The linear regression model, and model assumptions

Econometrics for minor Finance, Lecture 3

Tunga Kantarcı, Fall 2025

An empirical challenge: Educational outcomes

Can we improve student outcomes?

What are the policy tools?

Changing class size? Improving teacher quality? Investment in school?

An empirical challenge: Educational outcomes: Policies

Can smaller classes yield to higher test scores? What are the mechanisms?

Classroom management: Fewer students can mean fewer disruptions, leading to a better learning environment.

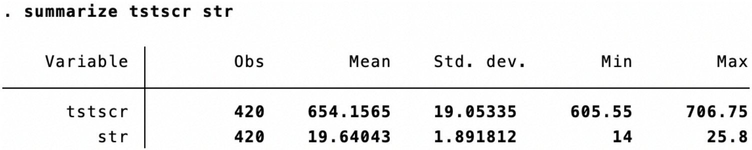
Teacher workload: With fewer students, teachers may be less overwhelmed and better able to prepare and assess.

Individual attention: Smaller classes may allow teachers to give more personalized instruction and feedback. Students might feel more engaged and less overlooked.

An empirical challenge: Educational outcomes: Data

The California Standardized Testing and Reporting (STAR) dataset contains data on test performance and school characteristics. The data are from districts in California collected in 1999 by the California Department of Education. Test scores are the average of the reading and math scores on a standardized test administered to 5th grade students. The student-teacher ratio is the number of students divided by the number of teachers working full-time in a district. The data is analyzed in Kruger and Whitmore (Economic Journal, 2001).

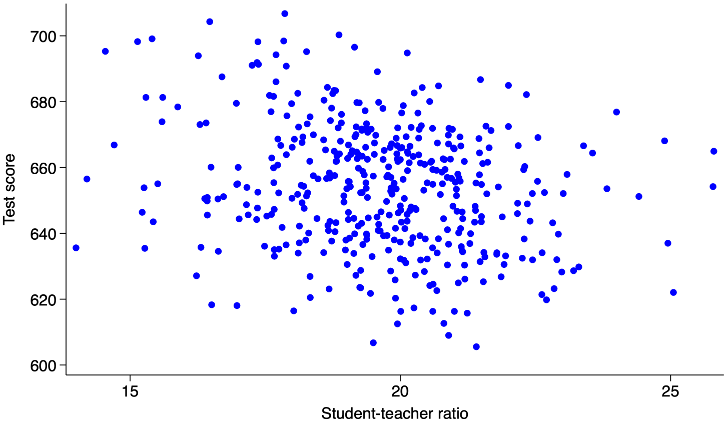
An empirical challenge: Educational outcomes: Data: Summary statsitics



```
. summarize tstscr str

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      tstscr |        420    654.1565    19.05335     605.55     706.75
         str |        420    19.64043    1.891812         14       25.8
```

An empirical challenge: Educational outcomes: Data: Scatter plot of test scores against student-teacher ratio

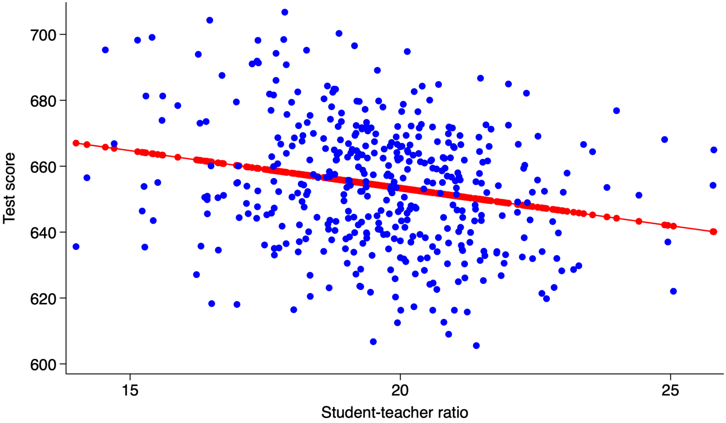


An empirical challenge: Educational outcomes: Model

Our aim would be to formally analyze how test scores vary with the student-teacher ratio. For this, we need to assume a model, but not just any model. It must be one that fits the data well and captures the underlying relationship we care about.

So there are two things. First, is the model, which provides a structured lens for understanding the relationship. Second is fitting, which ensures that this lens reflects the actual patterns in the data.

An empirical challenge: Model and fitted values



The linear regression model: Model

An observational unit is the entity on which measurements are taken, such as a student, worker, firm, or country. An observational unit is random. It becomes non-random once we sample and observe it.

The linear regression model: Model

Our model is

$$y_i = \beta_0 \cdot 1_i + \beta_1 x_i + u_i$$

for each of n observational units.

y_i : outcome: e.g., test score

β_0 : intercept term, unknown population parameter so needs to be estimated

1_i : a constant equal to 1, used to include the intercept term

β_1 : slope coefficient, unknown population parameter so needs to be estimated

x_i : predictor: e.g., class size.

u_i : error term, to capture factors not observed in the data, unobserved but we can approximate it

The linear regression model: Model

y_i also called dependent variable, or response variable

x_i also called independent variable, regressor, or explanatory variable

These names reflect convention, they do not imply causation. Regression describes conditional relationships, and sometimes causal ones.

