Introduction to Regression Analysis

Week 1 (Introductory course)

Duration: 12 hours

Format: In person and online

Instructor: Daniele Toninelli (daniele.toninelli@unibg.it)

Course description:

Regression is one of the most commonly used method for quantitative analyses: it can be applied in almost any research field. Nevertheless, there is the risks of applying it without taking into account important assumptions about data or without knowing how this technique exactly works. This could affect the results interpretation or it could lead to confounding (or even misleading) conclusions and decisions.

This course aims at providing a basic but solid introduction about regression analysis. It starts from the basis, i.e. from raw data (carefully exploring them) and it explains how a regression model estimation works. Then, step by step, with a simple but rigorous approach, this course explains how to better set and develop a regression analysis. Attendees will be guided in fully interpreting obtained results. Moreover, they will become able to evaluate results' quality and reliability (generally and comparing alternative models and approaches).

Almost every lecture is structured as follows. First, a practical example (with a knowledge gap to be filled) is proposed. Then, starting from this, a theoretical introduction of the regression methodology follows. Finally, using Stata and with a full involvement of participants, we set and run the regression analysis and we focus on interpreting results (always with a practical perspective), evaluating and using them in order to provide an answer to the starting problem.

In this course, an active involvement of the attendees will be constantly encouraged and stimulated, proposing tasks and discussions during lectures as well as homework.

Learning schedule:

Day 1	**Theory**	***Getting ready for the analysis*** Introduction of the regression analysis by means of a practical example and proposing a practical scope: how this technique can help us? The problem we want to study is, first, translated into technical terms, introducing the official notation used in a regression analysis framework. Then, we provide suggestions about how to prepare and explore data: potential issues with raw data are checked and fixed. Best practices to choose variables and to assign them a role are proposed.
	Practice (with Stata)	Exploratory analysis of data on a sample dataset. Then, we will choose and assign roles to available variables, setting our research scope. Do we have the necessary information?
		Break
	Theory	***Estimating my very first model*** Starting from the basis, we provide a practical introduction of simple linear regressions, including its general purpose. We then describe how the analysis works, introducing the Ordinary Least Squares (OLS) parameter estimation method. Then, we explore what we can obtain as main results (considering different outputs) and what we can conclude.
	Practice (with Stata)	Set and estimate a simple linear regression model, using Stata. We start practically reading the main part of a simple regression analysis output, interpreting the regression parameters and understanding how to practically use them. We ask to participants to set a personal simple regression

		analysis and to interpret results using alternative/different datasets we provide.
Day 2	**Theory**	***Introducing statistical tests*** Introduction of the statistical tests used to draw conclusions on regression parameters. Hypotheses, reference distributions, significance level, decision rules and test conclusions. We read and practically interpret the full simple regression analysis output. We provide a short overview on different software's outputs (SPSS, SAS, R, Stata, …)
	Practice (with Stata)	Interpretation of different regression output examples drawn by participant. Read and use practically results of statistical tests on regression parameters: what is the conclusion?
		Break
	Theory	***Simple to multiple: towards the final model*** Multiple linear regression analysis: introduction of the model notation, hypotheses and 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.
	Practice (with Stata)	Estimate of a multiple linear regression model and overview on the obtained outputs: how to read them and how to act basing on drawn conclusions. Each student will obtain and interpret a different model using Stata.
Day 3	**Theory**	***How to evaluate your model***

		How to obtain the final reduced model step by step. Introduction of the main iterative estimation methods (backward, forward, stepwise). We evaluate the goodness of fit of a model: ANOVA, *F* distribution and test, R^2. Compare alternative models using AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). Test of nested models. Outputs coming from various software are shown, in order to enhance the flexibility in reading and interpreting regression results.
	Practice (with Stata)	From a full to a reduced (final) multiple linear regression model using iterative methods. We evaluate and compare alternative regression models. Full overview of the analysis output and discussion of conclusions, also comparing alternative models estimated by attendants using Stata.
		Break
	Theory	***Last checks on residuals*** Final check about the regression model: tools used to evaluate and study residuals and their assumptions. We explain how to practically interpret these tools, in order to judge the regression analysis. We propose best practices helping to overcame potential issues on regression residuals (e.g., high leverage observations or outliers, skewed distribution, …): data transformation and treatment.
	Practice (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 about the estimated model? Strategies to enhance the model performance and goodness of fit.

Prerequisites: a very basic knowledge of Statistics (base course) is preferred, as well as a basic knowledge of Stata.

Software: Stata

Readings: Agresti, A. (2018), "Statistical methods for the social sciences" - 5th edition, Pearson, ISBN: 9781292220314.

Instructor short bio:

Daniele Toninelli is Associate Professor in Economic Statistics at the Department of Economics (University of Bergamo, Italy) since 2021 and formerly Assistant Professor since 2007 (University of Bergamo, Italy).

He graduated in Statistics (2003, University of Milan-Bicocca) and he obtained a Master degree in «Statistics for Marketing Researches and Survey» (2004, University of Milan-Bicocca) and a PhD in «Marketing for Enterprise Strategies» (2009, University of Bergamo).

Previously, he was research fellow at the University of Bergamo from 2003 to 2007. He was 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 (Research and Expertise Centre for Survey Methodology, 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). He is directing, from 2022, the post-university course «Data Analyst for Strategic Decisions - DADS» (2023, 2024; https://sdm.unibg.it/corso/dads/).

His past professional experiences include employments at IBM Italia/Celestica (1994–2001), PiTre S.r.l. (2000–2001) and Multiplex Arcadia (2002–2003).

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», «Quantitative Methods for Business Data Analysis» and «Advanced Probability and Statistics for Business and Finance». He teaches also short courses about SAS, SAS Programming and Data Visualization.

His main research interests include: survey and web survey methodology, collection and use of big data (or social media data) in estimating economic and social indicators (wellbeing, gender differences and discrimination), price indexes and composite index estimation.

Contacts:

- Email: daniele.toninelli@unibg.it.
- Webpage (with full CV): https://www.unibg.it/ugov/person/1582.