THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL OF PUBLIC POLICY

Mini-lecture 4

DAC 2023

Today's **Agenda**

1. Bivariate Regression
2. Goodness of Fit
3. Confidence Intervals
3. T-statistics and P-values

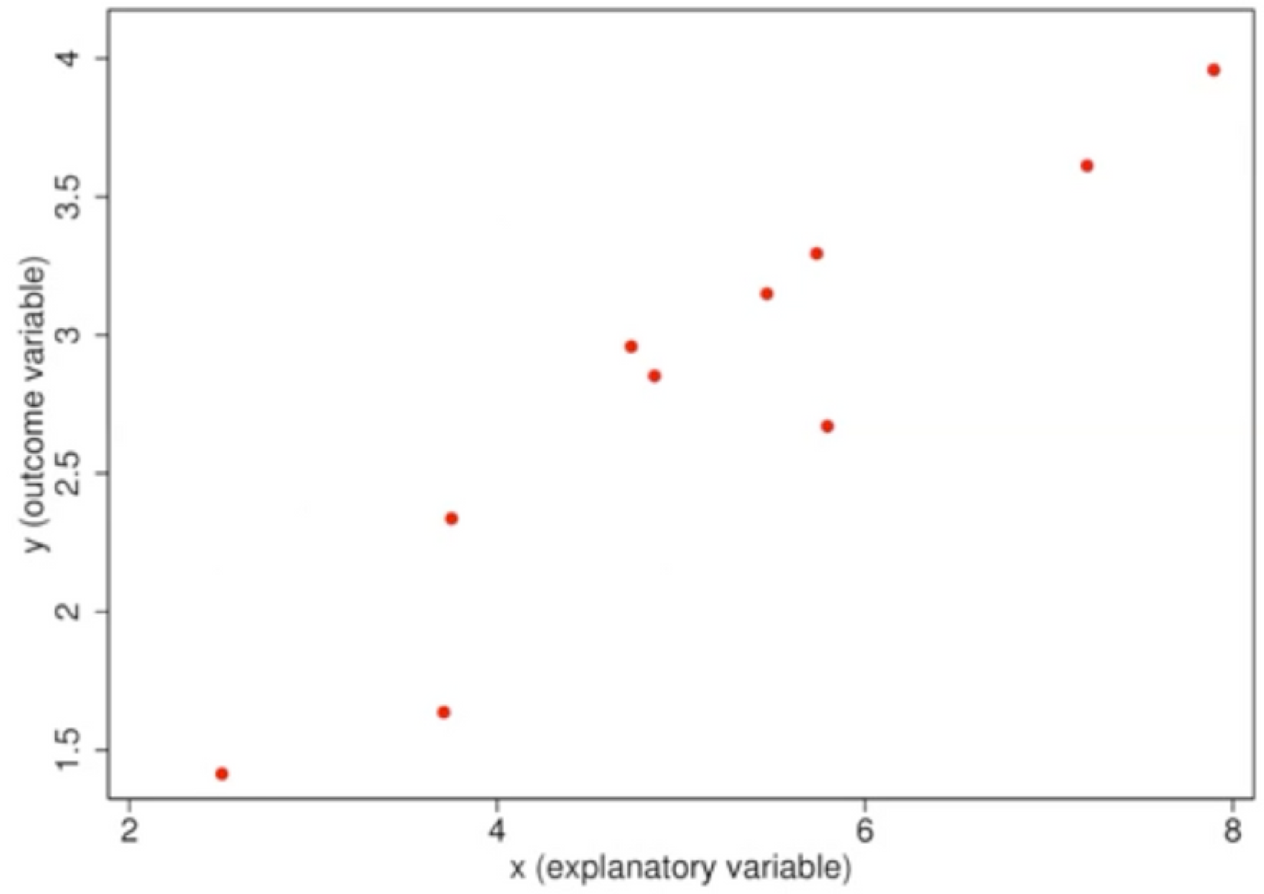
Bivariate Regression

Bivariate regression

- **Regression** is a statistical method to assess:
  - The relationship between two variables and its strentgth
- **When** is regression a good model (a model with explanatory power)?
  - When we can assume a linear relationship between our variables

Bivariate regression

- **Regression** is a statistical method to assess:
  - The relationship between two variables and its strentgth
- **When** is regression a good model (a model with explanatory power)?
  - When we can assume a linear relationship between our variables



- What does linear mean here?
  - This means that a change in one variable corresponds to a proportional change in the other variable

Bivariate regression

- **The formula:** $Y_i = \beta_0 + \beta_1 X_i + u_i$

Bivariate regression

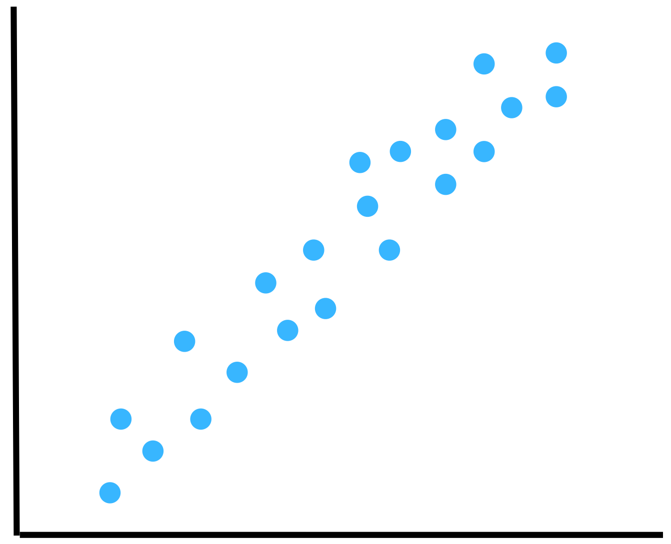
- **The formula:** $Y_i = \beta_0 + \beta_1 X_i + u_i$
- **What is it saying to us?**
  - Plain english interpretation

Bivariate regression

- **The formula:** $Y_i = \beta_0 + \beta_1 X_i + u_i$

- **What is it saying to us?**
  - Plain english interpretation

If the relationship between
the two variables is **linear...**

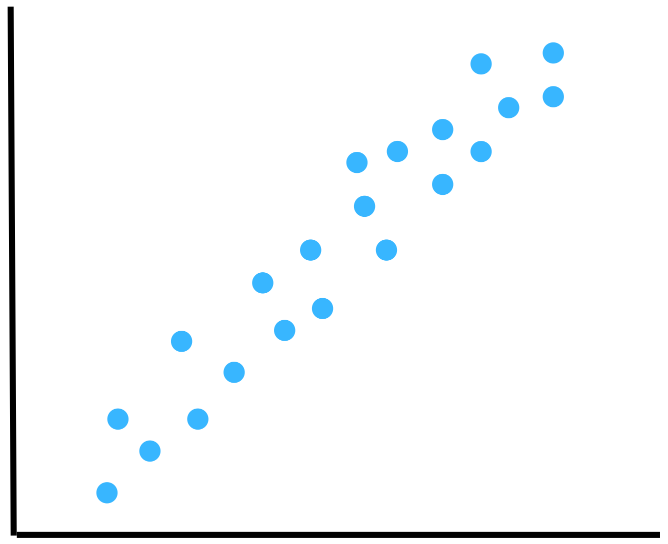


Bivariate regression

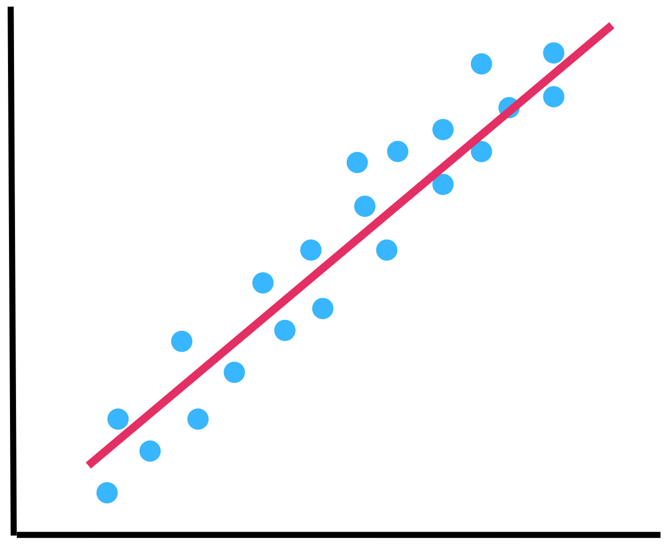
- **The formula:** $Y_i = \beta_0 + \beta_1 X_i + u_i$

- **What is it saying to us?**
  - Plain english interpretation

If the relationship between the two variables is **linear...**



Regression can fit a line...

