ANALYST

Co-funded by the European Union

MODULE 3/Multivariate, Temporal and Network Data Analysis

Carlo Drago PhD

University Niccolo Cusano, Rome

MASSIVE OPEN ONLINE COURSE (MOOC)

Project N. 2023-1-IT02-KA220-HED-000161770

ANALYST - A New Advanced Level for Your Specialised Training

HOZINT HORIZON INTELLIGENCE | fyg consultores | UNIVERSITY OF WESTERN MACEDONIA | Centro Studi AMIStaDeS APS | UNIVERSITÀ CUSANO | European Institute for Local Development | UNIVERSIDAD DE ALMERÍA

ANALYST

Co-funded by the European Union

MULTIVARIATE DATA ANALYSIS

This module introduces three topics: multivariate data analysis, time series analysis, and network analysis.

Multivariate data analysis study of data sets characterized by several variables. These techniques aim to identify patterns and relationships between variables. Tools such as principal component analysis (PCA) simplify and interpret complex data structures. See Härdle & Hlávka (2007)

Time series analysis analyzes data points collected over time. This allows the analysis of seasonal fluctuation, the prediction of the future, and the understanding of temporal patterns. Time-based data is often analyzed using ARIMA models and exponential smoothing.

See Brockwell and Davis (2002)

ANALYST

Co-funded by the European Union

NETWORK ANALYSIS

Network analysis can be used to analyze complex systems. Graph theory and social network analysis are used to investigate network properties, identify communities, and understand how information flows through networks.

- Overall these techniques empower analysts:
- to detect relevant relationships, understand systems of relationships (for instance in community detection)
- support evidence-based strategic decisions considering networks

Wasserman & Faust (1994), Fortunato (2010)

SIMPLE LINEAR REGRESSION

- Simple linear regression is a way of understanding the relationship between two variables. One variable, called the independent variable, influences the other, called the dependent variable.
- This method is about finding the straight line that most accurately represents the relationship between the data points. This line helps predict how changes in the independent variable will affect the dependent variable, making it an important tool for analyzing relationships.

See Greene (2000)

SIMPLE LINEAR REGRESSION

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Y is defined as dependent variable, where X are the independent variables. At the same time the coefficients are the betas. Note there is just a single explanatory variable explaining the Y

See Greene (2000)

MULTIPLE LINEAR REGRESSION

Multiple linear regression is a statistical technique for modeling the relationship between a dependent variable and two or more independent variables. This method aims to predict the outcome of a dependent variable based on the values of the independent variables (two or more independent variables).

This approach mainly involves analyzing data containing more than one variable to understand relationships and patterns. Multiple linear regression is an important part of this process as it helps to understand how multiple factors simultaneously influence a particular outcome.

Multiple linear regression is a powerful tool for making predictions and identifying significant relationships between variables in complex data sets.

MULTIPLE LINEAR REGRESSION

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

Y is defined as dependent variable, where X are the independent variables. At the same time the coefficients are the betas. Note there is mord than a single explanatory variable explaining the Y

See Greene (2000)

ASSUMPTIONS ON REGRESSION MODEL

Relevant assumptions of the regression model are:

- Linearity
- No perfect multicollinearity
- Correct structural form
- No autocorrelation
- Exogeneity
- Homoscedasticity
- Normality

See Greene (2000), Verbeek (2017)

MODEL SELECTION

Model selection in the context of multiple regression involves choosing the most suitable model from a set of potential models to best explain the relationship between the dependent variable and multiple independent variables. This process is crucial to ensure that the model is both accurate and interpretable. It involves several steps and considerations:

Criteria for Selection: Common criteria include the R square and the adjusted R-squared, which help determine the model with the best fit while penalizing for complexity to avoid overfitting. See Greene (2000).

MODEL SELECTION

Evaluation of Assumptions: Ensuring that the chosen model meets the assumptions of linear regression, such as linearity, independence, homoscedasticity, and normality of errors, is crucial for reliable results.

