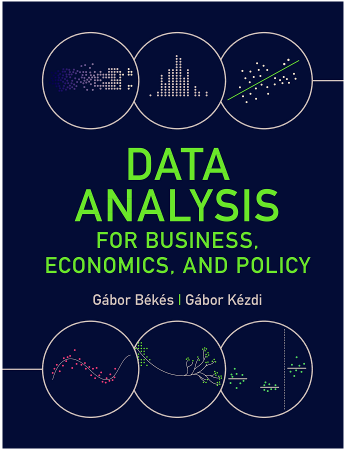
04 Comparison and correlation

Gábor Békés

Data Analysis 1: Exploration

2020

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- **gabors-data-analysis.com**
  - Download all data and code `gabors-data-analysis.com/data-and-code/`
- This slideshow is for **Chapter 04**

Motivation

Are larger companies better managed? Answering this question may help in benchmarking management practices in a specific company, assessing the value of a company, or estimating the potential benefits of a merger between two companies.

To answer this question you downloaded data from the World Management Survey. How should you use the data to measure firm size and the quality of management? How should you assess whether larger companies are better managed?

Topics for today: Comparison and correlation

Topics for today

The y and the x

- Much of data analysis is built on comparing values of a y variable by values of an x variable, or more x variables.
- Uncover the **patterns of association**: whether and how observations with particular values of one variable (x) tend have particular values of the other variable (y).
- The role of y is different from the role of x.
  - it's the values of y we are interested in
  - compare observations that are different in their x values.
- It is our decision to pick y

The y and the x

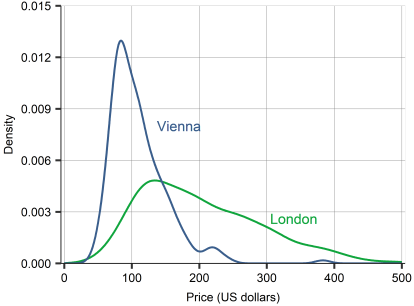
- This asymmetry comes from the goal of our analysis.
- Goal 1: predicting the value of a y variable with the help of other variables - many x variables, such as x_1, x_2, ...
- The prediction itself takes place when we know the values of those other variables but not the y variable.
- Goal 2: learn about the effect of a causal variable x on an outcome variable y.
- What the value of y would be if we could change x

Comparison and conditioning

- We compare y, by values of x –> we **condition** y on x or y *given* x
  - x (by the values of which we make comparisons) –> is the conditioning variable.
  - y –> outcome variable.
- Compare prices of hotels (y) with different cities (x) –>
  - price of hotel is the outcome
  - type of city is the conditioning variable.

Comparisons and conditional distributions

- The conditional distribution of a variable is the distribution of the outcome variable given the conditioning variable.
- Straightforward concept if the conditioning variable is qualitative (simple if binary)
- Comparing histograms



Conditional statistic

- Conditional mean = mean of a variable for each value of the conditioning variable.
- The conditional expectation of variable y for different values of variable x is

$$E[y|x]$$

- This is a function: for a value of x, the conditional expectation gives number that is the expected value (mean, average) of variable y for observations that have that x value
- It gives different values based on the conditioning variable x

Conditional and joint distributions of two quantitative variables

- Two variables, many values
- The joint distribution of two variables shows the probabilities (frequencies) of each value combination of the two variables.
- A scatterplot is a two-dimensional graph with the values of each of the two variables measured on its two axes, and dots entered for each observation in the dataset with the combination of the values of the two variables.
- Works when dataset relatively small.
- For larger samples, we can bin values, and use "bin scatter" – Bin scatter shows conditional means for bins we created

Case Study - Management quality and firm size

- Management quality and firm size: describing patterns of association
- Whether, and to what extent, larger firms are better managed.
- Answering this question can help understand why some firms are better managed than others.
- Data from the World Management Survey to investigate our question.

Case Study - Management quality and firm size

- Interviews by CEO/senior managers, based on that a score is given
- Management quality is measured as management score.
- Each score is an assessment by the survey interviewers of management practices in a particular domain
  - tracking and reviewing performance or
  - time horizon and breadth of targets, etc
- Measured on a scale of 1 (worst practice) to 5 (best practice).

Case Study - Management quality and firm size

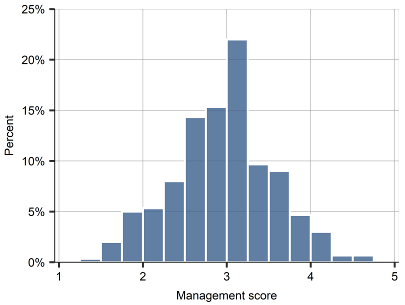
- Take 18 individual measures and average
- Measure of the quality of management is the simple average of these 18 scores = "the" management score.
- By construction, the range of the management score is between 1 and 5.

