UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Linear Models for Categorical Data

2324-1-F8203B010-F8203B011M

Learning objectives

The course introduces the linear models for categorical data according to two different settings. The first concerns the general linear model (GLM), including several special cases such as ANOVA and ANCOVA models. The second setting deals with the generalized linear models, particularly Poisson log-linear models for count data and binomial logistic models, from a GLM perspective. Analyses of empirical cases are carried out through the SAS software.

Knowledge and understanding. This teaching will provide knowledge and understanding concerning:

- the most common linear models used for categorical variables, be they on the side of the dependent variable or the independent variables or on both sides
- the conventional methods for deepening analyses through appropriate definitions of model parameter functions, which aimed in particular at comparing specific groups or categories of statistical units
- the main procedures implemented in the SAS software for the construction of linear models for categorical data and the subsequent deepening of the analyses with the relative graphical representations
- the reading and interpretation of the analysis outputs produced by the SAS software.

Ability to apply knowledge and understanding. At the end of the course, the students will be able to:

- choose the most proper linear model based on the dependent variable properties
- assess the opportunity of using a specific linear model for categorical data according to a priori formulated goals also concerning the type of study and the nature of the available data
- interpret the meaning of the interaction parameters included in the model for two categorical variables jointly considered and deepen the results of the analyses relying on strategies that take into account the significance or not of these interactions
- use the main SAS procedures addressed to categorical data modelling by overcoming the default settings and using the most advanced statements to customise and deepen the analyses.

The course allows the student to acquire the main theoretical and applicative bases relating to the specification and

set-up of linear models for categorical data necessary in any working context where data files are used and for advancing university studies.

Contents

General Linear Model (GLM), one-way or more than one-way ANOVA and ANCOVA models. Generalized Linear Models (GzLM), binomial logistic model and Poisson log-linear model. Applications to real and experimental data with the SAS software.

Detailed program

- Theory of general linear model (GLM): model specification, assumptions, generalized inverse, estimable functions, testable hypotheses. Link with the constrained least-squares estimation method. Sum-to-zero and set-to-zero linear constraint approaches. Effect parameterization vs reference category parameterization. Contrasts
- Special cases of GLM: one-way or more than one-way fixed-effects ANOVA models, ANCOVA model. SAS PROC GLM
- Generalized Linear Model (GzLM): probability distribution function of response variables, link function, model specification, maximum likelihood estimation method, estimator properties, criteria for goodness-of-fit, confidence limits and statistical testing hypotheses
- Special cases of GzLM: Poisson log-linear model for count data and binomial logistic model, in a GLM perspective. SAS PROC GENMOD

Prerequisites

Knowledge of the topics covered in undergraduate Statistical Inference and Statistical Models courses is recommended.

Teaching methods

Theoretical lectures in the classroom and practical exercises in the statistical-informatics laboratory with the SAS software.

Assessment methods

The exam consists of preparing statistical data analysis with the SAS software (according to the rules specified in the course e-learning platform), whose output has to be discussed during the examination, and in a written test (duration: 2 hours) concerning both theoretical and practical topics.

The theoretical questions are general and aim at verifying the theoretical knowledge acquired on the logic and

advanced aspects underlying the model specification in the presence of categorical data (on the side of the dependent variable and/or of the independent variables), the various types of model parameterization, the notions of estimability and testability of parameter functions, and the drawing of statistical inference for such models. They also allow verifying the ability to use the symbolic-formal statistical language and appropriately provide definitions. The parts with a more methodological nature are the object of an optional issue that allows verifying the ability to prove the most advanced theoretical results analytically.

The practical questions concern the identification, the construction and the use of the most appropriate modelling for real situations and real data, and the definition of the analysis design most suited to satisfying a priori defined study objectives. The statistical data analysis, which has to be prepared before the examination and then presented during the test, constitutes the part of the exam in which these aspects have considerable emphasis since it requires the student to work critically and in total autonomy, especially in the definition and in the achievement of the study objectives. The practical questions ultimately allow verifying the ability to understand real problems and propose solutions in data analysis, the competence in reading and interpreting the analysis results, and the ability to carry out the required analyses using SAS procedures.

Furthermore, the methodology for preparing the statistical data analysis with SAS is assigned nominally and randomly (using a random number generator) to each student enrolled in the e-learning platform of the course. This analysis must be prepared before the exam. The student must follow a specific track relative to the methodology assigned and published at the end of the course on the e-learning platform. During the examination, the student must present the printed output according to the modalities specified in the e-learning platform of the course.

Given the abundance of teaching material uploaded on the e-learning platform of the course, no distinction is made between exams for attending students and exams for non-attending students. Finally, there is no ongoing test.

Textbooks and Reading Materials

- Teaching material uploaded on the course e-learning website (restricted access)
- Agresti, A. (2002), Categorical Data Analysis, Second Edition, New York: John Wiley & Sons
- Dobson, A., and Barnett, A. (2018), An Introduction to Generalized Linear Models, Boca Raton, FL: Chapman Hall/CRC, Fourth edition
- Littell, R. C., Freund, R. J., and Spector, P. C. (2002), SAS for Linear Models, 4th Edition, Cary, NC: SAS Institute Inc.
- Searle, S. R., and Gruber, M.H.J. (2017), Linear Models, 2nd Edition, John Wiley & Sons, Hoboken, New Jersey

Semester

First semester, second period

Teaching language

Italian

Sustainable Development Goals

QUALITY EDUCATION