Understanding Statistics

Outline

- First regression, correlations, etc.
- Then scaling

Regression

- y= output, such as income, support for welfare

- c= constant

- b= coefficient (an increase of one year of education increases income by 1000 crowns)

- e=error

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
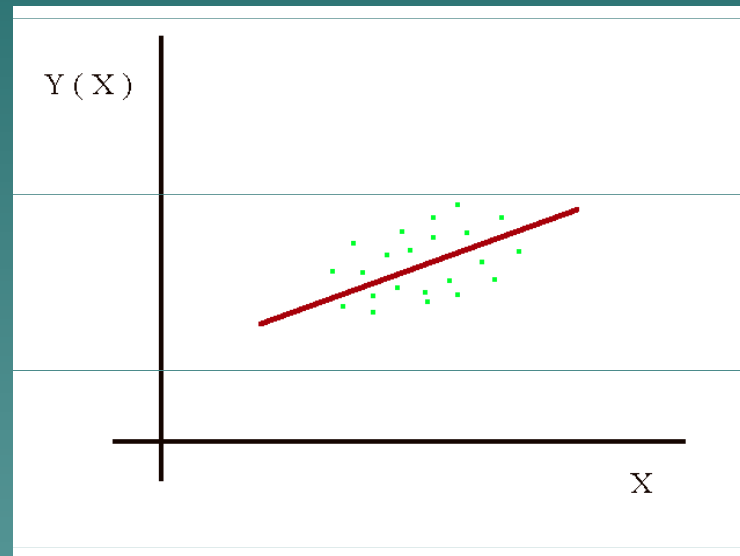
Case 2
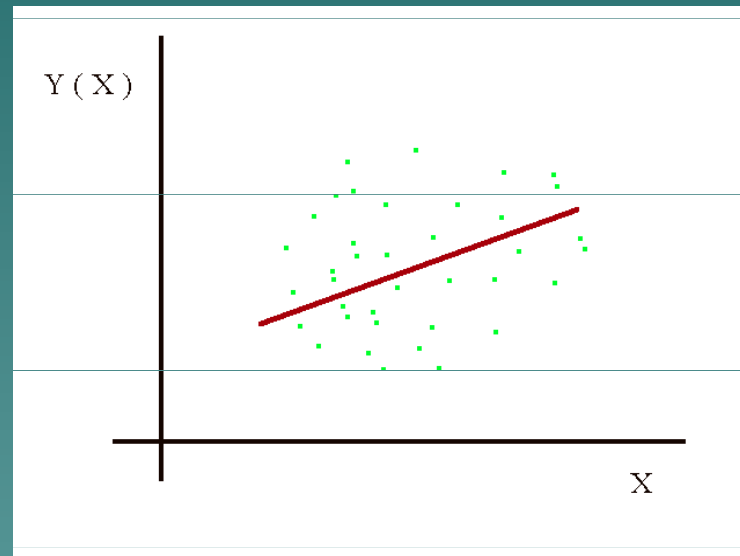
Factors influencing significance

◆ 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

Multiple Regression

- Usually several variables influence the dependent variable
- Example: income is influenced by years of education and gender
- What does it mean if years of education has a coefficient of 1000
- What does it mean if gender has a coefficient of 5000?
- Which variable explains income better?
- Dummy variables = 0 or 1

Control variables

- We want to control whether our variable really is explaining the result or whether some other underlying variable is really at work

- Example: we might want to control if the labor market really discriminates against women who have the same jobs as men or whether the problem is that women choose different types of jobs

- So we can add "working in the private sectors" as a control variable, since women are more likely to work in the private sector

- If gender is still significant then it means that even women, who work in the private sector receive lower salaries than men

- If gender is no longer significant it means that the real problem is that either women choose to work in the public sector (but why do they choose this?)

- or that women cannot get jobs as easily in the private sector (again the question is why?)

Comparing Models

◆ R-square= the amount of variance that a model explains

◆ The higher the r-square the better

◆ But if we have 1,000 variables it will normally explain more variance than if we have one. What is the problem?

◆ We want models to be as "parsimonious" as possible.

◆ Adjusted r-square takes into account also the number of variables

◆ Also we want all the variables to be statistically significant at the 5% level.

Scaling

- Often we cannot find one question that can measure exactly what we want to measure
- But a collection of questions can measure it more clearly
- An example is support for welfare
- We also get a more exact answer if we can create a scale from 1-50 than from 1-4
- We can also use simpler statistical methods if we have a "continuous" dependent variable
- If the dependent variable is 0-1 we need to use things like "logit" and "probit"
- If it is 0-4 we should use things like "ordinal logit" or "ordinal probit"

One Dimensional Scales

◆ Cronbach alfa

◆ Test to see if all the questions are consistent

◆ We might think a group of questions belong together, but the respondents could interpret them differently

◆ Cronbach's alfa expects all the questions to measure approximately the same thing

Mokken Scale

- Is not used much, but it is also valuable
- It allows us to test the consistency for questions that we can rank
- For example: do you support the women's right to abortion anytime she wants it?
- Do you support the right for women to have an abortion if they are too poor to take care of the child?
- Do you support the right for women to have an abortion if they were raped?
- Do you support the right for women to have an abortion if their own life is threatened?

Multidimensional scaling

- Factor analysis
- There can be several dimensions to an issue
- For example: "support for big government"
- One dimension could be support for welfare programs
- Another dimension could be support for the military and police
- A third dimension could be support for education
- We use statistical programs to see which questions scale well together.
- The most common is called "principle components analysis"