SECTION: ECONOMETRICS

Gulara Rahimova,
Baku State University, Azerbaijan
ORCID: 0000-0002-7820-5837
rahimova.gulara.ali@bsu.edu.az;

Farhad Mirzayev,
Baku State University, Azerbaijan
ORCID: 0000-0001-8794-9187
farhadmirzeyev@bsu.edu.az;

PROBLEMS OF USING BIG DATA IN ECONOMETRIC RESEARCH

https://doi.org/10.30546/900510.2025.01.0106

Abstract
We live in the age of Big Data and it is no coincidence that big data has been dubbed the 'new oil'. The term 'big data' (or metadata) was first used by Clifford Lynch in 2008. Emphasising that the volume of information in the world is growing rapidly, he defined big data as any heterogeneous data matrix exceeding 150 GB (75*231 Gb) per day. Until 2011, big data analysis was carried out only within the framework of scientific and statistical research. However, in early 2012, the volume of data started to be expressed in very large numbers and serious needs arose for their systematisation, classification and practical use. These needs raised the question of developing new methods of analysis to make more accurate and justified economic decisions. Thus, the need to build clear econometric models and detect patterns as a result of collecting, processing and analysing regular and large amounts of data arose. Since econometrics is a science that quantitatively and qualitatively studies economic relationships using mathematical methods, statistical models and computer science, in this paper we examine the problem of applying big data, which is part of Data Science, in econometric scientific research and in building econometric models. Because only by using econometric methods can the digital economy make more reliable forecasts, evaluate cause-effect relationships and calculate risks with high accuracy. This paper aims to highlight the importance and potential of using big data in econometric research and to provide recommendations for researchers and practitioners who are trying to use these data effectively in their work.

Keywords: Econometric models, Big Data, econometric forecasting, Data Science, Information Technology, Machine Learning methods.

Introduction

In the modern era of digital revolution, the volume of data created and collected in various fields of activity is increasing exponentially. This exponentially increasing quantity is called "Large volumes of datas" or 'Huge datas', in other words 'Big Data'. Big data is different from the data volumes we are used to and represents voluminous, diverse, rapidly changing and structured or unstructured data that requires new approaches to its analysis and use. Big data is a generalised term that combines a group of concepts, technologies and methods for processing very large data volumes in distributed information systems, enabling the organisation of qualitatively new useful information (knowledge).

Big data:
- is not only the data itself, but also the technologies for processing and usage scenarios, which can be divided into several main categories;
- new algorithms suitable for searching and processing large data sets;
- new information management technologies that allow working with complex, heterogeneous and distributed data sources;
- high-performance computing systems;
- architectures and algorithms that allow processing information flows from high-speed networks, devices and sensors;
- high-performance and highly reliable distributed file systems capable of handling petabyte-scale data;
- technologies that allow integrating heterogeneous data from different sources - data synthesis and integration.

