Regression

Regression

- y= output, such as income, support for welfare
- c= constant, does not change
- b= coefficient (an increase of one year of education increases income by 1000 crowns)
- e=error

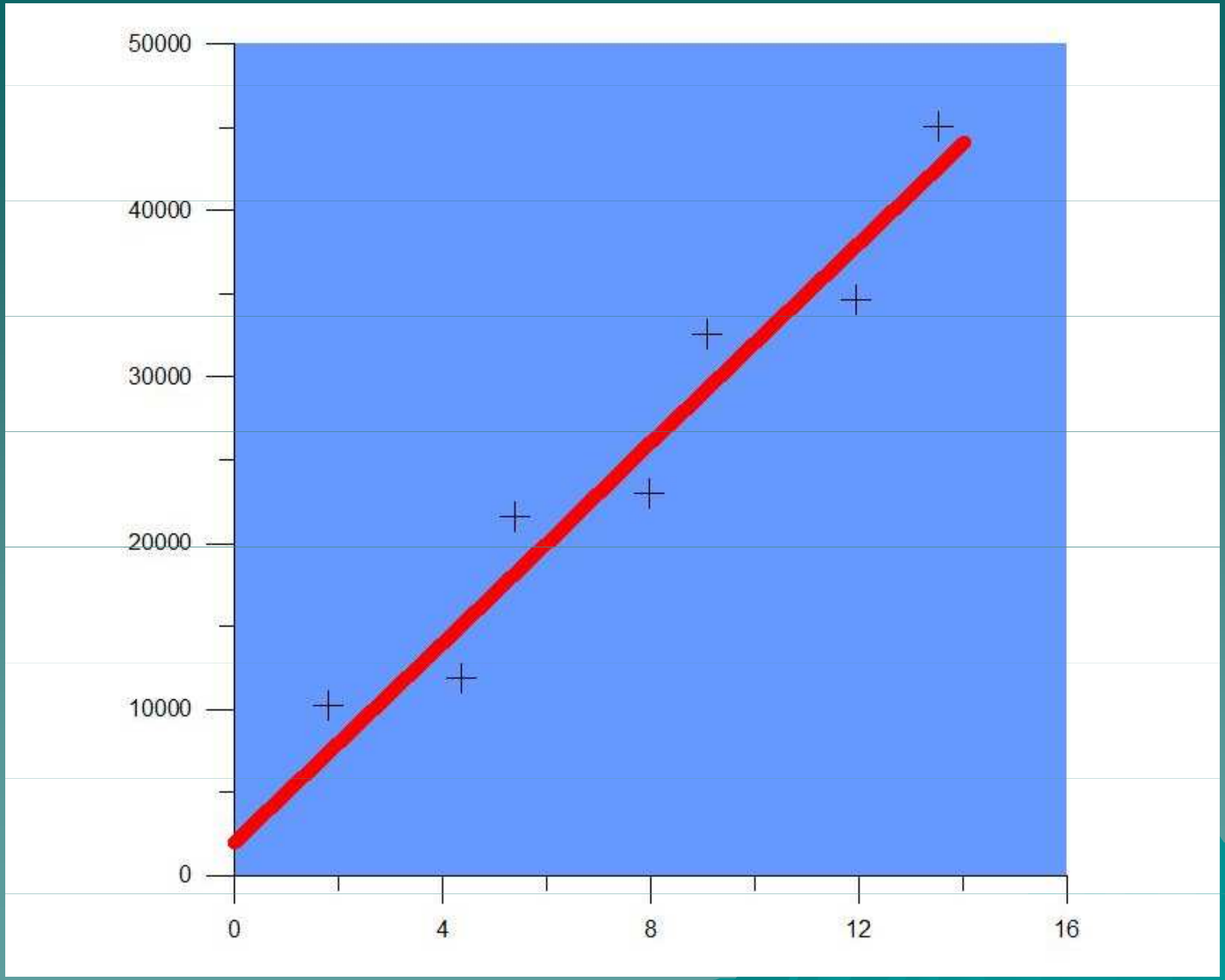
Example

- Y = c +aX + e
- Y= 2000 + 3000X
- Income = 2000$ + 3000 (years of education)
- If somebody studies for 10 years her expected income will be

2000$ +3000$ * 10years = 32,000$

- If she studies for 2 years her expected income will be

2000$ +3000$ * 2years = 8,000$


50000
40000
30000
20000
10000
0
0
4
8
12
16

New Example

- Y = c +aX + e
- Y= 10 + 200X
- y= Live births in Prague that year
- X = Amount of hours they show films with Tom Cruise on TV that year
- How many children will be born if they show 5 of his films and each film is 2 hours?