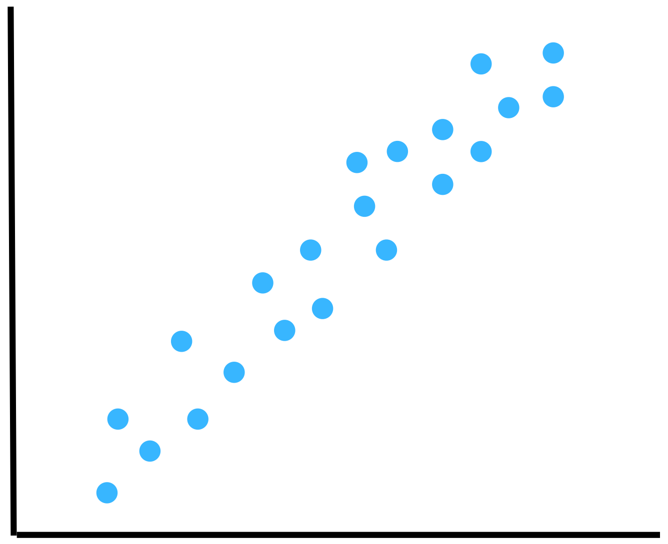


Bivariate regression

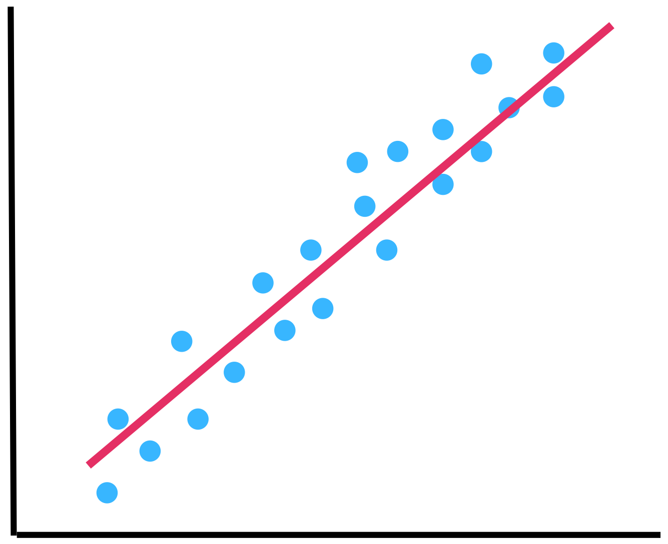
- **The formula:** $Y_i = \beta_0 + \beta_1 X_i + u_i$
- **What is it saying to us?**
  - Plain english interpretation

By numerically, we should understand it gives us an **intercept** (where does the line intercept the Y-axis) and a **slope** (by how much Y is increased by a one-unit increase on the X (independent variable)

If the relationship between the two variables is **linear...**

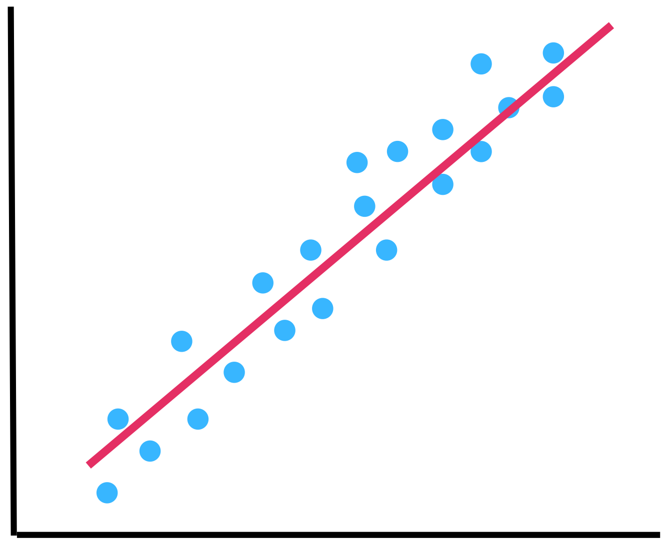


Regression can fit a line...



That **numerically** explains the relationship between our two variables

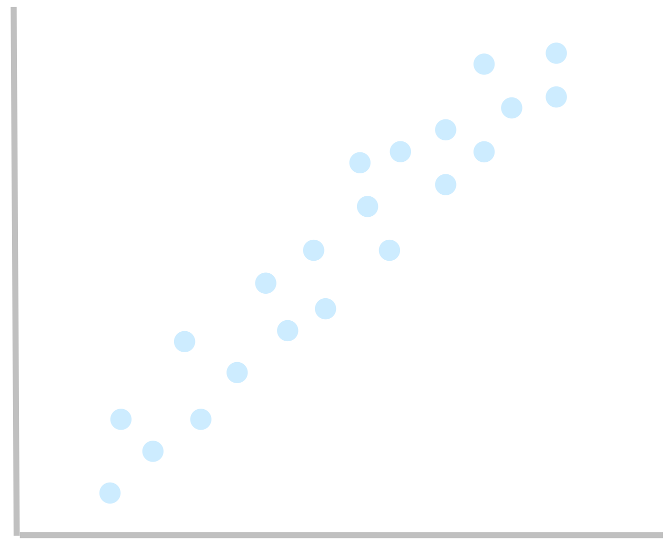
*Y = Bo + B1*X1 +ui*



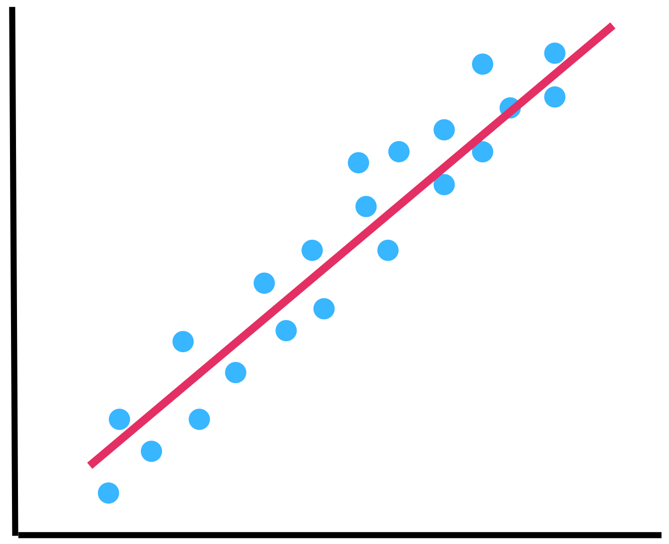
How does it uncover the numeric relationship?

