Chapter 10: Regression

Context

- In chapter 8 we looked at summarizing the relationship between two variables, using:
  - average of x, SD of x
  - average of y, SD of y
  - correlation coefficient r
- We also looked at the SD-line: the line that goes through the middle of the football
- Now we will discuss another line: the regression line.
- The regression line estimates the average value of y for each value of x. Hence, it can be used for predicting y from x.
- Other new terminology: regression estimate, graph of averages, extrapolation, regression effect, regression fallacy

2 / 21

Regression line

3 / 21

Graph of averages

- Goal: we want to describe how one variable depends on the other.
- More precisely: we want to estimate the average value of y for a given value of x.
- Example: what is the average number of children per woman for countries with a contraceptive use of 15% (between 10% and 20%)?
- Example: what is the average number of children per woman for countries with a contraceptive use of about 75% (between 70% and 80%)?

4 / 21

Regression estimate

- We can make such estimates in a more systematic way. See overhead.
- With an increase of 1 SD in x, we do not expect that y increases by a full SD. Instead, y is expected to increase by only r SDs.See overhead.
- This is the regression method and results in a regression estimate

5 / 21

Example

- HANES study: height and weight of 988 men age 18-24
  - Height: average = 70 inches, SD = 3 inches
  - Weight: average = 162 pounds, SD = 30 pounds
  - Correlation coefficient $r = 0.47$
- Estimate the average weight of men that are 73 inches tall
- Regression method: (see overhead)
  - Step 1: Draw picture
  - Step 2: Convert x to standard units z_x: how many SDs are these men above average in height?
  - Step 3: Compute $z_y = z_x \times r$: average weight in standard units
  - Step 4: Convert z_y back to original units: average weight in pounds

Regression line

- All the regression estimates fall on one line
- This line is called the regression line
- For an increase of 1 SD in x, there is an increase of only r SDs in y. See overhead.
- Note that the regression line is always less steep than the SD line, because $-1 \leq r \leq 1$.
- Both the SD line and the regression line go through the point of averages.

Graph of averages

- The graph of averages shows the average y value for each value of x. See overhead.
- If the graph of averages follows a straight line, it is the same as the regression line.
- If the graph of averages is close to a straight line, then the regression line is a smoothed version of the graph of averages.
- Warning: If the graph of averages doesn't look like a straight line at all, then don't use the regression line! See overhead.

Regression method for individuals

Example

- Math SAT scores and first year GPAs of students:
  - SAT score: average $= 550$, SD $= 80$
  - first year GPA: average $= 2.6$, SD $= 0.6$
  - $r = 0.4$
  - The scatter diagram is football shaped
- A student is chosen at random. Predict his/her first year GPA. Our best guess is the average GPA: 2.6.
- A student is chosen at random and has SAT score 650. Predict her/his first year GPA. Our best guess is the average GPA for students with an SAT score of 650. We can find this new average using the regression method. See overhead.

10 / 21

Method

- If you know nothing, the average is the best guess
- If you know the x-value, the best guess for the y-value is the average of the y-values for this x. You can find this new average using the regression method.

11 / 21

Extrapolation and generalization

- In SAT example, the university has only experience with the students it has admitted. Does the regression estimate also work for students who were not admitted?
  Not necessarily, but it is often used like that.
- Be careful with generalizations:
  - Sample: Does the sample represent the population to which you want to generalize?
  - Range: The SAT scores in the sample ranged from 200 to 800. The regression method may work poorly for a student with SAT score 50.

12 / 21

Percentile (ranks)

- Example: SAT scores and first year GPA:
  - SAT score: average = 550, SD = 80
  - first year GPA: average = 2.6, SD = 0.6
  - $r = 0.4$
  - The scatter diagram is football shaped
- A student is chosen at random, and is at the 90th percentile of the SAT scores. Predict his/her percentile rank on the first-year GPA.
- Method (see overhead)
  - Step 1: Draw a picture
  - Step 2: Find z_x (use normal table)
  - Step 3: Compute $z_y = z_x \times r$
  - Step 4: Convert z_y to percentile rank (use normal table)

Percentile (ranks)

- Note that we did not use information about average and SD!
- We only used the normal table and r. That is because the whole problem is worked in standard units.
- Why can we use the normal table? Because the scatter diagram is football shaped. See overhead.

Regression to the mean

- Note that student who was at 90th percentile for SAT scores, was at the 69th percentile for first-year GPA scores.
- So student was well above average for SAT, and still predicted to be above average for GPA, but less so.
- Why? Why don't we predict the student to be at the 90th percentile?
- This is because the scores are not perfectly correlated.
  - If $r = 1$, then we would predict the percentile ranks to the same.
  - If $r = 0$, then the SAT score does not help us in estimating the GPA. So we would predict the percentile rank to be 50%.
  - If r is between 0 and 1, we predict something in between, and the regression method tells us precisely what.

Regression fallacy

Example

- Preschool program for boosting children's IQs
  - Children are tested when they enter (pre-test)
  - Children are tested when they leave (post-test)
  - Results:
    - Pre-test: average = 100, SD = 15
    - Post-test: average = 100, SD = 15
- So it seems the program didn't have much effect.
- A closer look at the data showed:
  - Children who were below average on the pre-test had an average gain of 5 IQ points
  - Children who were above average on the pre-test had an average loss of about 5 IQ points

What is going on?

- It seems that the program equalizes intelligence
  - Perhaps the brighter kids play with the dull kids, and the difference between the two groups diminishes?
- Reality: not much is going on:
  - The children cannot be expected to test exactly the same on both tests
  - These differences make the scatter diagram for the test scores spread around the SD line, in a football shape
  - The spread around the line makes that the bottom group comes up and the top group comes down
  - See overhead

Regression effect

- Regression effect: in almost all test-retest situations
  - the bottom group on the first test will on average show some improvement on the second test
  - the top group on the first test will do a bit worse on the second test
- Regression fallacy: thinking that the regression effect must be due to something important, not just spread around the SD line

Another explanation

- Example: repeated IQ tests
- Basic fact: scores will differ a bit, due to chance error
- If someone did very well on the first test, that suggests that he/she was lucky, and will probably score a bit lower on the second test
- If someone did very badly on the first test, then that suggest he/she had a bad day, and will probably do a bit better on the second test

20 / 21

Two regression lines

- Two regression lines: y on x, and x on y
- See overhead
- Example: IQ scores:
  - Women: average = 100, SD = 10
  - Men: average = 100, SD = 10
  - Correlation between IQ of husbands and wives is about 0.5
- Large study found that men with IQ of 140 had wives whose average IQ was 120
- They also found that women with IQ of 120 had men whose average IQ was 110
- What is going on? See overhead. We talk about different groups of people.

21 / 21