

Data Mining in Economic Science

Ad Feelders

Let neither measurement without theory
Nor theory without measurement dominate
Your mind but rather contemplate
A two-way interaction between the two
Which will your thought processes stimulate
To attain syntheses beyond a rational expectation!

Arnold Zellner [Zel96]

1 Introduction

Data mining is commonly defined as the computer-assisted search for interesting patterns and relations in large databases. It is a relatively young area of research that builds on the older disciplines of statistics, databases, artificial intelligence (machine learning) and data visualization. The emergence of data mining is often explained by the ever increasing size of databases together with the availability of computing power and algorithms to analyse them. Data mining is usually considered to be a form of secondary data analysis. This means that it is often performed on data collected and stored for a different purpose than analysis; usually for administrative purposes.

In this chapter we consider the possibilities of applying data mining in economic science. In doing so, we must naturally be aware of the considerable amount of research that has already been done in economic data analysis. To what extent can data mining contribute to the analysis of economic data? In answering this question we could consider data mining as a collection of techniques and algorithms that have been developed in this area of research. In doing so we could compare data mining algorithms to data analysis techniques more commonly used in economics, and see if they allow us to answer different questions, or to answer existing questions in a better way. Alternatively, we can also consider data mining as a highly exploratory form of data analysis that is data driven rather than theory driven. The latter aspect of data mining is most important in this contribution.

This chapter is organized as follows. In section 2 we give a brief description of the object of study of economics. Then we consider economic modelling as a way to apply economic theory to particular problems and as a tool to deduce the consequences of particular assumptions. In order to give empirical content to

economic models data is required. In section 4 we give an overview of the types of data typically available in economics. In section 5 we briefly explain how economic data are used to quantify economic models. In section 6 we discuss data mining approaches to the analysis of economic data. In this section we also consider arguments for and against the use of data mining in the analysis of economic data. Finally, in section 7 we summarize the main points of our discussion and give an outlook on possible future developments.

2 Economic Science

Individuals, households, groups, whole economies, can be seen as facing the same problem: they have objectives but limited resources to achieve them. This limit on resources or constraints forces them to choose a course of action to achieve their objectives. Economics is the study of such choices: how they are made and what their implications will be. A classic illustration of an economic problem is the situation facing a consumer. The objective can be thought of as that of obtaining the greatest satisfaction from one's purchases of goods. The constraint is one's income and the prices of goods. If one's income were infinite or all prices zero then the economic problem would largely disappear. But with a limited income and positive prices one has to choose how, through one's purchases, one can achieve the greatest level of satisfaction.

It is common to divide economics into two main branches, micro- and macroeconomics. Microeconomics deals with the behaviour of single, or small units such as the individual consumer, the single firm, and the individual worker. It tries to answer questions such as what determines the price of a particular good, what determines the output of a particular firm or industry, and what determines the amount of hours of labour a particular worker is willing to supply. The defining characteristic of microeconomics is that the unit being analysed is relatively small.

Macroeconomics is the branch of economics which deals with the behaviour of aggregate or average variables, such as total output of the economy, total unemployment, and the average price of all goods produced in the economy. It attempts to explain the behaviour of these aggregates or averages and their interrelationships. The defining characteristic is that the unit being analysed is an aggregate or total.

3 Economic modelling

In order to apply economic theory to particular problems, and to deduce the consequences of particular assumptions economists often construct mathematical models. Such models typically take the form of (a system of) equations describing the relations between economic variables such as income, consumption, and interest rate. To give a simple example: in macroeconomics it is often assumed that in the short run total consumption C depends on national income

Y . We write this as $C = f(Y)$, where f is some function that we leave as yet unspecified. Other factors, such as interest rate presumably also affect consumption but for the purpose of this analysis we choose to ignore them. This is an example of the pervasive *ceteris paribus* condition often invoked in economic reasoning and modelling. Basically it means that we assume that all other relevant variables (such as interest rate in this case) remain the same. Now to state that consumption depends on income is a rather weak statement, what is meant actually is that when income goes up (down) consumption will *ceteris paribus* go up (down) as well. More specifically it is often assumed that the relation can be described by the linear equation

$$C = a + bY \tag{1}$$

The economic interpretation of the symbols in this equation is

- C : consumption
- b : marginal propensity to consume ($0 < b < 1$)
- Y : national income
- a : autonomous consumption ($a > 0$).

The linear form is often used for convenience, or as a first approximation, but is usually not implied by economic theory itself. Economic theory is usually qualitative in nature. It does for example not specify the exact values of a and b in the above equation, although it does constrain a to be positive and b to take a value between 0 and 1. The qualitative nature of such general models is understandable: one would guess that the parameter values will differ from country to country, and that within the same country their values will change over time. Nevertheless, the behaviour of a complex economic system may depend crucially on the specific values of these parameters. When economic models are used to support policy decisions, it is usually important to know their approximate values. To give empirical content to qualitative economic theories, statistical techniques are used to estimate the parameters of economic models from data.