Although Big Data and Data Science are related to data, they are different fields. While Big Data focuses on the management and processing of large amounts of data, Data Science operates as an interdisciplinary field that uses scientific methods, algorithms and systems to extract information and make decisions in analysing this data. This includes statistics, machine learning, predictive modelling and data visualisation. Data scientists collect, clean, analyse and interpret data from a variety of sources. Big Data and Data Science are interdependent. Big Data defines the data and Data Science provides the tools and expertise to analyse it. Together, they enable organisations to make data-driven decisions, innovate and solve real-world problems. In general, the most advanced methods of using data are called Data Science, and this science is considered one of the most powerful tools for working with data. We have already mentioned that the concept of Data Science combines several subject areas. Firstly, it is statistics (socio-economic and mathematical). Socio-economic statistics studies various economic indicators, the methods of their collection and calculation. For example, the main purpose of these statistics is to study how various economic indicators such as Gross Domestic Product and inflation are calculated. Mathematical statistics, on the other hand, deals with the estimation of unknown parameters and the testing of statistical hypotheses. Data Science also includes econometrics, the study of the relationships between variables. For example, the Central Bank needs to understand what it needs to do to reduce inflation in a country. An econometrician builds a model by analysing the relationships between inflation and the variables that affect it, and can thus identify the variables that have the greatest influence on the acceleration of inflation. Information econometrics is emerging as a result of the application of information technologies - spreadsheets, general purpose statistical packages, econometric software packages, time series analysis programmes - in econometrics. In addition to econometrics and statistics, machine learning is also considered a very important part of Data Science. The most common machine learning problems are prediction and classification problems. All three disciplines mentioned above within Data Science have one thing in common: any study is based on retrospective data from the past. This approach is based on the assumption that cause-effect relationships persist over time. If cause-and-effect relationships break down, this becomes a separate research study, and economists look for reasons why regularisation doesn't work. Now we will explain what Big Data, econometrics studies and the relationship between Big Data and Econometrics are and how Econometrics can extract useful information from big data. Big Data, or big data analytics, is the use of data (data analytics) to extract useful information and make management decisions. Econometrics is a scientific discipline that studies economic phenomena and develops methods of economic measurement and analysis. Nowadays, since large amounts of data that are carefully examined (analysed) in economic research are presented in digital form, the need to use statistical and econometric methods has arisen. Econometrics is a scientific discipline that studies economic phenomena and develops methods of economic measurement and analysis. Nowadays, the need to use statistical and econometric methods has arisen as large amounts of data that are carefully examined (analysed) in economic research are presented in digital form.

And in the meantime, they say that econometric analysis of big data should be carried out. In the process of analysis, the developed methods of econometrics reveal the problems of working with big data. Because standard methods of analysis designed to work with small data volumes are ineffective in big data conditions, their application does not give the expected results (Chen, L., Wang. Y., & Lee, L., 2022). Therefore, the development of information technologies, methods of using big data and their analysis within the framework of modern econometrics research are becoming an integral part of the development of econometrics as a science. The aim of this research study is to examine and discuss the application of big data in the context of econometrics research, to illustrate the various methods, advantages and limitations of working with big data and the areas of economic research where the application of big data has shown significant results and impact, as well as to reveal the problems that researchers face when working with big data and to suggest possible ways to overcome these limitations. Understanding and using big data in econometrics research opens new horizons for building more accurate models, making predictions based on these models, uncovering hidden patterns, identifying cause-effect relationships and developing effective decision-making strategies in the economic field (Kumar S., & Gupta R.,2023). To understand the application of big data in econometrics in more depth, it is necessary to understand its definition and characteristics. The concept of "big data" includes several key features of data with different characteristics that require special approaches and methods of analysis:

- The first aspect of big data is its volume. Whereas before we worked with relatively small data volumes, today we have to deal with the processing of huge data volumes measured in terabytes and petabytes. For example, social networks, financial markets, medical diagnostics, transport (taxi) services and other domains generate large data volumes that require adaptation of traditional analysis methods (Liebowitz J., 2017);
- The second aspect is data diversity. Big data can be presented in various formats and structures. These can be structured data, such as databases and tables, and unstructured data, such as text documents, audio and video recordings, images and other formats. This includes data from a variety of sources - social media, sensors, online platforms and other sources. The variety of data requires different methods and tools for processing and analysing them;
- The third characteristic of big data is the speed at which it is generated. Big data can enter at a high speed and its realisation may require a quick analysis. For example, financial markets or social media monitoring generate large amounts of data in real time. To analyse them, it is necessary to have methods that allow data to be processed and analysed at high speed (Liebowitz J., 2017);
- The fourth characteristic of big data is the reliability of the data and the degree to which they come from reliable sources. Big data can be heterogeneous (inhomogeneous) due to high levels of noise, errors and various data sources. This can result from errors (inaccuracies) in data collection, inaccuracies and the presence of pseudonyms and inconsistent data. Therefore, data cleaning, preprocessing and reliability and quality control are considered important steps when working with big data.