How does it uncover the numeric relationship?

If the relationship between the two variables is **linear…**



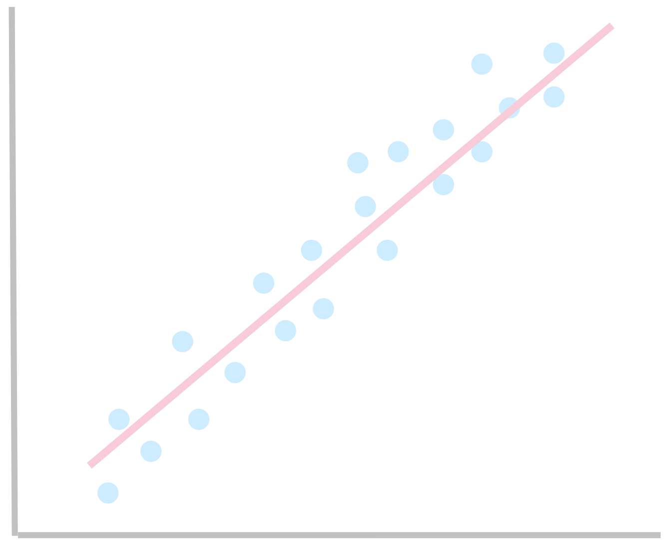
Regression can fit a line…



Step 2!

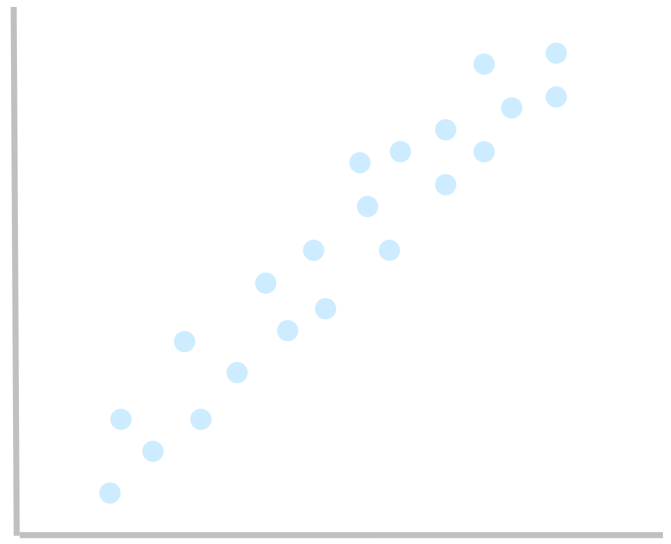
That **numerically** explains the relationship between our two variables

*Y = Bo + B1*X1 +ui*

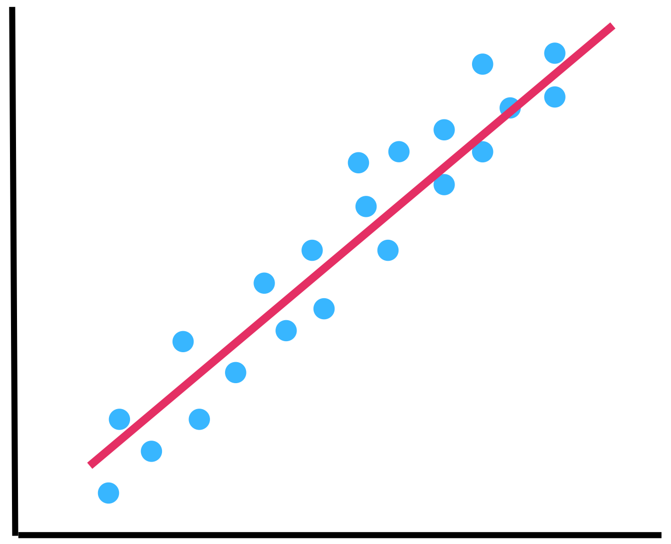


How does it uncover the numeric relationship?

If the relationship between the two variables is **linear...**

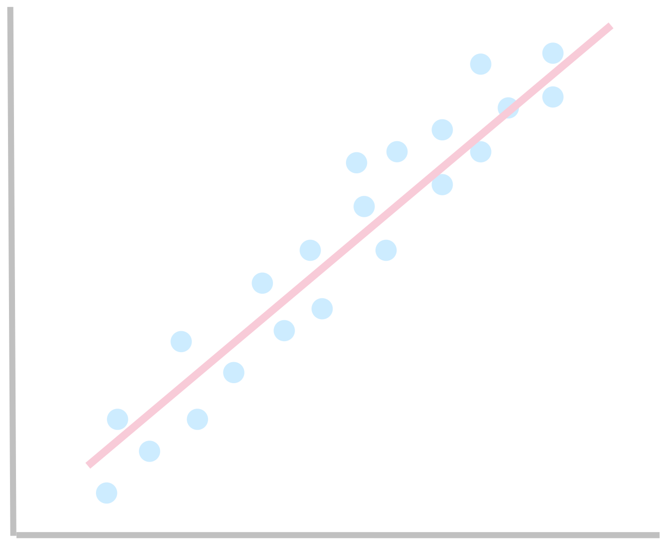


Regression can fit a line...



That **numerically** explains the relationship between our two variables

*Y = Bo + B1*X1 +ui*

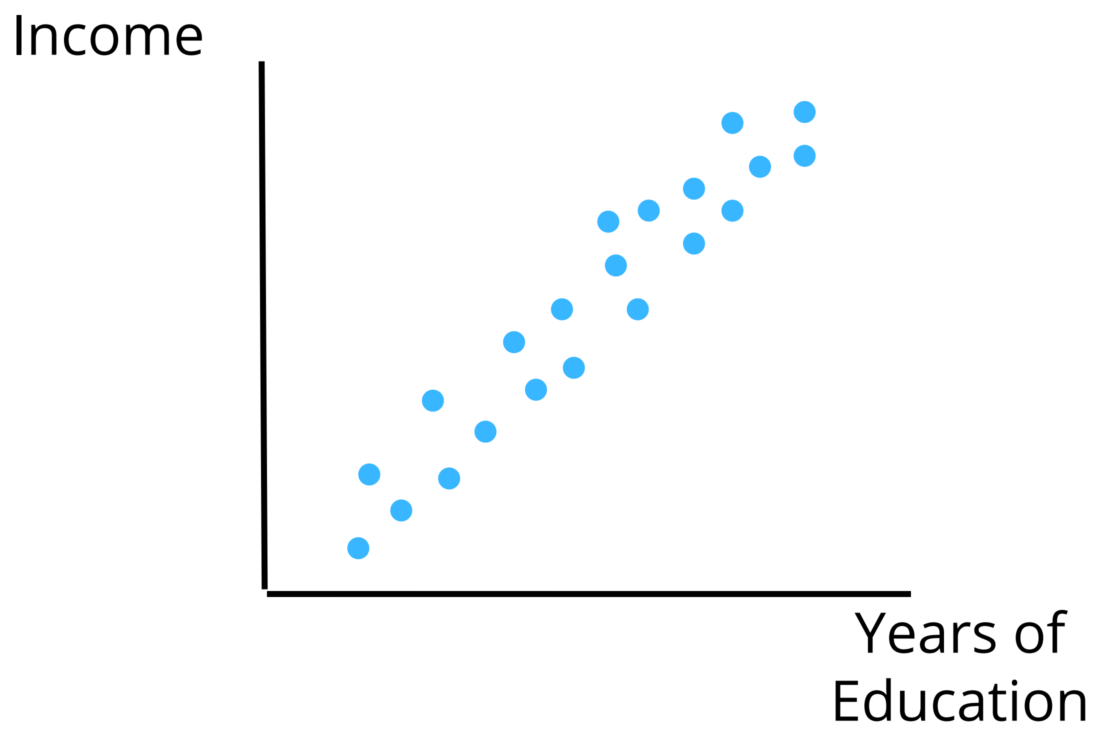


Step 2!