The linear regression model: Model

The model

$$y_i = \beta_0 \cdot 1_i + \beta_1 x_i + u_i$$

is a population model. It provides the conceptual recipe for how the data we observe are generated in the population: the data generating process. That is, it assumes how the outcome y is produced from the predictor

$$x$$

together with unobserved influences captured by the error term

$$u$$

The population parameters

$$\beta_0, \beta_1$$

summarize the systematic part of this relationship in the population. Randomness enters through the sampling process of y_i and x_i.

The linear regression model: Model: Notation note

The model for observational unit i is

$$y_i = \beta_0 \cdot 1_i + \beta_1 x_i + u_i$$

As we typically have sample data on n observational units, we can stack the observations to obtain a compact model for n units

$$y = \beta_0 \cdot 1 + \beta_1 x + u$$

Moreover, although the intercept is technically multiplied by a vector of ones, we suppress it in notation and write:

$$y = \beta_0 + \beta_1 x + u$$

The linear regression model: Model: Assumptions

Like any other model, a regression model abstracts away complexity by reducing reality to a functional form. Models are approximations of the reality.

The linear regression model: Model: Assumptions

"All models are wrong, but some are useful."
— George E. P. Box

The linear regression model: Model: Assumptions

What all this implies is that we are making assumptions under which any regression model yields interpretable, credible, and sometimes causal insights about the world. A regression model is useful not because it is true, but because under the right assumptions it provides good approximations of relationships. If these assumptions fail, the model is not a good approximation of reality.

The linear regression model: Model: Assumptions: Linearity

Linearity. The model is linear in the parameters.

The linear regression model: Model: Assumptions: Linearity

The model

$$y = \beta_0 + \beta_1 x^2 + u$$

is

- linear in the parameters
- nonlinear in the regressor
- linear in the squared regressor

What makes a regression model linear is that it is linear in the parameters.

The linear regression model: Model: Assumptions: Linearity: Intuition

The model

$$y = x^{\beta} + u$$

is nonlinear in the parameter. This is not the linear regression model. Linearity means parameters enter additively and proportionally. If the true relationship looks more curved or complex, as it is here, this straight, proportional form is not a good assumption.

The linear regression model: Model: Assumptions: No perfect collinearity

No perfect collinearity. If there are multiple explanatory variables, they must be linearly independent. In other words, no column vector corresponding to one variable can be expressed as a linear combination of the column vectors of the other variables.

The linear regression model: Model: Assumptions: No perfect collinearity

If we include several explanatory variables in a regression model, each variable must contribute unique information. This means that no variable can be exactly predicted from the others. If one variable is just the sum or multiple of other variables, then it does not add anything new. When this happens, the regression model cannot separate the individual effects of the variables, and the estimator we want to use, to estimate the unknown parameters of the model, breaks down because it cannot distinguish which variable is responsible for changes in the outcome.

The linear regression model: Model: Assumptions: No perfect collinearity: Example

Consider the regression model

$$wage = \beta_0 \cdot 1 + \beta_1 d^{female} + \beta_2 d^{male} + u$$

where

$$d_i^{female} = \begin{cases} 1 & if \quad i = female \\ 0 & if \quad i = male \end{cases}$$

$$d_i^{male} = \begin{cases} 0 & if \quad i = female \\ 1 & if \quad i = male \end{cases}$$

Sum of the values in each row of d^{female} and d^{male} is equal to the value in that row of 1. Hence, one value can be perfectly predicted from other values. This is perfect collinearity.

The linear regression model: Model: Assumptions: No perfect collinearity: Example

$$\begin{bmatrix} 1 & d^{female} & d^{male} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

The linear regression model: Model: Assumptions: No perfect collinearity: Intuition

If we include a constant and both male and female dummies, the female dummy is always equal to 1 – male dummy. This redundancy makes it impossible for the regression model to separate their effects. Each regressor must bring unique information to the model.

The linear regression model: Model: Assumptions: Exogeneity

Exogeneity.

$$E[u \mid x] = 0$$

This condition says that, once we account for x in the model, the remaining error, u, is just random noise. It does not systematically lean up or down depending on x. In other words: given any value of x, the model is right on average. This assumption is also sometimes called the zero conditional mean assumption.