Case Study - Management quality and firm size

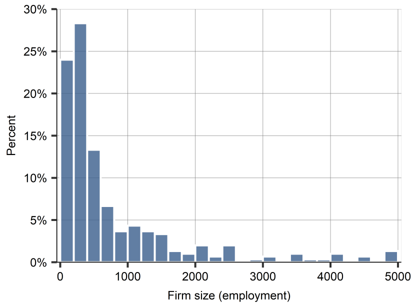
- Data from the World Management Survey to investigate our question.
- In this case study we analyze a cross-section of **Mexican** firms from the 2013 wave of the survey.
- Only firms with 100 – 5000 employees, N=300
- The y = measure of the quality of management. The x = measure of firm size.
- Firm size = number of employees

Case Study - Management quality and firm size

(a) Management score



(b) Firm size (number of employees)



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

- Management score: The mean is 2.9, the median is 3, and the standard deviation is 0.7.

- Firm size: The range of employment is 100 to 5000. The mean is 760 and the median is 350, skewness with a long right tail. Some large firms, but not extreme, kept as is.

Case Study - Management quality and firm size

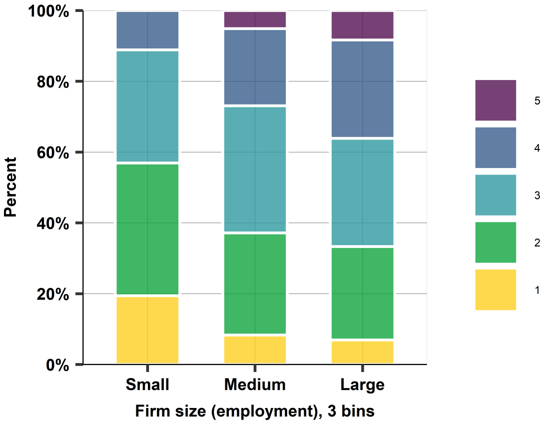
Conditional probabilities in data.

- Three bins of firm size. By number of employees: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Take a single measure: Lean management score, with values 1,2,3,4,5.
- Thus, for each score variable we have 15 conditional probabilities: the probability of each of the 5 values of y by each of the three values of x – e.g., $P(y = 1|x = small)$.

Case Study - Management quality and firm size

- Lean management score 1–5
- Firm size: small, medium, large
- Conditional probability:
  - share of score=1 conditional on being a small firm is about 20%.
  - share of score=5 conditional on being a large firm is about 10%.
- Shows a pattern of association

Note: Source: Management quality is an average score of 18 variables. Firm size is number of emoployees. `wms-management-survey` *data. Mexican sample, n=300.*



Case Study - Management quality and firm size

Conditional statistic - conditional mean.

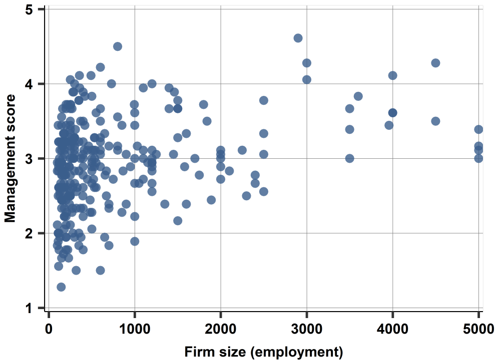
- Can calculate the mean given firm size.
- Three bins of employment: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.18 for large.
- First simple evidence: larger firms have better management.

Case Study - Management quality and firm size

- Conditional mean and joint distribution
- How our management quality variable
  - y: the management score

  is related to our firm size variable
  - x: employment
- Scatterplot
- Bin-scatter

Case Study - Management quality and firm size



- Scatterplot
- Both x and y axis qualitative
- Each dot is an observation
- Full information on association

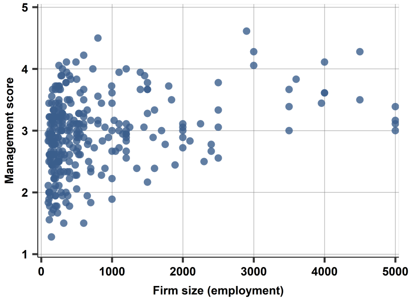
Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

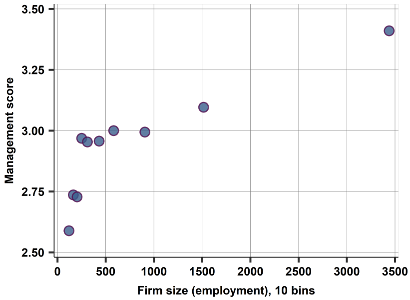
- Bin-scatter: calculate the mean of y conditional on ten bins of x.
- Bin-scatter: cut x's distribution into 10 parts, with equal number of firms. (remember - percentiles)
- Show average management score as a point corresponding to the midpoint in the employment bin (e.g., 110 for the 100–120 bin).
- Dots NOT equally spread out - more frequent where more observations!