- This is when the model will discover what is the numeric relationship
  - The "line of best fit" is precisely a method to plug numbers into the linear model
  - **OLS!**
  - It turns out that minimizing the squared sum of the residuals (=OLS) is a very good guess of what Y (dependent variable) would be given our X (independent variable)

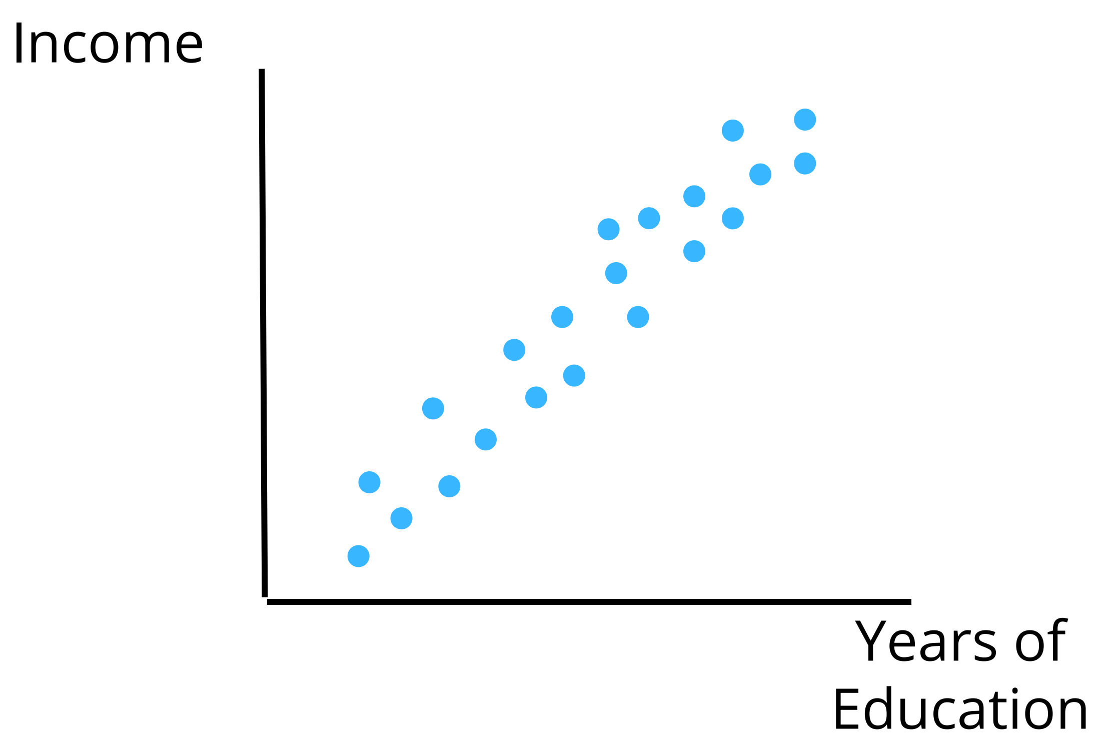
Example!

If we were trying to understand by **how much** increasing the years of education of a citizen impacts his/her income

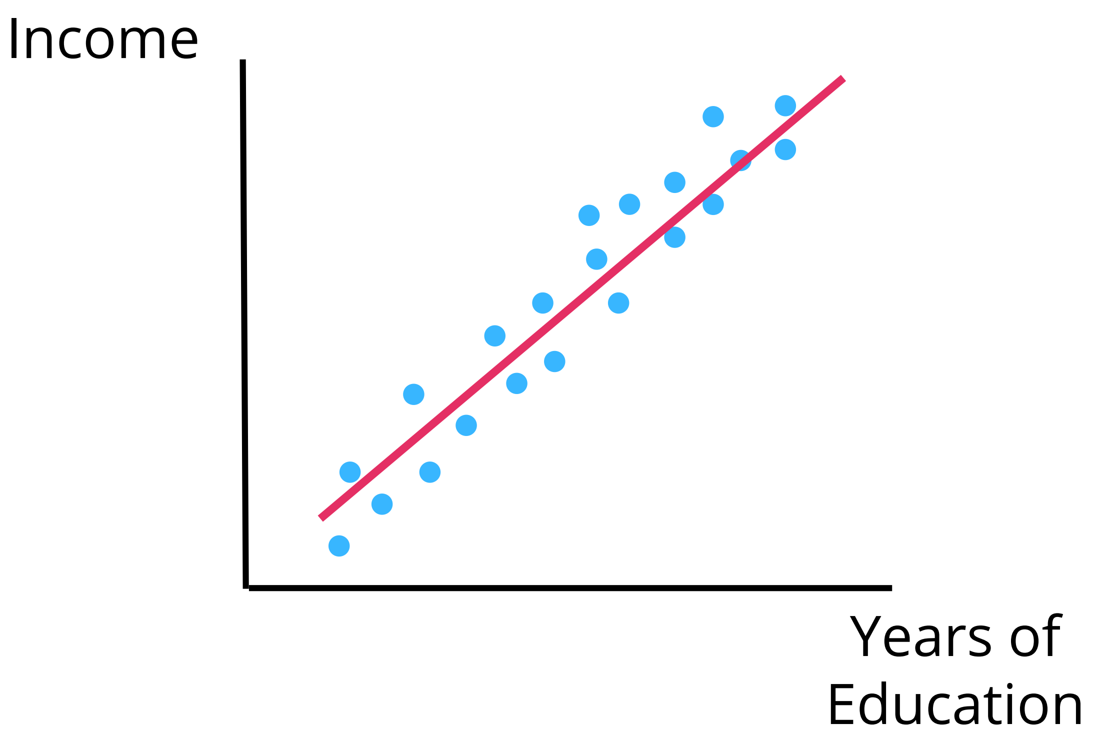


Example!

If we were trying to understand by **how much** increasing the years of education of a citizen impacts its income