The Answer

- Y= 10 + 200 (5 x 2)
- = 10 + 200 (10)
- = 10 + 2000
- = 2010

Variables

- Variables = vary
- Both income and education change, so they vary
- Variation = how much they vary
- Explained variation (r-square) = we want to explain how much they vary

Dependent and Independent Variables

- Y (income) is the dependent variable, because its outcome ***depends*** on a different variable
- X (Education) is an independent variable, because it exists independently of income and it influences income
- A clearer case: gender influences income, but income does not influence gender

The constant

- Y= 2000 + 3000X
- This means that even if somebody would not study any years at all he or she would have some income
- Even people without an education can do some jobs like cleaning toilets, emptying garbage
- (or becoming a politician?)
- Even if we do not have a job, we get money from the state, family or friends (or through stealing?)
- Education cannot explain everything, so what is left over comes under the constant

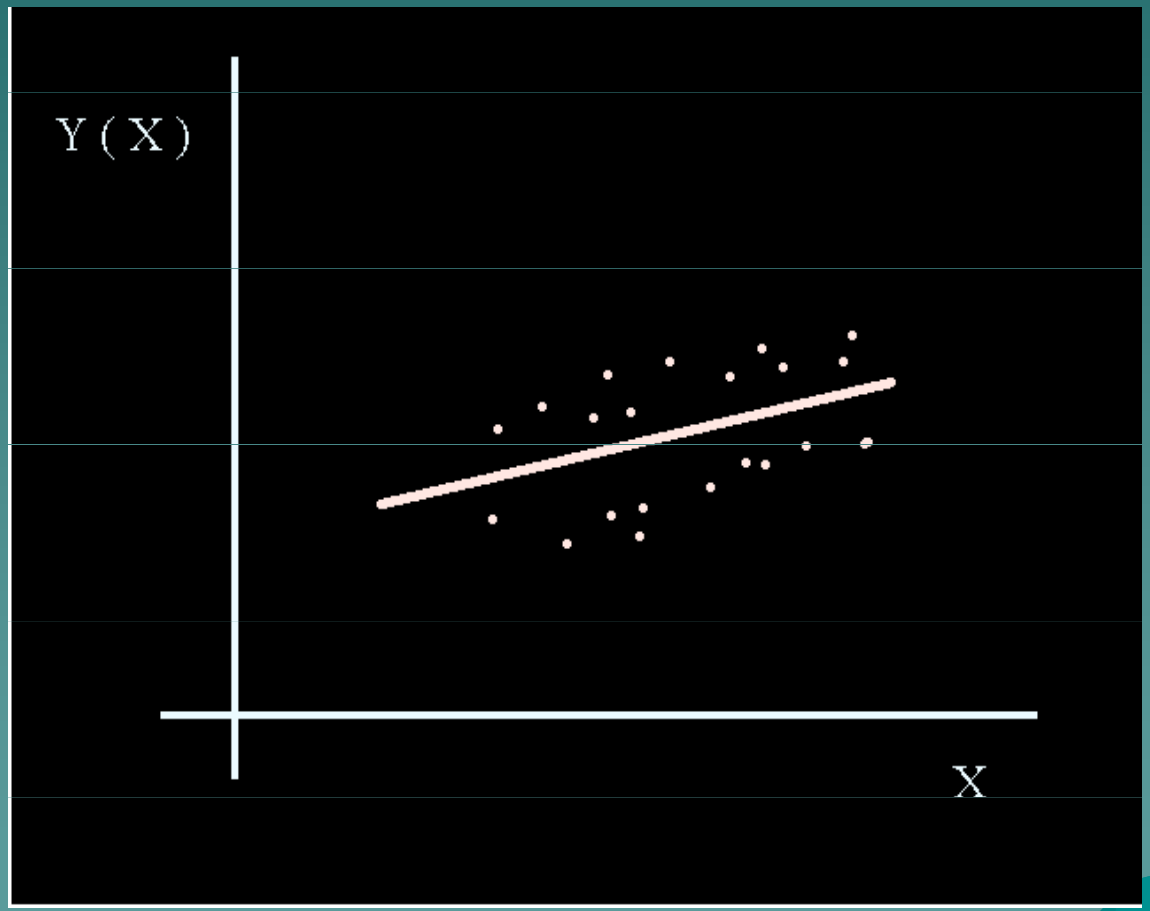
The Errors

- We expect that in each specific case there will be some error
- But on the average the errors will be equal to 0
- That is, in some cases the actual income will be a little bit higher than expected, but in other cases it will be a little bit lower than expected
- So the cases will cancel out each other

