Introduction to regression

Week 1 (Introductory course)

Duration: 12 hours/3 days

Course Description

Regression is one of the most commonly used method for quantitative analyses: it can be applied in almost any field for research purposes, for driving informed decisions or, more generally, for describing our society. Regression analysis is so widely and frequently used, that sometimes one risks to apply it without taking into account important assumptions and prerequisites (for example about the used data). This could lead to misleading, confounding or even "wrong" conclusions (and, consequently, to "bad" decisions).

Thus, it is essential to deeply understand and carefully apply regression analysis methods. In order to reach this objective, we should be aware, first, about how this technique works. In addition, we should be able to explore the available data (and their quality), for making the best decisions about the models we aim at estimating. This knowledge would also enable us to assess the risks we could meet, to evaluate the limits of our results and, finally, to understand whether these findings are reliable.

This course starts from a very simple basis, i.e. from the "raw material" we have at our disposal: data. From this point, step by step, with a simple but complete and rigorous approach, it explains how to better set and develop a regression analysis. Attendees will be able to fully interpret the regression analysis results, evaluating their quality and understanding how "solid" their conclusions are. They will also be able to evaluate and compare potential alternative models and approaches, identifying the best one(s).

Every lecture is structured as a triple-step approach. First, a practical theme is proposed, starting from an example (usually based on a dataset and/or linked to a practical problem or decision to be made). Then, starting from this practical framework, an introduction of the statistical methodology takes place. This is done in order to show how, practically, one could face the issues linked to regression analysis. Finally, a practical part takes place, with the support of Stata. The analysis is set and results are read, interpreted and evaluated, in order to provide an answer to the starting problem.

Thus, the approach starts and is developed from a practical perspective. However, a complete technical introduction (despite in simple terms) will be also made.

All lectures will be organized trying to maintain a continuous interaction between the teacher and the attendees.

Prerequisites

Preferred prerequisites are a basic knowledge of statistics (despite the teacher will reprise the main concepts, when used) and a basic knowledge of Stata (the software that will be used for this course).

Schedule

The **12 hours** of this course will be shared into **three days/blocks** (of course, strongly linked). Each block consists of two **theoretical sessions** of about 90 minutes each; each of them is followed by a **practical Lab session** of about 30 minutes.

BLOCK 1: (est.: 4 hours)

THEORY: **our "bricks" & setting the simplest level of regression analysis**

Setting properly our "raw material": introduction to the regression analysis by means of a practical example and proposing a practical scope. This problem is, first, translated into technical terms, introducing the "language" used in a regression analysis framework and its technical notation. Then, potential issues with data are checked and fixed, preparing the dataset for the analysis. Starting from the basis, a practical introduction of simple linear regressions follows, including variables roles definition and analysis general purpose.

LAB SESSION (with STATA)

Exploratory analysis of available data on a sample dataset and assigning roles to regression variables. Setting the simple linear regression model estimation and defining the variables' role. Reading and practically interpreting the main part of the simple regression analysis output.

THEORY: **how simple linear regression analysis works and how to obtain analysis results**

How the analysis works: parameter estimation method (Ordinary Least Squares) and graphical interpretation of outcomes. What we can obtain as main results, what we can conclude.

LAB SESSION (with STATA)

Setting a simple linear regression analysis on STATA: comparison of different alternative models and of alternative outcomes; short overview on different software's outputs; discussion of different estimation examples drawn by participants.

BLOCK 2: from simple to multiple regression models: reading output and using its tests (est.: 4 hours)

THEORY: **understanding regression analysis parameters' test (and distributions)**

Introduction of the statistical tests used to draw conclusions on regression parameters and of the linked distributions (t test, confidence intervals and significance level, inference conclusions). Using statistical tests in general and in order to draw conclusions on regression parameters. Reading and practically interpreting the full simple regression analysis output.

LAB SESSION (with STATA)

Estimation of a simple linear regression model and discussion of different estimation examples: how to judge the regression parameters and how to practically use them. Practically reading and using the results of statistical tests on parameters: what is the consequence?

THEORY: **shifting from simple to multiple regression analysis: what changes?**

The multiple linear regression analysis: introduction on the model hypotheses and of the estimation method. How the inference on parameters is performed and how it can be interpreted. Criteria to read the full multiple regression model output in practical terms.

LAB SESSION (with STATA)

Estimating of a multiple linear regression model and overview on the obtained outputs: how to read them and how to act basing on drawn conclusions.

BLOCK 3: estimating procedures, models comparison, checking assumptions (est.: 4 hours)

THEORY: **how the multiple regression analysis leads to the final model**

The multiple regression analysis step by step: how to obtain the final reduced model. Introduction of the main iterative estimation methods and of estimation options (different order, different models). Treating factor variables and interactions among regressors. How to evaluate a model and how to compare different alternative models with few tools assessing the goodness of fit: ANOVA test, F distributions and R-squared. What are the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion); how the test for nested models works.

LAB SESSION (with STATA)

From a full to a reduced (final) multiple linear regression model. Evaluating and comparing alternative regression models using the main tests and statistics (ANOVA test, R-squared, AIC, BIC, test for nested models). Full overview of the analysis output and discussion of linked conclusions. As further examples, outputs coming from various software are shown, in order to enhance the flexibility in reading and interpreting regression results.

THEORY: **final check of obtained regression model and troubleshooting**

Importance of a final check about our regression analysis. We describe the main graphical and numerical tools used to evaluate and study the regression residuals and their assumptions. We explain how to practically use findings coming from these tools, in order to set properly the analysis. We propose best practices helping to properly choose regression variables and to explore and transform (if needed) the available variables in order to overcame potential issues on regression residuals. Issues such as the presence of high leverage observations or outliers will be checked and faced.

LAB SESSION (with STATA)

How to obtain regression residuals (and where to find them). How to study residuals and to interpret the analysis on residual assumptions. What can we conclude? Is our model, definitely, a “good model”? Strategies to enhance the model performance and goodness of fit. Treating and transforming (if needed) the available variables in order to overcame potential issues on regression residuals and other challenges.

TO CONCLUDE: In the last part of the course, further methods (alternative or complementary to the linear regression analysis) are also shortly proposed (logistic regression, multilevel models, and so forth); these are useful in specific frameworks and can be learned through a deeper, autonomous study of participants.

Short biography

Daniele Toninelli is Associate Professor in Economic Statistics at the Department of Economics (University of Bergamo, Italy). He graduated in «Statistics and Economics» (2003) and he obtained a first level Master degree in «Statistics for Marketing Researches and Survey» (2004) and a PhD in «Marketing for Enterprise Strategies» (2009).

He was research fellow at the University of Bergamo (2003 to 2007) and PhD student or visiting researcher at Statistics Canada (2008, 2009, 2012, 2013), at the University of Ottawa (2012, 2013), at the VŠB-Technical University of Ostrava (2012-2013), at the RECSM (Universitat Pompeu Fabra, Barcelona; 2014). He was guest lecturer at the University of Ljubljana (2018). He was a Management Committee member of the "WEBDATANET" network (European Project - COST Action IS1004; 2011-2015). His teaching activity, started in 2003 at the University of Bergamo, includes the following main courses: «Index Numbers Theory», «Statistics for Financial Markets», «Economic Statistics for Marketing Research», «Advanced Business Statistics», «Economic Statistics», «Data Production and Analysis» and «Advanced Probability and Statistics for Business and Finance».

His main research interests include: survey & web survey methodology, price indexes estimation, statistics for finance, use of big survey data and use of data collected through social networks in estimating economic and social indicators (wellbeing, in particular).