4 Data in economics

In economics almost all available data are of observational nature; the data was not obtained by performing controlled economic experiments, but by passively observing economic reality. One of the consequences of the limited possibility to experiment in economics is a gap between theory, with its frequent invocation of the *ceteris paribus* clause, and the available data. For example to estimate the demand curve for oranges - the relation between the price of oranges and the quantity demanded - it is not sufficient to observe prices of oranges in different time periods and the corresponding quantities purchased. The reason is that the "other things" (e.g. income and the prices of other products) have the nasty habit of not remaining equal. In order to make a proper estimate of the

price-elasticity of oranges we would have to include other important influences on the demand for oranges in our analysis as well. If we were in a position to construct an experiment to obtain the relevant data, we could control for those other variables to make sure that the *ceteris paribus* clause is fulfilled.

Having noted that economic data are primarily of observational nature, we turn our attention to the different types of data structures typically encountered. Usually the following data structures are distinguished

1. Cross-section data: the observation of variables on different units (e.g. people, households, firms). For example, the observation of income of many different people results in cross-section data. The ordering of data does not matter.
2. Time-series data: the observation of variables at different points in time. For example, the observation of an individual's income at different points in time yields a time-series.
3. Panel data: the observation of variables on different units at several points in time. For example, the observation of income of different people at several points in time results in panel data.

Another useful subdivision of economic data is into micro-data and macro-data. Micro-data are collected on individual decision making units, such as individuals, households and firms. Macro-data results from aggregating over individuals, households or firms at the local or national level.

Lots of data on economic activity is collected on a routine basis. For macro economic data this is usually done by government bodies such as Statistics Netherlands and the Bureau of Economic Analysis and the Bureau of Labor Statistics in the US. A large amount of data is currently available on the World Wide Web. For example, Statistics Netherlands (www.cbs.nl) has an electronic database called StatLine that contains information on many economic and social topics. The Bureau of Economic Analysis (www.bea.doc.gov) and the Bureau of Labor Statistics (<http://stats.bls.gov/>) provide similar services as do government bodies in many other countries. A good place to start the search for economic data on the World Wide Web is Resources for Economist (www.rfe.org/Data/), edited by Bill Goffe.

5 Econometrics: quantifying economic models

As mentioned in section 3, the relations between variables postulated by economists are usually of a qualitative nature. Consider again the relation between total consumption and national income

$$C = a + bY \tag{2}$$

with the meaning of the symbols as specified in section 3. In order to give empirical content to such a model, we must have observations that allow us to estimate the unknown parameters a and b of this equation.

The discipline of econometrics concerns itself with the application of tools of statistical inference to the empirical measurement of economic models. Regression analysis is by far the most widely used technique in econometrics. This is no surprise, since economic models are often expressed as (systems of) equations where one economic quantity is determined or explained by one or more other quantities. Note that the economic model we start with is deterministic, i.e. it specifies an exact relationship between consumption and income. When we use observed economic data, for example a time series of consumption and income, we do not expect an exact relationship between the two. Equation (2) would lead to the following econometric model specification:

$$C_t = a + bY_t + e_t \quad (3)$$

Where t is an index for different time periods, and e is the error term. The error term accounts for the many factors that affect consumption but have been omitted from the model. It also accounts for the intrinsic uncertainty in economic activity.

According to the received view empirical economic research should proceed along the following steps (see for example [HGJ01], section 1.6)

1. The process starts with an economic problem or question. On the basis of economic theory we consider what variables are involved in the problem and what is the possible direction of the relationship(s) between them. From this we obtain an initial specification of the model and a list of hypotheses we are interested in.
2. The economic model must be transformed into an appropriate econometric model. One must choose a functional form (e.g. linear) and make assumptions about the nature of the error term.
3. Sample data are obtained, and an appropriate method of econometric analysis is chosen.
4. Estimates of the unknown parameters are made and hypothesis tests are performed, using some statistical software package.
5. Model diagnostics are performed to check the validity of our assumptions concerning relevant explanatory variables, functional form and properties of the error term.
6. The economic consequences of the empirical results are evaluated.

Note the dominant role of economic theory and the modest role of the sample data in this procedure. In the next section we discuss this issue in more detail.

6 Data mining in economics

In the previous section we have sketched a picture of economic data analysis that is largely driven by economic theories and models. In practice economic theory

is rarely so detailed that it leads to a unique model specification. There may for example be many rival theories to explain a certain economic phenomenon. Also, the usual *ceteris paribus* clauses of economic theory yield some choices to be made when it comes to the empirical estimation and testing of relationships.