The linear regression model: Model: Assumptions: Exogeneity: The theoretical implication

This assumption has two implications. First, the law of iterated expectations requires that

$$\mathrm{E}\left[u\right] = \mathrm{E}_X\left[\mathrm{E}\left[u \mid x\right]\right]$$

Since

$$\mathrm{E}\left[u \mid x\right] = 0$$

we have

$$\mathrm{E}\left[u\right] = 0$$

The linear regression model: Model: Assumptions: Exogeneity: The theoretical implication

Second, the law of iterated expectations requires that

$$\mathrm{E}\left[ux\right] = \mathrm{E}_X\left[\mathrm{E}\left[ux \mid x\right]\right] = \mathrm{E}_X\left[x\mathrm{E}\left[u \mid x\right]\right]$$

Since

$$\mathrm{E}\left[u \mid x\right] = 0$$

we have

$$\mathrm{E}\left[ux\right] = 0$$

The linear regression model: Model: Assumptions: Exogeneity: The Theoretical implication

Now, from See Math refresher B in Wooldridge, we know that

$$\mathrm{Cov}\left[u, x\right] = \mathrm{E}\left[ux\right] - \mathrm{E}\left[u\right]\mathrm{E}\left[x\right]$$

Since

$$\mathrm{E}\left[ux\right] = 0$$

and

$$\mathrm{E}\left[u\right] = 0$$

we have

$$\mathrm{Cov}\left[u, x\right] = 0$$

So the assumption implies that u is not correlated with x.

The linear regression model: Model: Assumptions: Exogeneity: Intuition

Suppose we want to predict exam scores from study hours. Suppose students who study more also have hidden advantages, like natural ability. Since they are hidden, they show up in the error term. This makes the error correlated with study hours. The result is that the regression model mixes hidden ability with effort. So the slope is biased.

The linear regression model: Model: Assumptions: Homoskedasticity

In Greek, homo means same, and skedasis means dispersion. Hence, homoskedasticity means same dispersion. Constant variance.

The linear regression model: Model: Assumptions: Homoskedasticity

Homoskedasticity.

$$\text{Var}\left(u \mid x\right) = \sigma^2$$

This condition says that the error term u has the same variance at all values of x. It means that, regardless of the level of x, the variability of u is constant.

The linear regression model: Model: Assumptions: Homoskedasticity: Intuition

Suppose we want to predict wages from education. Individuals with little education face limited job options, so their wages vary less. The more educated can access a wider variety of jobs paying different wages, so their wages vary more. This means the variance of u depends on x. The result is that the regression misjudges its own precision: when errors vary with education, it becomes harder to get a precise prediction of how wages vary with education.

The linear regression model: Model: Assumptions: Normality of errors

The errors are normally distributed. u has the mean and variance given by the exogeneity and homoskedasticity assumptions, and has a normal distribution:

$$u \mid x \sim N\left[0, \sigma^2\right]$$

We will use this assumption if n is small but drop it if n is large. This is a technical assumption. We will get back to this.

Sampling: Assumption: Random sampling

Random sampling. The data we collect

$$\{(x_i, y_i) : i = 1, 2, \ldots, n\}$$

is a random sample of the population that follows the population model. It says that each unit in the population has equal probability of selection, and selections are independent: observations are independently and identically distributed.

Sampling: Assumption: Random sampling: Intuition

Suppose we want to know about the average education in the population. If only sample university students, the sample no longer represents the population. Then the model predictions will be systematically biased.

Sampling: Assumption: Random sampling: It does not guarantee representativeness

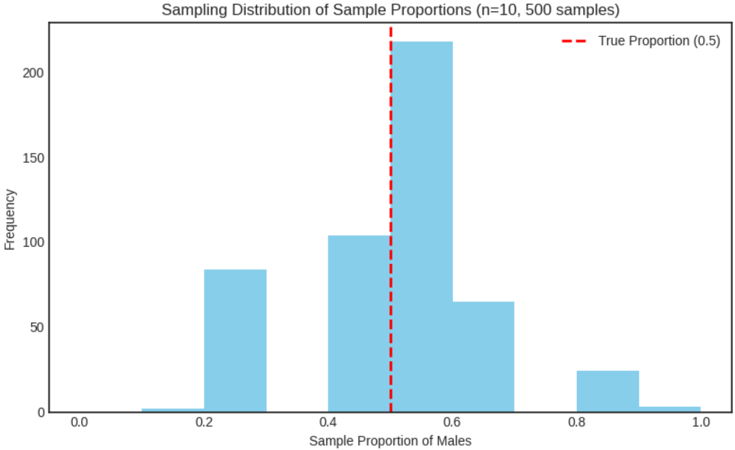
- Randomness does not eliminate sampling variation. Any single random sample may be unrepresentative, we can be unlucky. Representativeness is probabilistic, not absolute.
- Random sampling ensures the sample is expected to represent the population. Across repeated samples, the sampling distribution of statistics, e.g., means, proportions, etc. converges to the population values.
- Randomness protects against systematic bias, but not against chance variation in one particular sample.

Sampling: Assumption: Random sampling: It does not guarantee representativeness: Example

A single random sample can be unrepresentative, but the distribution of a sample statistic across repeated samples centers on the true population value:

- In a 50/50 male-female population, one random sample of size $n = 10$ might yield 7 males and 3 females ($\hat{p} = 0.7$), which is not representative of $p = 0.5$.
- If we repeat this process hundreds of times, the sampling distribution of $\hat{p}$ clusters around the true proportion $p = 0.5$.
- Random sampling guarantees representativeness in expectation, not in every single draw.

Sampling: Assumption: Random sampling: It does not guarantee representativeness: Example


Sampling Distribution of Sample Proportions (n=10, 500 samples)
True Proportion (0.5)
Frequency
Sample Proportion of Males