Big data analysis presents specific problems that require specialised methods and techniques. Some standard methods and tools may not be able to cope with large amounts of data, and new approaches must be developed to ensure the efficiency and accuracy of the analysis.

Below we will examine specific methods of working with big data in econometrics and their capabilities and limitations.

Machine learning methods provide powerful tools for working with big data in econometrics. Machine learning algorithms such as regression, random forests, gradient boosting (decomposition) and neural networks can be successfully applied in forecasting economic indicators, analysing correlation dependencies and classifying data (Smith J., & Brown A., 2022).

They are based on the use of a large number of observations and can automatically detect subtle patterns and complex relationships in data. However, the effective application of machine learning methods requires a sufficiently large amount of data and high quality variables to train the models (Mayer-Schönberger V., Cukier K., 2013).

Panel data methods allow us to analyse data collected over time from many observation units. This allows us to account for temporal and inter-event dependencies and to control for the influence of individual characteristics. Panel data analysis can be performed using different models, such as random effects or fixed effects models. In this case, working with large datasets may require the use of efficient computational methods and algorithms, such as additional estimation and bootstrapping (Hastie T., Tibshirani R., Friedman J., 2016).

Data compression methods are approaches that aim to reduce data size without losing information. They are based on the use of various compression techniques such as matrix factorisation, technical truncation methods and data aggregation (concatenation). Data compression approaches can be useful for processing and storing large data sets, especially when computing power and memory resources are limited. However, it should be noted that data compression may result in the loss of some information, which may reduce the accuracy of the analysis. The application of big data methods to econometric analysis has been demonstrated in many studies. For example, by analysing large amounts of financial data, researchers can predict stock prices or classify market trends using machine learning techniques (Newbold P., Carlson W., Thorne B, 2014). Econometricians can also use panel data methods to analyse the impact of economic policies in different countries on domestic consumption growth. Data compression techniques can be useful in analysing large time series such as climate change data (Smith J., & Brown A., 2021). We have shown that econometrics uses statistical methods and mathematical models to study economic phenomena and analyse data. The incorporation of big data into the field of econometrics has led to an expansion of analytical capabilities and increased accuracy. One area where the application of big data has made a significant impact is in the analysis of financial markets. With the large amount of data available, researchers can more accurately analyse market movements, identify trends and predict changes in price indices and assets. Such data can include information on stock trading, retail sales, medicine, the Internet of Things, the property market, financial news and other factors that can influence financial markets.

The forecasting of economic indicators has also benefited significantly from the use of big data. By analysing large amounts of data, economists and financial analysts can make more accurate forecasts of important variables such as GDP, inflation, unemployment and other economic indicators. This helps managers make informed decisions and allows them to better plan economic activities. Macroeconomic analysis has also benefited significantly from the use of big data. The study of economic data on a national and regional scale allows us to identify interrelationships and dependencies between various economic factors. Analysts can use data on GDP, investment, trade, unemployment and other indicators to identify trends and inform strategic planning. Market competition research can also benefit significantly from big data analysis. Using data on prices, demand, sales and other market factors, researchers can identify competitive trends and determine the factors that influence the success of businesses in the market. This is essential for developing effective marketing and pricing strategies. However, there are some challenges associated with the use of big data in econometrics. Working with large amounts of data requires sufficiently powerful computing resources and data processing experience. It is also crucial to take into account the confidentiality of data and ensure data protection during their collection and analysis.

Analysing big data in econometrics requires taking into account a number of challenges and limitations that need to be taken into account when conducting research.

One of the main nuances is the processing and storage of large amounts of data. Analysing big data requires powerful computing resources and efficient algorithms. To overcome this, distributed systems for parallel computing and data storage can be used.

Gradient boosting is built sequentially and each base model learns from the mistakes of the previous one. However, some base models can be trained simultaneously using parallel computing to speed up learning.