In applied econometrics often alternative specifications are tried, and the specification, estimation and testing steps are iterated a number of times. Leamer (Leamer 1978) gives an excellent exposition of different types of specification-searches used in applied work. The search for an adequate specification based on preliminary results has sometimes been called data mining within the econometrics community [Lea78, Lov83]. In principle, there is nothing wrong with this approach, its combination however with classic testing procedures that do not take into account the amount of search performed have given data mining a negative connotation. Spanos [Spa00] uses the vivid analogy of shooting at a blank wall and then drawing a bull's eye around the bullet hole: the probability of the shot being in the bull's eye is equal to one. The proper way according to the classical view is to specify the model (i.e. drawing the target) before looking at the data (seeing where the bullet hole is).

In this section we discuss two approaches to economic data analysis that part from the classical approach outlined in section 5. They part from this approach in the sense that

1. The data is used extensively to search for a good model.
2. The models are “atheoretical” in the sense that received economic theory plays a minor role in the analysis.

The first issue is addressed in section 6.1, the second in section 6.2.

6.1 General-to-specific modelling

In the introduction, we characterized data mining as the search for interesting relations and patterns in data bases. We have seen that such a data based search for a good model specification is rejected by the traditional approach in econometrics. In practice however, researchers would start from their favoured theoretical model and “patch” the model, for example by including additional variables, if the data didn't agree (e.g. if a parameter estimate has the “wrong” sign). In this procedure one starts with the favoured model, which is usually a relatively simple theoretical model, and repair and extend it to uphold the favoured hypothesis if any data problems are encountered. Different researchers starting from different initial hypotheses will very likely end up with different models at the end of the day.

Others have argued that it is defensible to search the data for a good specification as long as this search is performed in a systematic and justifiable manner. An approach to econometric modelling that explicitly incorporates search is the general-to-specific modelling approach [HR82]. The main idea is to start with a complex model and to simplify it through the repeated application of statistical

tests on the significance of model parameters. The complex model we start with should ideally include the rival models concerning some economic phenomenon.

The model is taken to be of autoregressive distributed lag (ADL) form

$$y_t = \sum_{j=0}^m (\beta_j x_{t-j} + \delta_j y_{t-1-j}) + e_t \quad (4)$$

where m is the maximum number of lags considered. Such models are called autoregressive distributed lag models because they are a combination of an autoregressive model and a distributed lag model. In an autoregressive model the dependent variable y is explained by its own history (so called lagged values of y). A distributed lag model is a regression model in which the current value of y is explained by current and past values of one or more independent variables x .

Hoover and Perez [HP99] describe a simulation study in which they formulate a mechanical search procedure (one could say: data mining algorithm), which mimics some aspects of the search procedures used by practitioners of general-to-specific modelling. Full mechanization of the search procedures is very hard, because consistency with economic theory is also used by them to judge the acceptability of a candidate model. In this simulation study the true "data-generating process" (i.e. the model that generated the data) is known, and the data mining algorithm is assessed for its ability to recover this true model. They report fairly positive results, and in as far as the algorithm is shown to have some defects (such as a tendency of overfitting, i.e. inclusion of extra variables in the final specification) they suggest adaptations to overcome these problems. In a reaction to this study Hand [Han99] argues that the assumption that the true model is contained in the initial model is not realistic. Therefore one should not measure success by how often the search procedure yields the true model, or a model that includes the true model, but rather by how accurate the predictions of the final model are. Hand voices the opinion that the structure of the model is irrelevant because one can never know the true structure, but that models should be judged exclusively on their predictive performance. Needless to say that this is a very controversial point within economics.

6.2 VAR models

The discussion concerning the pros and cons of VAR (Vector Auto Regression) models provides a good illustration of the arguments for and against data mining in economics. As mentioned above the "traditional" approach to learning from economic data relies heavily on economic theory to provide a specification of the model. Economic models typically consist of a number of equations, one for each dependent variable, where each equation describes the relation between the dependent variable and a number of explanatory variables. These models are often referred to as structural models to emphasize that the mathematical equations depict (without exploiting possibilities of algebraic simplification) the detailed economic behaviour postulated by the model. Each equation in the

structural model either describes a hypothesized pattern of economic response or embodies a definition. A large body of econometrics is concerned with the estimation of such systems of equations on the basis of observed data. In practice it turned out that

1. Economic theory is usually not specific and detailed enough to arrive at a unique model specification (in other words many different specifications are consistent with economic theory).
2. Because of technical problems with the consistent estimation of such systems of equations, in many cases "incredible" (from the viewpoint of economic theory) assumptions have to be added to make consistent estimation feasible.