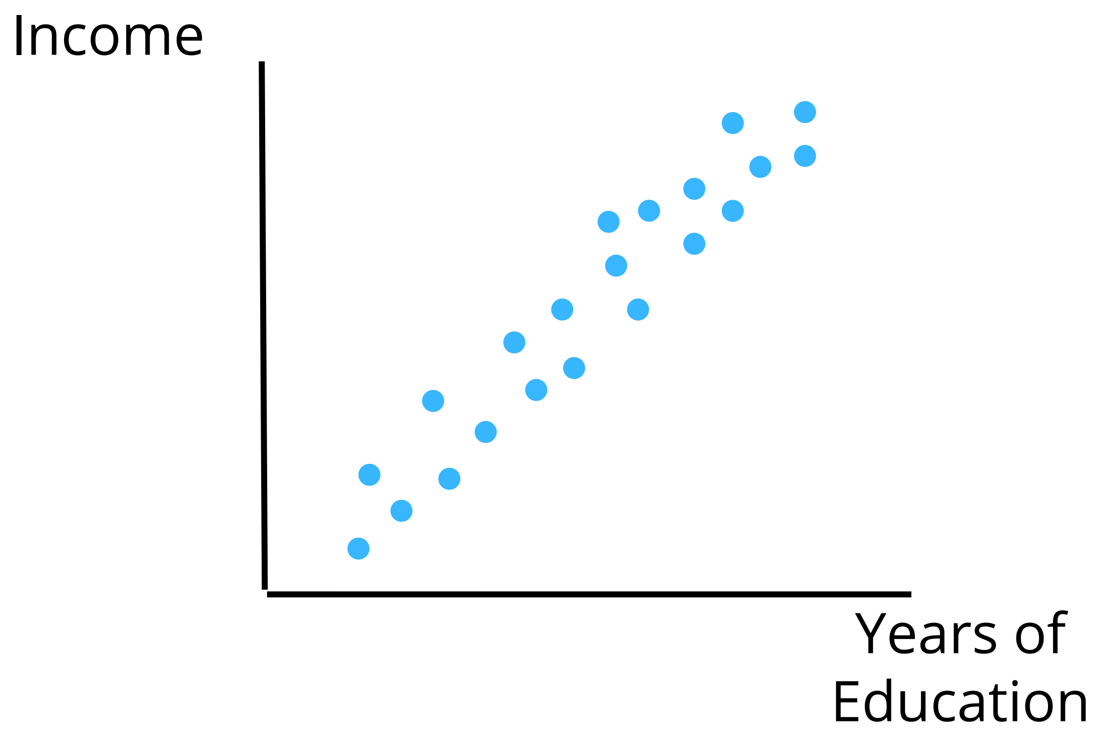


We would run our OLS model - that would fit a line between the variables

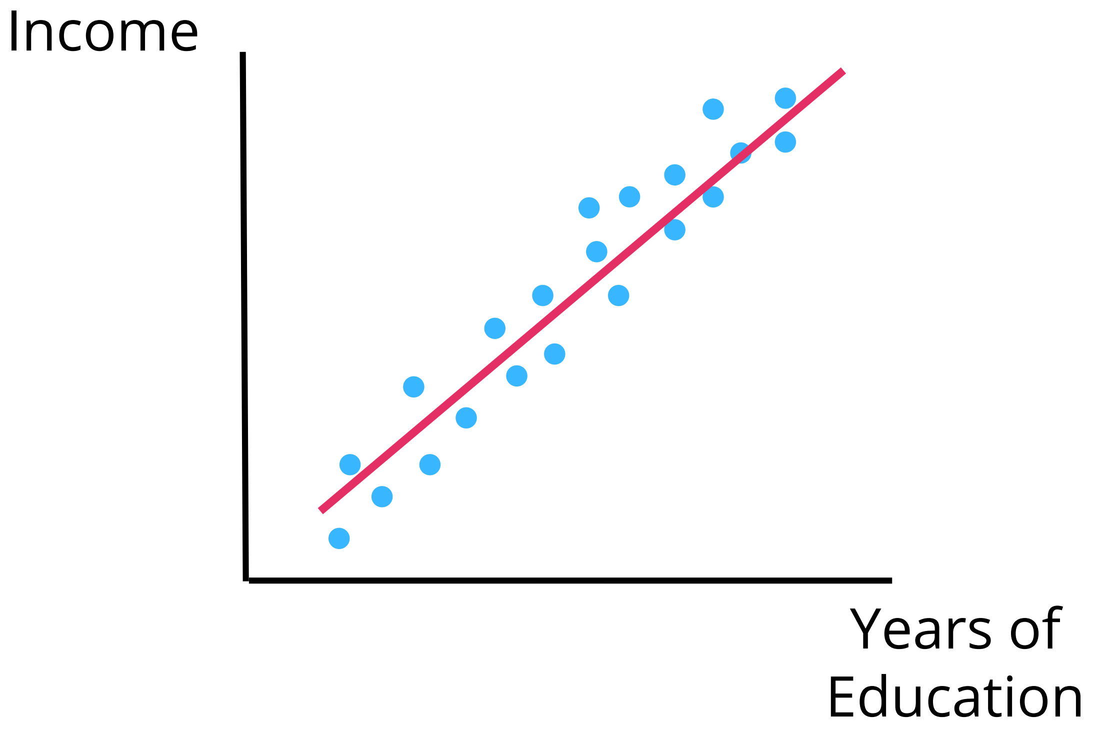


Example!

If we were trying to understand by **how much** increasing the years of education of a citizen impacts its income

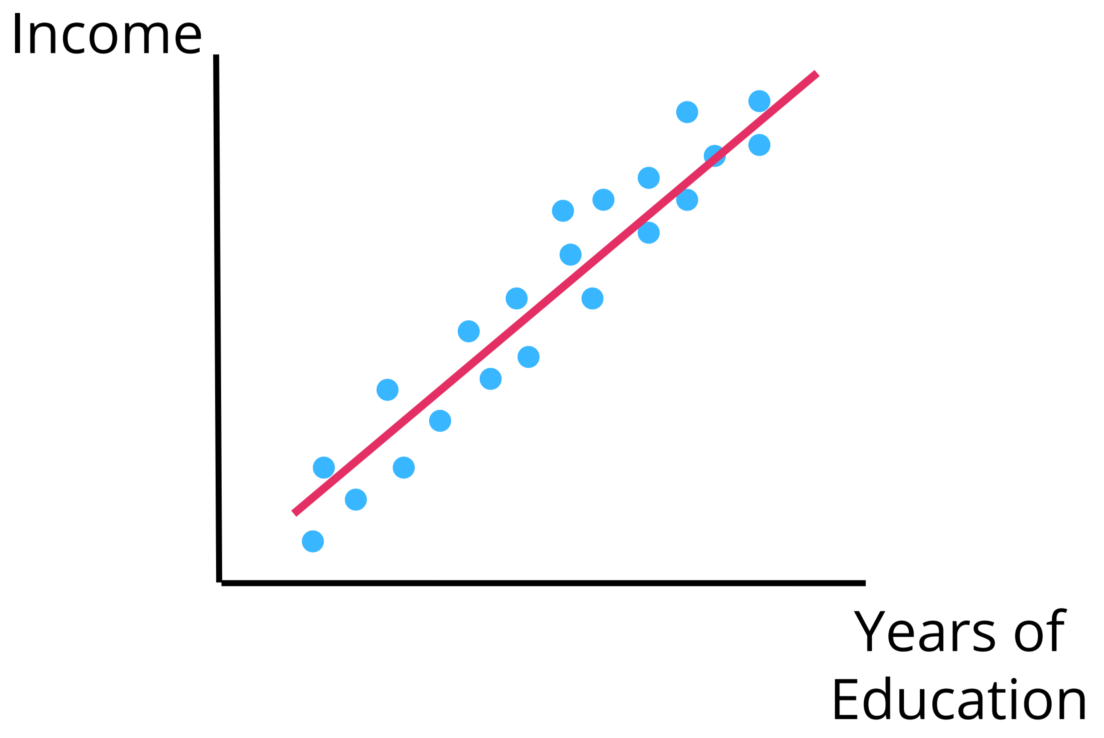


We would run our OLS model - that would fit a line between the variables



That **numerically** explains the relationship between our two variables

*Income = 10,000 + 15,000 * Years of Education + ui*



Goodness of Fit

Goodness of Fit

- **R-squared** measures the proportion of the total variation in the dependent variable (TSS) that is explained by our independent variable (ESS)

