Research methodology and effective writing

Lecture III - **Models and data**

Matej Lorko
matej.lorko@euba.sk
www.lorko.sk

Suggested reading:

- Dudenhefer, P. (2009). A guide to writing in Economics. EcoTeach Center and Department of Economics, Duke University.
- Neugeboren, R. H., & Jacobson, M. (2005). Writing Economics. Harvard University.
- Johnson, J. B., Reynolds, H. T., & Mycoff, J. D. (2015). Political science research methods. Cq Press.
- Friedman, S., Friedman, D., & Sunder, S. (1994). Experimental methods: A primer for economists. Cambridge University Press.

Economic models

- Economists build models the way curious scientists do: Reduce the phenomenon to its basic elements and recombine these elements so as to produce a model that resembles the original in relevant respects. Take it apart, figure out how it works, then put it back together and see if it goes.

- Economic models specify relationships between two kinds of variables: exogenous variables and endogenous variables

- Exogenous variables are inputs to the model, factors that influence what happens but are themselves determined "outside" the model. They are givens, fixed values that are assumed not to change over the period of analysis. Endogenous variables are outputs of the model, determined "within." Usually, a mathematical function is used to represent the relationship between exogenous and endogenous variables

- Applying basic models allows one to make predictions about the real world economy. By building and using models, economists are able to focus on simple, sometimes subtle, relationships in the data and explain the causal links at work. Finding the pattern in the data allows one to say something about how the economy works.

- The fit between a model and reality is never perfect. When the fit is good, we can make better predictions about the future and better understand the past. In the former case, the passage of time will fail to disconfirm the prediction; in the latter case, historical research will match our expectations.

- As in any science, our theories can really only be disproved. However, when our predictions are correct, the weight we place on our models increases. When our predictions are wrong, we are left either looking for more data or perhaps a new or revised model. That model may be used to generate new predictions, which can then be confronted with new data, which may again bring disconfirmation of the prediction and suggest a revision of the current model.

Models and data

- In presenting your hypothesis, you need to discuss the data set you are using and, in most cases, the model (usually some type of regression) you will run. You should say where you found the data, and use a table, graph, or simple statistics to summarize them.

- You should explain how the data relate to your hypothesis and note any problems they pose. If you have only a small set of observations, or have to use proxies for data you cannot directly observe, you should explicitly acknowledge this.

- In your thesis, it may not be possible to reach conclusive empirical results. You may have incomplete data, or your regression coefficients may not be significant, or you may not have controlled for significantly all the factors involved. It is better to acknowledge these shortcomings than to make overly broad and unsupported statements.

Measurement

- Before testing hypotheses, we must understand some issues involving the measurement of the concepts we have decided to investigate and how we record systematic observations using numerals or scores to create variables that represent the concepts for analysis.

- How researchers measure their concepts can have a significant impact on their findings; differences in measurement can lead to totally different conclusions.

- It is useful to think of arriving at the definition of the variables as being the last stage in the process of defining a concept precisely. We often begin with an abstract concept (such as democracy), then attempt to define it in a meaningful way, and finally decide in specific terms how we are going to measure it.

- At the end of this process, we hope to attain a definition that is sensible, close to our meaning of the concept, and exact in what it tells us about how to go about measuring the concept.

- To be useful in providing scientific explanations for behavior, measurements of phenomena must correspond closely to the original meaning of a researcher's concepts.

- They must also provide the researcher with enough information to make valuable comparisons and contrasts. Hence, the quality of measurements is judged in regard to both their accuracy and their precision.

Accuracy of measurement

- There are two major threats to the accuracy of measurements. Measures may be inaccurate because they are unreliable and/or because they are invalid.

- Reliability describes the consistency of results from a procedure or measure in repeated tests or trials. In the context of measurement, a reliable measure is one that produces the same result each time the measure is used. An unreliable measure is one that produces inconsistent results-sometimes higher, sometimes lower.

- The reliability of social science measures can be calculated in many different ways.

    - The test-retest method involves applying the same "test" to the same observations after a period of time and then comparing the results of the different measurements.

    - The alternative-form method of measuring reliability also involves measuring the same attribute more than once, but it uses two different measures of the same concept rather than the same measure.

    - The split-halves method of measuring reliability involves applying two measures of the same concept at the same time. The results of the two measures are then compared. This method avoids the problem that the concept being measured may change between measures.

- A valid measure is one that measures what it is supposed to measure. Unlike reliability, which depends on whether repeated applications of the same or equivalent measures yield the same result, validity refers to the degree of correspondence between the measure and the concept it is thought to measure.

Sampling

- Suppose we want to assess national level of support for some proposed government policy. Since it is impossible to interview everyone, a more practical approach is to select just a "few" members of the population for further investigation. This is where sampling comes in.

- A sample is any subset of units collected in some manner from a population. The sample size and how its members are chosen determine the quality (that is, the accuracy and reliability) of inferences about the whole population.

- A researcher's decision whether to collect data for a population or for a sample is usually made on practical grounds. The advantages of taking a sample are often savings in time and money. The disadvantage is that information based on a sample is usually less accurate or more subject to error than is information collected from a population.