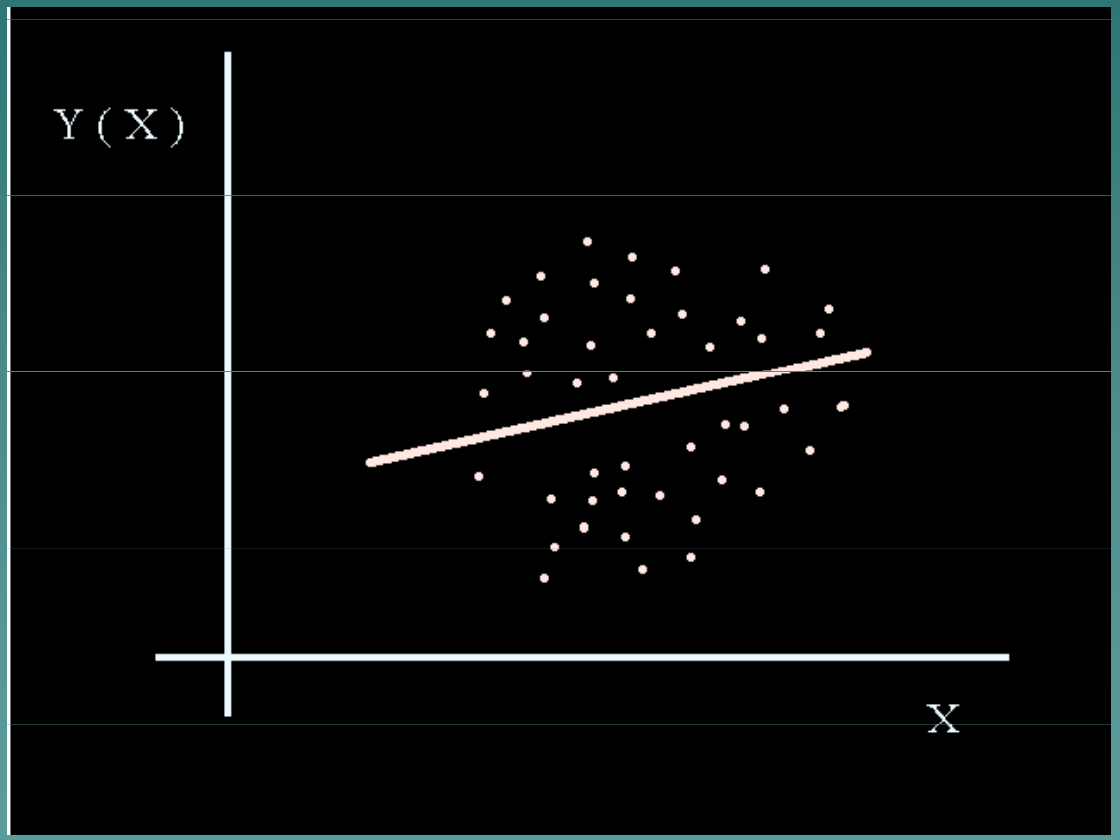
Significance

- p=probability that the relationship is ***NOT*** real
- $p<.05$
- $t>1.96$
- Socially constructed norm
- Would you accept a 10% risk of being wrong if you bought a stock for one crown and expected a profit of one million?
- Would you accept a 1% risk that pressing a button could end the world?

Case 1



Case 2



Factors influencing significance

- Size of the population
- Size of the error

Strength of a variable

- Coefficient
- Standardized coefficient ($0 \leq b \geq 1$)
- Correlation (similar to standardized coefficient, but only used with two variables and no constant)
- r=variance (how much the dependent variable varies)
- r^2=explained variance

More on Coefficients

- The standardized coefficient makes it easier to compare the relative strength of two variables.
- Example: income = 2000 + 1000 *Education
- Or income = 2000 + 5000*gender
- Gender has a higher efficient, but can only be 0 or 1, while education can be 1-25
- So if we can make both coefficients be between 0 and 1 then they are easier to compare

Disadvantages of Standardizing

- It is more difficult to interpret standardized coefficients
- It is clear to say that income increases by 1000 dollars for evey year of education then to say that the standard deviation of income increases by .7 for every standardized increase in education.
- Note: you do not have to know here what a standard deviation is, but it has to do with how much a measure deviates from the expected value

Straight line?

- Normally one should make a plot of the dependent and independent variables to see if it really makes a straight line
- Sometimes it could be curved
- Then we can use a log-function
- But since we are working with attitudes, this is not necessary
- Instead we will either use ordinal regression or scaling, which we will discuss later

An example of why it only makes sense to use this kind of regression if we group questions together to a larger scale. Here I have created a chart for the bivariate regression where LESSREG is the dependent variable and SEX is the indpendent variable