Complexity vs. Interpretability: Balancing model complexity with interpretability is essential. A simpler model may be preferred if it offers similar predictive performance with greater ease of interpretation. See Zellner (1979)

ANALYST

Co-funded by the European Union

CAUSALITY

Causality analysis in regression examines the cause-effect relationships between variables. It aims to determine whether changes in one variable lead to changes in another, rather than just determining correlations.

About causality see Pearl (2009)

CAUSALITY

There exists various different approaches in identifying causality:

- Instrumental Variables (IV)
- Difference-in-Differences (DiD)
- Regression Discontinuity Design (RDD)
- Granger Causality (used in time series data)
- Propensity Score Matching
- Causal Graphical Models

LOGISTIC REGRESSION

Logistic regression is a statistical method used for binary classification problems where the outcome variable is categorical and has two possible outcomes. It models the probability that a given input belongs to a certain category.

The approach uses a logistic function to map the predicted values to probabilities and ensure that they lie between 0 and 1. In contrast to linear regression, which predicts continuous values, logistic regression is particularly well suited to problems where the outcome is dichotomous, such as yes/no or true/false modalities (see Greene 2000, Verbeek 2017)

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical technique that simplifies complex data sets. It involves transforming the original variables into a new set of principal components.

These components are uncorrelated and ordered so that the first few retain most of the variation in the original data set. PCA helps to reduce the dimensionality of the data while retaining as much information as possible, making the data easier to visualize and analyze. It is used in machine learning and pattern recognition.

See Härdle & Hlávka (2007)

ANALYST

Co-funded by the European Union

TIME SERIES ANALYSIS

Time series analysis is a statistical technique for analyzing and interpreting data points collected or recorded at successive points in time.

This method can be used to identify patterns, trends and seasonal fluctuations in the data. It is usually used to predict future values based on historical data. By using models such as ARIMA, exponential smoothing, and seasonal decomposition, analysts can gain insight into both short-term fluctuations and long-term trends.

In order to predict some time series it is possible to use the Box & Jenkins procedure. For a comprehensive views on forecasting see Hyndman & Athanasopoulos (2018)

ANALYST

Co-funded by the European Union

TIME SERIES ANALYSIS

Relevant frameworks in time series analysis can be considered

- Classical time series analysis with the aim to identify structural characteristics analysis of the time series
- Modern univariate time series analysis with the aim to model different time series
- Modern multivariate time series analysis with the aim to model different time series considering multiple time series.
- Forecasting

See Brockwell & Davis (2002), Greene (2000) and Hyndman & Athanasopoulos (2018)

ANALYST

Co-funded by the European Union

NETWORK ANALYSIS

Network analysis examines how different nodes, which usually represent people, organizations, or computers, are connected within a network. It examines the structure and dynamics of these connections to understand how information flows, identify influential nodes, and uncover patterns or clusters.

By applying mathematical and computational techniques, network analysis helps uncover complex systems' underlying architecture and functioning.

See Wasserman & Faust (1994)

ANALYST

Co-funded by the European Union

NETWORK ANALYSIS

Typical network analyses consider:

- Structural Characteristics Analysis (the overall shape of the network)
- Centrality Analysis
- Community Detection
- Network Dynamics Analysis

See Wasserman & Faust (1994) and Fortunato (2010)

ANALYST

Co-funded by the European Union

REFERENCES

- Brockwell, P. J., & Davis, R. A. (Eds.). (2002). Introduction to time series and forecasting. New York, NY: Springer New York.
- Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.
- Greene, W. H. (2000). Econometric analysis 4th edition. International edition, New Jersey: Prentice Hall, 201-215.
- Härdle, W., & Hlávka, Z. (2007). Multivariate statistics. Barlin and Praue.
- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

ANALYST

Co-funded by the European Union

REFERENCES

- Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.
- Pearl, J. (2009). Causality. Cambridge university press.
- Tufte, E. R (1983). The visual display of quantitative information (Vol. 2, No. 9). Cheshire, CT: Graphics press.
- Verbeek, M. (2017). A guide to modern econometrics. John Wiley & Sons.
- Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.
- Zellner, A. (1979). Statistical analysis of econometric models. Journal of the American Statistical Association, 74(367), 628-643.