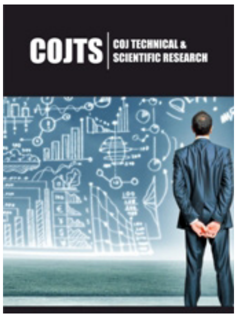
Crimson Publishers
Wings to the Research

Opinion

Econometrics and Data Science: An Econometric Perspective

Sunil K Sapra*

Department of Economics and Statistics, California State University, Los Angeles, CA 90032, USA



***Corresponding author:** Sunil K Sapra, Department of Economics and Statistics, California State University, Los Angeles, CA 90032, USA

Submission: May 20, 2019

Published: May 29, 2019

Volume 2 - Issue 2

How to cite this article: Sunil K S. Econometrics and Data Science: An Econometric Perspective. COJ Technical & Scientific Research.2(2). COJTS.000531.2019.

Opinion

Econometrics is a collection of tools used for detecting and summarizing relationships among variables of interest and is used widely in business and economics for decision making and policy analysis. The most common tools employed by econometricians towards these ends are linear regression or its generalization, generalized linear models to deal with non-continuous response. These models nevertheless are inadequate to deal with nonlinear relationships in the data as well as situations involving massive datasets and a very large number of variables, which have become increasingly common in many fields in recent years. Since the nature of nonlinearity is unknown in practice and standard econometric methods cannot be scaled up to analyze massive datasets, econometricians need to look for solutions elsewhere, such as data science-a new and vibrant field.

Data science is an emerging brand of statistics, which comprises the collection, preparation, analysis, visualization, management, and preservation of large amounts of data Saltz and Stanton [1]. It emphasizes development of fast and efficient algorithms for computation of estimators and forecasts with good performance rather than inferential justification for these algorithms and formulas. With the arrival of electronic computation in the 1950s, John W Tukey argued for a more application and computer-intensive discipline of statistics, which ultimately led to the creation of the field of data science. Unlike Econometrics, it is a statistical discipline, which dispenses with parametric models or formal inference and chooses prediction performance over inferential performance of models and procedures. It emphasizes algorithmic processing of large data sets for discovering useful information.

Varian [2] outlines some potential areas of collaboration between econometricians and data scientists. While econometricians have developed several useful techniques for causal inference, including instrumental variables, regression discontinuity, difference-in-differences, and natural and designed experiments, data scientists have focused mostly on development of prediction algorithms with good out-of-sample performance by avoiding overfitting. Varian [2] demonstrates that discovering important predictors among the many available predictors through the afore-mentioned causal inference techniques and incorporating them into the prediction model can potentially lead to more accurate forecasts. Another potential area for collaboration between econometricians and data scientists suggested by Varian [2] is model uncertainty. Econometricians often study the sensitivity of parameter estimates to the choice of controls and instruments or how an estimated parameter varies as different models are used. With the availability of big data, machine learning techniques popular in data science can be used to study these sensitivity issues more fruitfully. With ever-increasing size of datasets, techniques developed for the visualization and analysis of small and regular datasets are proving to be inadequate for dealing with the newly available large datasets, such as data on economic transactions. Data scientists have developed excellent predictive techniques which improve upon linear regression. These techniques include decision trees or ensembles of decision trees in the form of bagging, random forests, and boosting as well as neural networks to model complex nonlinear relationships and penalized regression techniques, such as LASSO and LARS for variable selection to circumvent overfitting if large datasets are available. Each of these techniques tunes a complexity parameter, such as the number of predictors, to obtain good out-of-sample predictions. The optimal value of the tuning parameter is chosen using cross-validation. Hastie et al. [3] provide an excellent discussion of these techniques. In addition, these algorithms use variable importance measures to determine the variables, which are important in prediction in the sense of contributing the largest improvement in prediction accuracy. An important limitation of these machine learning techniques is that these are designed for independently, identically distributed observations while the data in economics and finance tend to display correlation over time as well as non-stationarity Efron and Hastie [4]. However, as the Bayesian Structural Time Series Model, a machine learning technique developed by data scientists illustrates, some of these techniques can be adapted for time series data.

References

1. Saltz J, Stanton J (2017) An Introduction to data science. Sage Publishing, Los Angeles, USA.
2. Varian RH (2014) Big data: New tricks for econometrics. Journal of Economic Perspectives 28(2): 3-28.
3. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, data mining, inference, and prediction. (3rd edn), Springer-Verlag, New York, USA.
4. Efron, Bradley and Trevor Hastie (2016) Computer-age statistics. Cambridge University Press, New York, USA.