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

(RSS): measures the unexplained variability in the dependent variable

(TSS): measures the total variability in the dependent variable

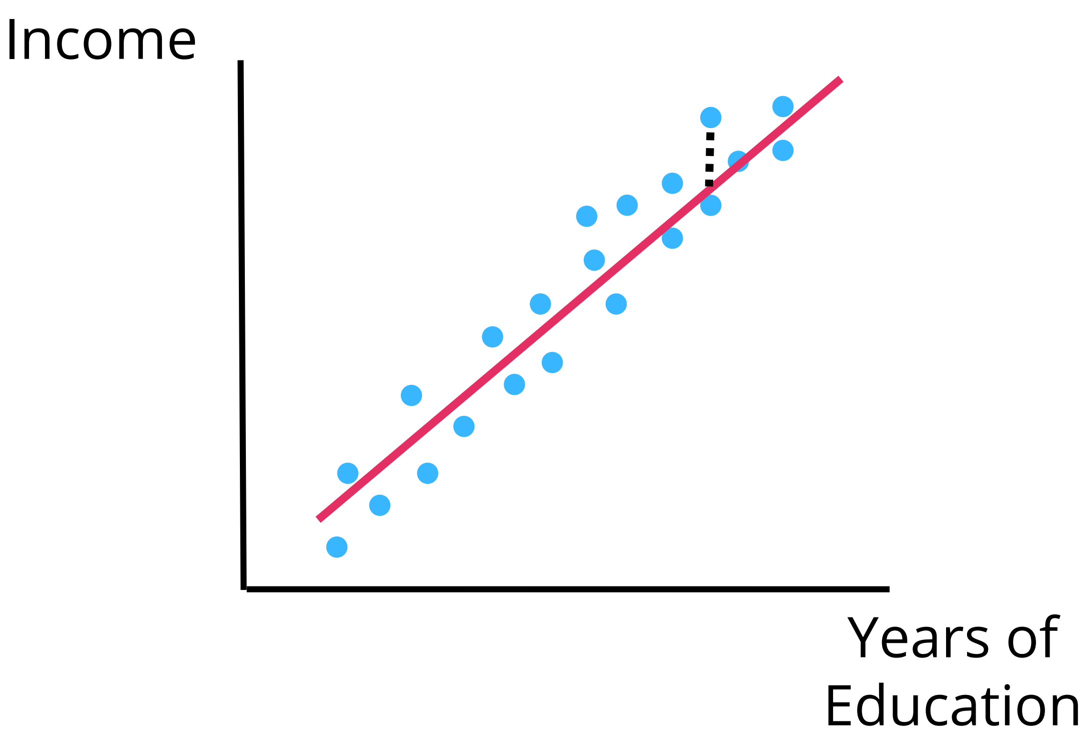
Goodness of Fit

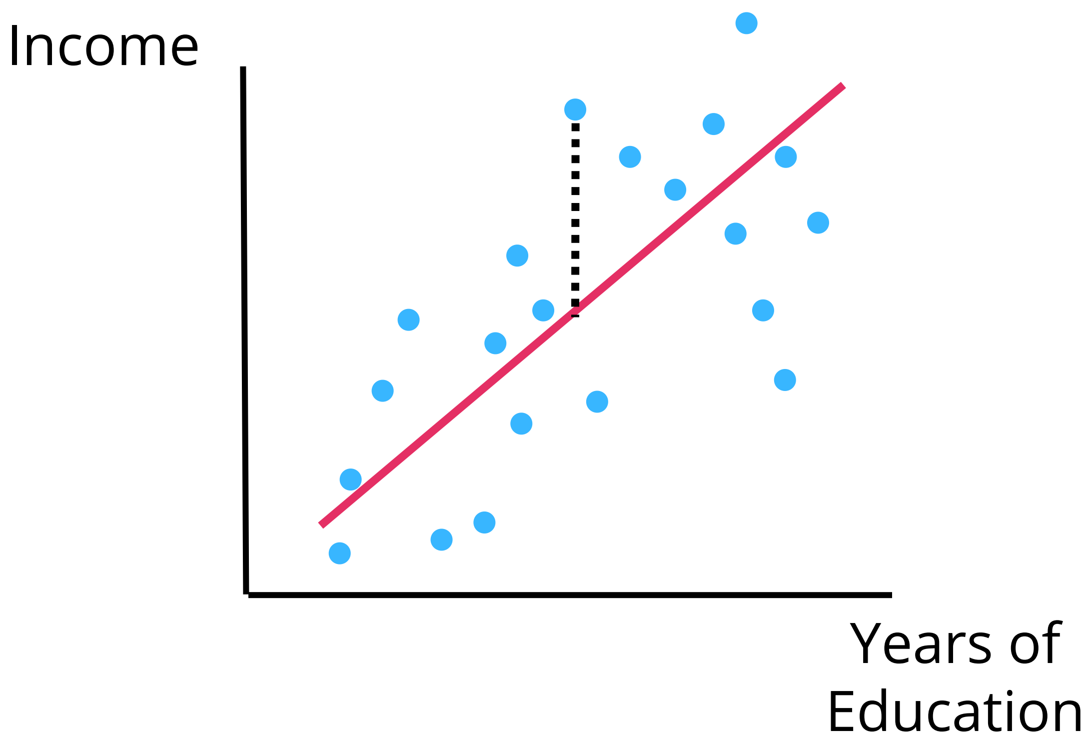
- **R-squared** measures the proportion of the total variation in the dependent variable (TSS) that is explained by our independent variable (ESS)

- It can also be understood as a way of assessing the size of the total mistakes made by our model when fitting the line.

Goodness of Fit

- **R-squared** measures the proportion of the total variation in the dependent variable (TSS) that is explained by our independent variable (ESS)
- It can also be understood as a way of assessing the size of the total mistakes made by our model when fitting the line.