- Once a sample has been gathered, features or characteristics of interest can be examined and measured. The attributes of most interest in empirical research are numerical or quantitative indicators such as percentages or averages. These measures, or sample statistics, as they are known - are used to approximate the corresponding population values, or parameters.

- In order to mitigate the sample bias, ideally each element in the total population should have a known probability of being included in the sample. This knowledge allows a researcher to calculate how accurately the sample reflects the population from which it is drawn.

What can be learned from samples

- Samples provide only estimates or approximations of population attributes. Occasionally these estimates may be exactly right, but most of the time, however, they will differ from the true value of the population parameter.

- When we report a sample statistic, we always assume there will be a margin of error, or a difference between the reported and actual values.

- Where does the loss of precision or accuracy come from? The answer is chance, or luck of the draw. If you flip a coin ten times, you probably won't get exactly five heads, even if the coin is fair or the probability of heads is one-half. Randomness seems to be an innate feature of nature, at least on the scale at which we observe it.

- Just as with our coin toss, a random sample of ten (or even much larger) is not likely to produce precisely the value of a corresponding population parameter. But if we follow proper procedures and certain assumptions have been met (for example, the sample is a simple random sample from an infinite population), a sample statistic approximates the numerical value of a population parameter.

Types of Data and Collection Techniques

- Researcher collects data on behavior by observing either the behavior itself (direct observation) or some physical trace of the behavior (indirect observation).

- Data collected through firsthand observation is an example of primary data, that is, data recorded and used by the researcher making the observations.

- Data from interviews or the written record can be primary data or secondary data - data used by a researcher who did not personally collect the data.

- Students will often find suitable data generated through interviews or the written record for free in publicly available data archives, but students wishing to use data generated through direct or indirect observation must usually rely on their own ability to make the observations.

- Structured and Unstructured Observation - In structured observation, the investigator looks for and systematically records the incidence of specific behaviors. The researcher will have decided, based on theory, the relevant behaviors before starting data collection.

- In unstructured observation, all behavior is considered relevant, at least at first, and recorded. Only later, upon reflection, will the investigator distinguish between important and trivial behavior.

Survey Research and Interviewing

- How to ensure validity and reliability of survey and interview data?
- Let R stand for the respondent and I for the interviewer:
    - The requested information must be available to R (that is, not forgotten or misunderstood).
    - R must know what is to I a relevant and appropriate response.
    - R must be motivated to provide I with the information.
    - R must know how to provide the information.
    - I must accurately record R's responses.
    - The responses must reflect R’s meanings and intentions, not I's.
    - Other users of the data must understand the questions and answers the same way R and I do.

Survey research

- A group of individuals respond to or fill out more or less standardized questionnaires. The questionnaires may take different forms to investigate different hypotheses, but they do not involve freewheeling or spontaneous conversations.

- Although surveys can be relatively quick and cheap mean to obtain data, the researcher needs to think carefully about:

- Completion rates - If the response rate is low, either because individuals cannot be reached or because they refuse to participate, the researchers' ability to make statistical inferences for the population being studied may be limited. Also, those who do participate may differ systematically from those who do not, creating other biases. Increasing the size of the survey sample to compensate for low response rates may only increase costs without alleviating the problem.

- Sample-population congruence - how well the sample subjects represent the population, is always a major concern. Here we are speaking of how well the individuals in a sample represent the population from which they are presumably drawn. Bias can enter either through the initial selection of respondents or through incomplete responses of those who agree to take part in the study.

- Questionnaire length - if a survey poses an inordinate number of questions or takes up too much of the respondents' time, the respondents may lose interest or start answering without much thought or care.

Response quality

- Response quality = the extent to which responses provide accurate and complete information. It is the key to making valid inferences.

- Response quality depends on several factors, including the respondents' motivations, their ability to understand and follow directions, their relationship with the interviewer and sponsoring organization, and, most important, the quality of the questions being asked.

- Engaging respondents - it is important to get off on a good footing by introducing yourself, your organization, your purpose, your appreciation of their time and .trouble, your nonpartisanship, your awareness of the importance of anonymity, and your willingness to share your findings.

- Since the whole point of survey research is to accurately measure people's attitudes, beliefs, and behavior by asking them questions, we need to spend time discussing good and bad questions.

- Good questions prompt accurate answers; bad questions provide inappropriate stimuli and result in unreliable or inaccurate responses. When writing questions, researchers should use objective and clear wording. Failure to do so may result in incomplete questionnaires and meaningless data for the researcher. The basic rule is this: the target subjects must be able to understand and in principle have access to the requested information.

- Certain types of questions make it difficult for respondents to provide reliable, accurate responses. These include double-barreled, ambiguous, and leading questions.

Closed-ended questions

- The main advantage of a closed-ended question is that it is easy to answer and takes little time. Another advantage is that answers are easy to compare, since all responses fall into a fixed number of predetermined categories. These advantages aid in the quick statistical analysis of data.

- With open-ended questions, by contrast, the researcher must read each answer, decide which answers are equivalent, decide how many categories or different types of answers to code, and assign codes before the data can be analyzed.