VAR models [Sim80] originated from a discomfort with this situation and also the observation that time-series models (not based on economic theory) were shown to have equal or better predictive performance than the so-called structural models. Essentially VAR models are an extension of time series models to multiple equation systems. For example, a two variable VAR(p) model looks like this

$$y_t = \alpha_1 + \gamma_1 t + \sum_{j=1}^p (\beta_{1j} x_{t-j} + \delta_{1j} y_{t-j}) + e_{1t} \quad (5)$$

$$x_t = \alpha_2 + \gamma_2 t + \sum_{j=1}^p (\beta_{2j} x_{t-j} + \delta_{2j} y_{t-j}) + e_{2t} \quad (6)$$

Here p denotes the lag length of the model. Thus if $p = 2$, we assume that all variables depend on the 2 previous values of all variables in the model, including itself. Of course we may use the data to search for a good value of p .

Koop [Koo00] gives the following example. Macroeconomic theorists have created many sophisticated models for the relationship between interest rates, the price level, money supply and real gross domestic product (GDP). A well-known example the IS-LM model extended for inflation, but there are many others as well. A VAR modeller would merely assume that interest rates, price levels, money supply and real GDP are related, and that each variable depends on lags of itself and all the other variables. Apart, perhaps, from the variables included in the model, there is no link between the empirical VAR and a theoretical macroeconomic model.

It is interesting to consider the arguments that have been brought to bear for and against VAR models as opposed to structural models. The major argument against their use is that they are not based on economic theory ("atheoretical") and therefore are useless in the advancement of economic science. They may be used for the purpose of prediction of economic variables but they do not shed any light on existing theory. The major argument in favour of VAR models is that they do not make incredible or arbitrary a priori assumptions concerning the relations between the economic quantities under study.

The basic points of disagreement then seems to be whether prediction per se is a legitimate objective of economic science, and also whether observed data should be used only to shed light on existing theories or also for the purpose of hypothesis seeking in order to develop new theories. Firstly, in our view prediction of economic phenomena is a legitimate objective of economic science. Models used for prediction may however be hard to interpret because they may have little connection with the way we understand economic reality. Secondly, it makes sense to use the data for hypothesis seeking. If not from empirical observation, how do scientists get their ideas for new theories?

7 Summary and outlook

In this chapter we have considered the role that data mining can play in economic data analysis. According to the "traditional" view of econometrics, the model to be estimated and the hypotheses to be tested should be specified a priori on the basis of economic theory.

The problem that practicing data analysts encountered is that economic theory is seldom specific enough to lead to a unique specification of the econometric model. For example, economic theory is always formulated with *ceteris paribus* clauses, does seldom specify the functional form of relations between variables and has little to say about dynamic aspects of economic processes.

Because of this practitioners have adopted ad-hoc methods of "data mining"; of using the data at hand for finding a good model and testing that model on the same data. Although it makes sense to use the data for finding a good model, using the same data for testing violates the assumptions of the testing procedure, unless the amount of search performed is somehow taken into account.

The pure hypothesis testing framework of economic data analysis should be put aside to give more scope to learning from the data. This closes the empirical cycle from observation to theory to the testing of theories on new data. Thus data mining is not a "sin" but can be made a valuable part of economic theory construction. A sample of data is mined to find interesting hypothesis, but the test of such hypotheses should be performed on data that was not used to create it in the first place. Of course such a procedure would require that enough data is available. This tends to be a problem in macroeconomic time series. In such cases some middle ground has to be found between complete a priori specification and a purely data based model search. The growing amounts of micro-data recorded about individual consumers and their purchasing behaviour however, provides great opportunities for data mining.

References

- [Han99] D.J. Hand. Discussion contribution on "data mining reconsidered: encompassing and the general-to-specific approach to specification

search” by hoover and perez”. *The econometrics journal*, 2(2):226–228, 1999.

- [HGJ01] R.C. Hill, W.E. Griffiths, and G.G. Judge. *Undergraduate Econometrics (second edition)*. Wiley, New York, 2001.
- [HP99] K.D. Hoover and S.J. Perez. Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The econometrics journal*, 2(2):167–191, 1999.
- [HR82] D.F. Hendry and J-F. Richard. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics*, 20:3–33, 1982.
- [Koo00] G. Koop. *Analysis of Economic Data*. Wiley, Chichester, 2000.
- [Lea78] E.E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental data*. Wiley, New York, 1978.
- [Lov83] M.C. Lovell. Data mining. *The Review of Economics and Statistics*, 65(1):1–12, 1983.
- [Sim80] C.A. Sims. Macroeconomics and reality. *Econometrica*, 48:1–48, 1980.
- [Spa00] A. Spanos. Revisiting data mining: hunting with or without a license. *Journal of Economic Methodology*, 7(2):231, 2000.
- [Zel96] A. Zellner. Past, present and future of econometrics. *Journal of statistical planning and inference*, 49:3–8, 1996.