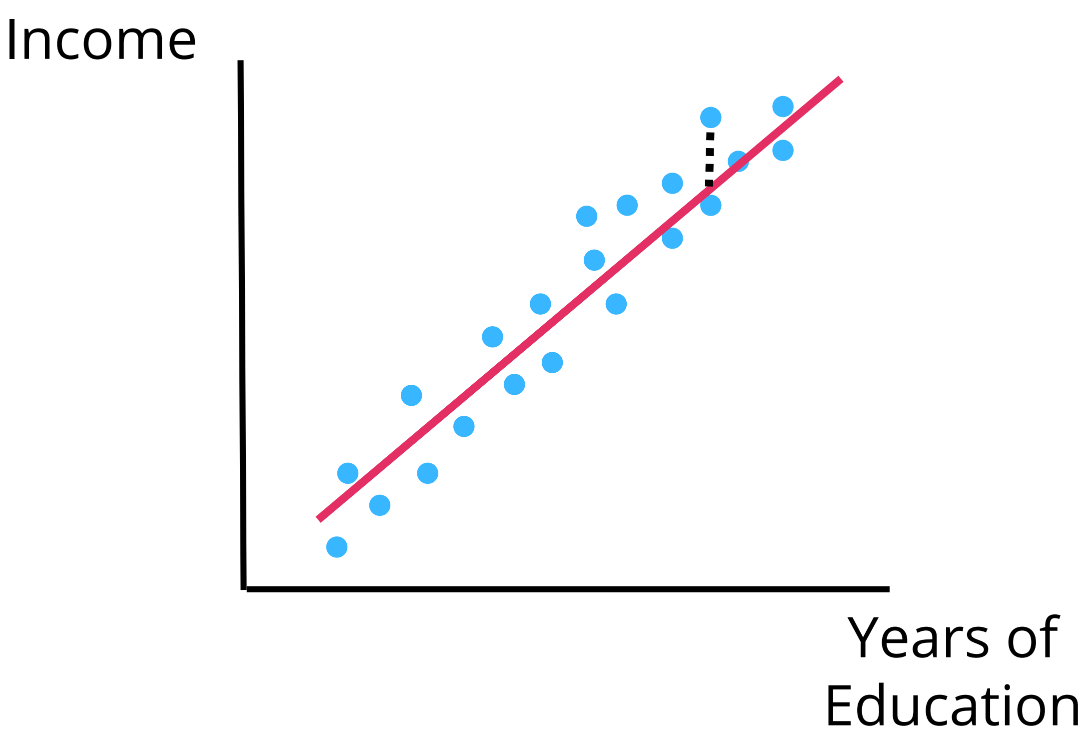




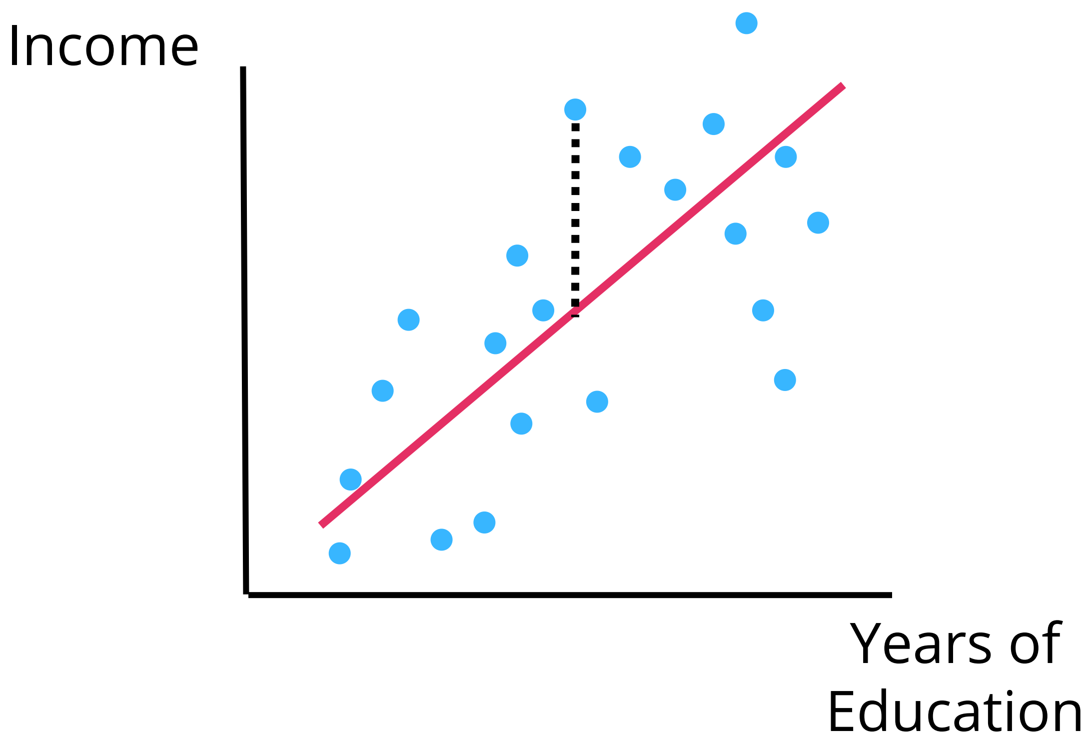
Goodness of Fit

- **R-squared** measures the proportion of the total variation in the dependent variable (TSS) that is explained by our independent variable (ESS)

- It can also be understood as a way of assessing how big are the bulk of mistakes made by my model when fitting the line?



High R-squared



Low R-squared

Confidence Intervals

Confidence Intervals

- We know Statistics is mostly about **estimation (inference)**
  - **Stats** provides us tools to make estimations (very educated guess?)

Confidence Intervals

- We know Statistics is mostly about **estimation (inference)**
  - **Stats** provides us tools to make estimations (very educated guess?)
  - **Confidence Intervals**: are just one additional tool (a statistical concept) used to estimate a range of values within which a population parameter

Confidence Intervals

- We know Statistics is mostly about **estimation (inference)**
  - **Stats** provides us tools to make estimations (very educated guess?)
  - **Confidence Intervals**: are just one additional tool (a statistical concept) used to estimate a range of values within which a population parameter

Without calculating
Confidence Intervals

Point estimate

24.5

"Giving the mean of my sample, my guess is that the average age of this class (population) is 24.5 years old"

Confidence Intervals

- We know Statistics is mostly about **estimation (inference)**
  - **Stats** provides us tools to make estimations (very educated guess?)
  - **Confidence Intervals**: are just one additional tool (a statistical concept) used to estimate a range of values within which a population parameter

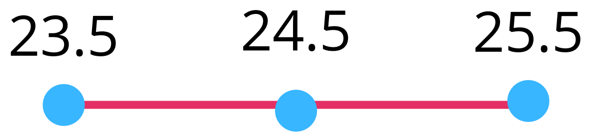
Without calculating
Confidence Intervals

Point estimate

"Giving the mean of my sample, my guess is that the average age of this class (population) is 24.5 years old"

Calculating
Confidence Intervals

Interval



"Giving the mean of my sample, I'm 99% confident that the average age of this class (population) is between 23.5 and 25.5 years old"

The questions then is: **how to calculate the CI?**

Formula

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

But we use the t-test (and not the z)! Why?

When the standard deviation (s) is estimated, then normalizing yields a t-distribution (similar shape to the bell curve)

t-stats and p-values

t-stats and p-values

- ***t*-stats:** a statistical measure we will need when constructing our confidence intervals (or hypothesis testing) when estimating our population parameters.
  - When running **regressions** our t-stats or t-value is calculated by:
    - t = B / SE
      - B = our estimate
      - SE = the standard error of the coefficient

- ***p-value***: is a statistical measure used in hypothesis testing to determine the strength of evidence against a null hypothesis. If the null hypothesis were true, then how likely (probability) would it be for me to obtain a value like this coefficient I've estimated?

t-stats and p-values

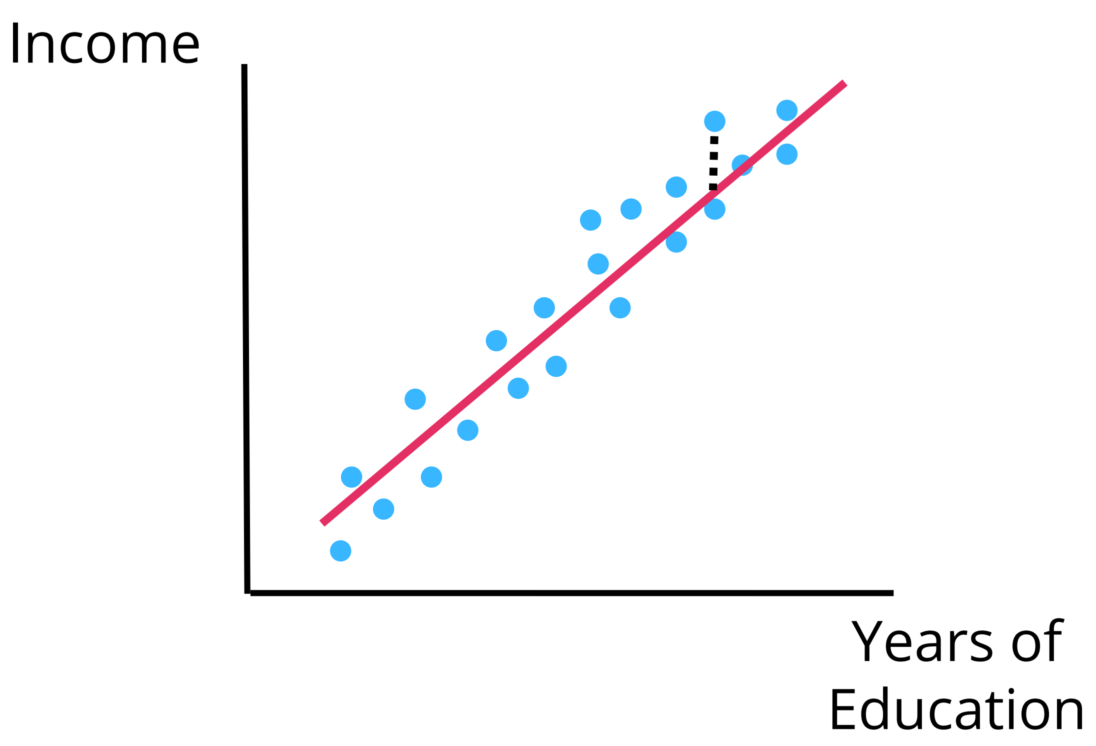
- ***p-value*: is a statistical measure used in hypothesis testing to determine the strength of evidence against a null hypothesis. If the null hypothesis were true, then how likely (probability) would it be for me to obtain an estimation like the one I got?