- Another advantage of closed-ended questions over open-ended ones is that respondents are usually willing to respond on personal or sensitive topics (for example, income, age, frequency of sexual activity, or political views) by choosing a category rather than stating the actual answer.

- Critics of closed-ended questions charge that they force a respondent to choose an answer category that may not accurately represent his or her position. Therefore, the response has less meaning and is less useful to the researcher.

- Also, closed-ended questions often are phrased so that a respondent must choose between two alternatives or state which one is preferred. This may result in an oversimplified and distorted picture of public opinion. A closed-ended question allowing respondents to pick more than one response (for example, with instructions to choose all responses that apply) may be more appropriate in some situations.

Open-ended questions

- Unstructured, free-response questions allow respondents to state what they know and think. They are not forced to choose between fixed responses that do not apply. Open-ended questions allow respondents to tell the researcher how they define a complex issue or concept.

- Disadvantage of the open-ended question is that respondents may respond too much or too little. Some may reply at great length about an issue - a time-consuming and costly problem for the researcher. On the other hand, if open-ended questions are included on mail surveys, some respondents with poor writing skills may not answer, which may bias responses.

- Furthermore, unstructured answers may be difficult to code, interpretations of answers may vary (affecting the reliability of data), and processing answers may become time-consuming and costly.

Question order

- The first several questions in a survey are usually designed to break the ice. They are general questions that are easy to answer.

- Complex, specific questions may cause respondents to terminate an interview or not complete a questionnaire because they think it will be too hard. Questions on personal or sensitive topics usually are left to the end.

Describing your data

- The best way to learn about writing a data section is to read several data sections in the literature on your topic and pay attention to the kinds of information they contain. Your data section should do at least the following.

- Identify the data source. This means a sentence that explicitly says where your data come from.

- Describe the data source. You should tell your readers such things as the number of observations, the population groups sampled, the time period during which the data were collected, the method of data collection, etc.

- State the strengths and weaknesses of the data source. How do your data compare with other data sources used in the literature? Does yours provide more observations, and/or more recent observations, than other sources? Was the data collected in a more reliable manner? Why is the data source particularly suited (or not) to your study? Note any features of the data that may affect your results. Were certain populations overrepresented or underrepresented? Is there attrition bias or selection bias? Did the method of data collection change?

- Explain any computations or adjustments you made. Sometimes, a data source does not give you something directly; you perhaps had to add/subtract/multiply/divide two given pieces of data to get a third. Describe how you constructed your sample. Did you have to eliminate certain kinds of observations, for instance?

Descriptive statistics

- Data sections often contain a table of descriptive statistics, statistics of relevance about the sample. These statistics usually include the mean (e.g., mean income, mean age, mean years of schooling, etc.) and standard deviation. For categorical data (like race), however, you do not report a mean; instead, you report the percentage of the observations in each group.

- Expected value - The mean or average value of a sample statistic based on repeated samples from a population.

- Standard errors - The standard deviation or measure of variability or dispersion of a sampling distribution. The larger the sample, the smaller the standard error.

- Sampling distributions - A theoretical (non-observed) distribution of sample statistics calculated on samples of size N that, if known, permits the calculation of confidence intervals and the test of statistical hypotheses.

- NOTE: The mean and standard deviation work well for normal (bell curve shaped) distribution. If dealing with other distributions, it may be more useful to use median or mode to describe central tendency (expected value).

Plotting your data

- A well-constructed graph can answer several questions at one time:
- Central tendency: Where does the center of the distribution lie?
- Dispersion or variation: How spread out or bunched up are the observations?
- The shape of the distribution: Does it have a single peak (one concentration of observations within a relatively narrow range of values) or more than one?
- Tails: Approximately what proportion of observations is in the ends of the distribution or in its tails?
- Symmetry or asymmetry (also called skewness): Do observations tend to pile up at one end of the measurement scale, with relatively few observations at the other end? Or does each end have roughly the same number of observations?
- Outliers: Are there values that,compared with most, seem very large or very small?
- Comparison: How does one distribution compare to another in terms of shape, spread, and central tendency?
- Relationships: Do values of one variable seem related to those of another?

Choosing the right chart

TABLE 11–10 Typical Presentation and Exploratory Graphs

Type of Graph	What Is Displayed	Most Appropriate Level of Measurement	Number of Cases	Comments
Bar chart	Relative frequencies (percentages, proportions)	Categorical (nominal, ordinal)	3–10 categories	Common presentation graphic
Dot chart	Frequencies, distribution shape, outliers	Quantitative (interval, ratio)	*Less than* 50 cases	Displays actual data values
Histogram	Distribution shape	Quantitative	$N > 50$ cases	Essential exploratory graph for interval or ratio variables with a large number of cases
Boxplot	Distribution shape, summary statistics, outliers	Quantitative	$N > 50$ cases	Can display several distributions; actual data points, an essential exploratory tool
Time series plot	Trends	Quantitative (percentages, rates)	$10 < N < 100$	Common in presentation and exploratory graphics