Case Study - Management quality and firm size

(a) Scatterplot



(b) 10 Bin-scatter



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

- Some positive association is shown, but not easy to read
- Bin-scatter - positive overall, but most for small vs medium.
- Difference in mean management quality tends to be smaller when comparing bins of larger size

Dependence and independence

- Dependence of two variables - y and x means that the conditional distributions of y - conditional on x - are not the same (x is the conditioning variable).
- Independence of y and x means the opposite: the distribution of y on x is the same, regardless of the value of x.
- Dependence of y and x, may take many forms.

Mean dependence

- Mean-dependence: conditional expectation $E[y|x]$ varies with the value of x.
- Mean-dependence is the extent to which conditional expectations (means) differ.
- Two variables are positively mean-dependent if the average of one variable tends to be larger when the value of the other variable is larger, too.
- Covariance and Correlation Coefficient are measures of mean dependence.

Covariance

The formula for the covariance between two variables x and y both observed in a data table with n observations is:

$$Cov[x,y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{1}$$

- for each observation $i = 1 \dots n$
- The product within the sum in the numerator multiplies the deviation of x from its mean $(x_i - \bar{x})$ with the deviation of y from its mean $(y_i - \bar{y})$
- The entire formula is the average of these products across all observations.

The correlation coefficient

$$Corr[x, y] = \frac{Cov[x, y]}{Std[x]Std[y]} \tag{2}$$

$$-1 \leq Corr[x, y] \leq 1 \tag{3}$$

- The correlation coefficient is the standardized version of the covariance.
- The covariance may be any positive or negative number, while the correlation coefficient is bound to be between negative one and positive one.

Dependence, correlation

- Covariance or the correlation coefficient allow for all kinds of variables, including binary variables and ordered qualitative variables as well as quantitative variables.
- However, they are more appropriate measures for quantitative variables. That's because the differences $y_i - \bar{y}$ and $x_i - \bar{x}$ make more sense when y and x are quantitative variables.

Case Study - Management quality and firm size

- The covariance between firm size and the management score is 177.
- The standard deviation of firm size is 977, the standard deviation of management score is 0.6.
- Positive mean-dependence: firm size tends to be higher at firms with better management.
- the correlation coefficient is 0.30 $(177/(977 * 0.6))$.
- This suggests a positive and moderately strong association.
- Management quality–firm size correlation varies considerably across industries?

Case Study - Management quality and firm size

Table: Measures of management quality and their correlation with size by industry

Industry	Management–firm size correlation	Observations
Auto	0.50	26
Chemicals	0.05	69
Electronics	0.33	24
Food, drinks, tobacco	0.05	34
Materials, metals	0.32	50
Textile, apparel	0.29	43
Wood, furniture, paper	0.28	29
Other	0.44	25
All	0.30	300

Note: *Employee retention rates: The probability of staying with the firm, in the two experimental groups. Source:* working-from-home *dataset*

Measuring a latent concept with many observed variables

- Often a concept is hard, even impossible, to measure.
- Latent variables - while we can think of them as a variable there is no single observed variable to measure them.
- Quality of management at a firm
- IQ

- The problem here is how to combine multiple observed variables

Condensing information: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless

Condensing information: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless
- Need bring it to common scale - standardization: subtracting the mean and dividing with the standard deviation
- The result is a series of variables with zero mean and standard deviation of one
- This standardized measure is called a **"z-score"** or "score"

Comparison and variation in x

- What is the "source of variation" in the conditioning variable
- Or put it differently, why values of the conditioning variable may differ across observations.
- Option 1: experimental data - perfect control
- Option 2: observational data - no perfect control

Comparison in Experimental data

- We have an intervention or treatment.
- Value of the conditioning variable differs across observations because the person running the experiment made them different. Hence the name: 'treatment variable'.
- There is controlled variation - a rule deciding treatment
- Experiment - comparing one or more outcome variables across the various values of a treatment variable
- Example: drug trial

Comparison with observational data

- Most data used in business, economics and policy analysis are observational.
- In observational data, no variable is fully controlled.
- Typical variables in such data are the results of the decisions
- The source of variation in these variables may have multiple sources
- People's choices, decisions, interactions, expectations, etc.
- Compare the value of the outcome variable for different values of the conditioning variable.
- Much harder interpretation

Summary

- For qualitative variables, correlation can be shown by summarizing conditional probabilities (frequencies).
- For quantitative variables, scatterplots offer a visual insight to the pattern of the relationship.
- The correlation coefficient captures a simple measure of mean dependence.
- In some cases, we measure a phenomenon with many variables. In such cases a standardized summary variable (the score) could be used to capture the essence.