t-stats and p-values

- ***p-value***: is a statistical measure used in hypothesis testing to determine the strength of evidence against a null hypothesis. If the null hypothesis were true, then how likely (probability) would it be for me to obtain an estimation like the one I got?

Thus, if our sample show this:

Coefficient estimate (slope) = 7,500

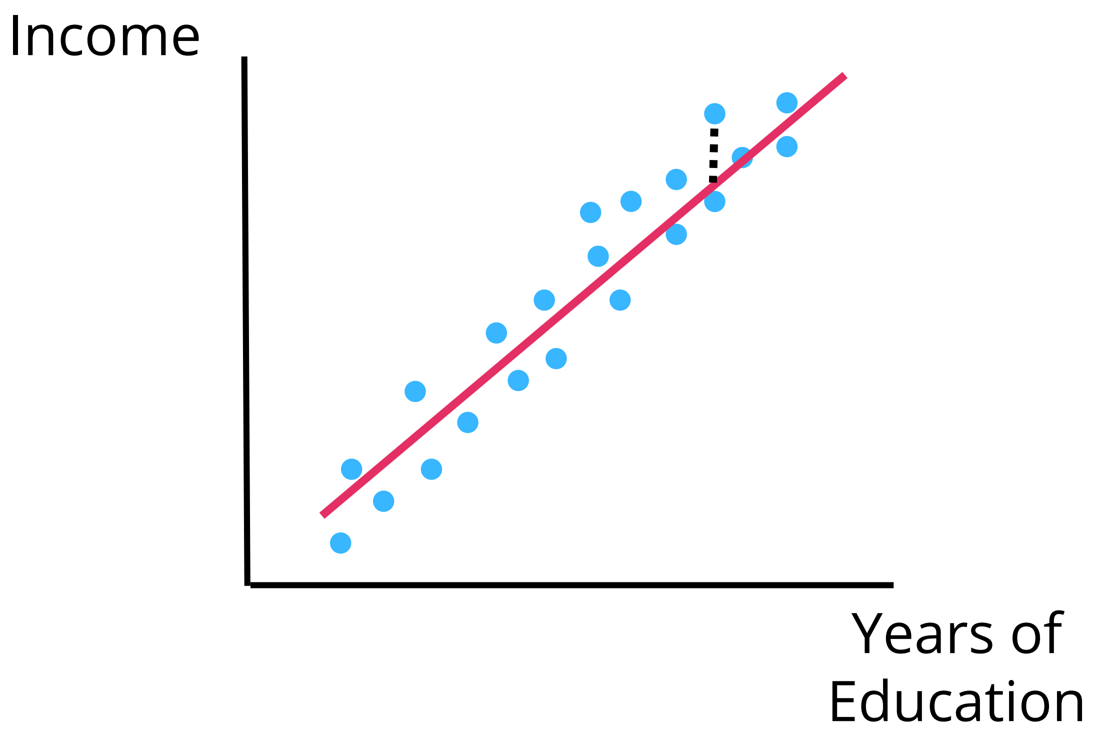


t-stats and p-values

- ***p-value***: is a statistical measure used in hypothesis testing to determine the strength of evidence against a null hypothesis. If the null hypothesis were true, then how likely (probability) would it be for me to obtain an estimation like the one I got?

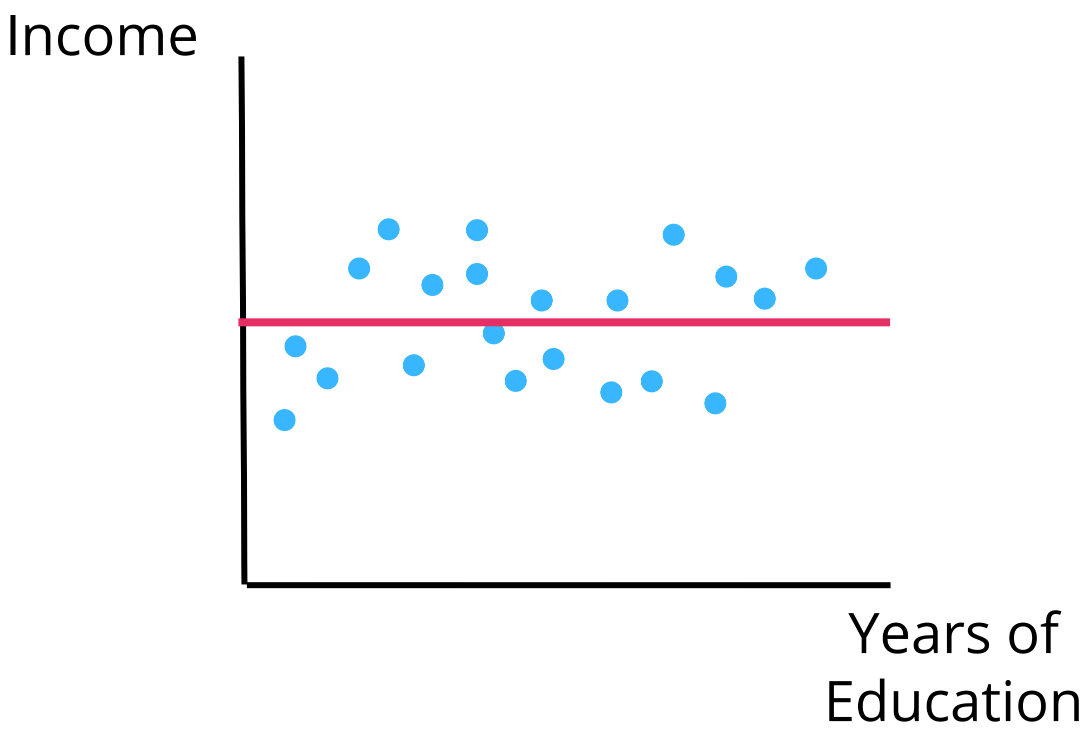
Thus, if our sample show this:

Coefficient estimate (slope) = 7,500



... but the null hypothesis says this

Coefficient estimate (slope) = 0.00



t-stats and p-values

- Then how can we be sure that our coefficient estimate is wrong?
  - **P-value!**
  - It will tell us the probability of getting a t-statistic for the value we estimated or even more extreme than that.

t-stats and p-values

- Then how can we be sure that our coefficient estimate is wrong?
  - **P-value!**
  - It will tell us the probability of getting a t-statistic for the value we estimated or even more extreme than that.

- **Interpreting the p-value**
  - When our p-value is lower than our rule of thumb (=0.05) or .0.01, we say that the probability is so low that we reject the null hypothesis (which is normally formulated in terms of the estimation being actually equal to zero.

Thank you!

Mini-lecture 4

DAC 2023