data_analysis/`
- OLS specific: `https://economictheoryblog.com/2015/04/01/ols_assumptions/`

Appendix - Fixed effects

There are some things we can't observe, but if they are constant over time, we can catch them by adding fixed effects.

Example

Philosophy students will have a different coding level than informatic students. The fixed effect α_i would be the study program if individual i.

$$\text{R-skill} = \text{program(i)} + \beta \cdot \text{Participation in Tutorium} + \epsilon \tag{6}$$

$\rightarrow$ Fixed effects model demean = subtract the mean value of a certain group, e.g. study program

$$y_{it} - \bar{y_i} = \beta(x_{it} - \bar{x_i}) + \epsilon_{it} - \bar{\epsilon_i} \tag{7}$$