It is also crucial to develop optimal methods for filtering and compressing data to reduce their volume without losing important statistical features. Another problem is that of sampling and representativeness. Due to the volume of big data and the diversity of sources, they may not be representative. It is important to develop methods that take this problem into account and allow analysis based on random selection to obtain statistically significant results. It is also necessary to carefully assess the degree of randomness of the data and systematic errors to avoid misinterpretation of the results. Issues of causality and interpretation of results are also important issues in analysing big data. Large data volumes may allow the discovery of statistically significant relationships between variables, but this does not necessarily mean that these relationships are causal. To overcome this problem, it is necessary to apply econometric methods that take into account the endogeneity of variables and test hypotheses about causal relationships (Liebowitz J., 2017).

It is also important to provide interpretable results that can be explained by economic theory and practical implications. For example, suppose we have a simple linear regression model:

$Y = \alpha 0 + \alpha 1 X + \varepsilon$,

where: Y is the dependent variable, X is the endogenous independent variable, $\alpha 0$- the drift coefficient, $\alpha 1$ the skewness coefficient, ε is the random error. However, if X is endogenous, ε correlated with the error, the values of the coefficients may be confounded. To eliminate endogeneity, we use an instrumental variable M that is correlated with X but not with the error.

$X = \pi 0 + \pi 1 M + \lambda$,

where: M is the instrumental variable, $\pi 0$ is the tool drift coefficient, $\pi 1$ is the tool deviation coefficient and λ is the tool error. We can now substitute the value of X from the second equation into the first equation: .

Now, if the instrument M fulfils the conditions for an instrumental variable, i.e. it is correlated with X and uncorrelated with the error, then we can use it to obtain robust and efficient estimates of the coefficients $\alpha 0$ and $\alpha 1$. This illustrates a simple example of the use of instrumental variables to account for endogeneity in econometrics. More complex models use various statistical tests and additional tools to ensure that the conditions for instrumental variables are valid (Newbold P., Carlson W., Thorne B, 2014).

To overcome these challenges, it is important to develop new methods and approaches in econometrics.

Recommendations.

1. The use of machine learning and artificial intelligence can help to analyse large amounts of data and find hidden patterns. However, it should be noted that these methods have limitations and require a careful approach when using them.
2. Information technology has undergone significant changes in recent years, resulting in the availability of large amounts of data that can be used for econometrics research. However, in order for these data to be useful, appropriate analysis methods need to be developed.
3. In general, the development of information technologies and analysis methods is an important factor for the successful use of big data in econometric research. Understanding the potential and limitations of big data will help researchers make informed decisions and make meaningful contributions to economic science and practice.

Conclusion. In the near future, Big Data will act as an important tool for decision-making from online businesses to entire countries and international organisations, becoming an indispensable tool for solving global problems such as fighting pandemics, curing cancer and preventing environmental crises, creating smart cities and solving transport problems. The application of Big Data in econometric research will allow more efficient use of economic indicators.

Referance:

1. Chen L., Wang. Y., & Lee L. "Blockchain Technology and Its Impact on Economic Data Integrity." Journal of Economic Perspectives, 36(3), 112-130 c., 2022
2. Kumar S., & Gupta R. "Data Security and Privacy Challenges in Economic Research: A Review." International Journal of Information Security, 25(1), 45-62 c., 2023
3. Smith J., & Brown A. "Machine Learning Applications in Econometrics: A Comprehensive Review." Journal of Applied Econometrics, 36(5), 689-715 c., 2021
4. Mayer-Schönberger V., Cukier K. – Big Data: A Revolution That Will Transform How We Live, Work, and Think/ Eamon Dolan Book, 272 с., 2013
5. Hastie T., Tibshirani R., Friedman J. – The Elements of Statistical Learning: Data Mining, Inference, and Prediction/ Springer, 767 с., 2016
6. Liebowitz J. – Big Data and Business Analytics/ Auerbach Publications, 304 с., 2017
7. Newbold P., Carlson W., Thorne B. – Econometrics: Statistical Foundations and Applications/ Springer, 592 с., 2014.