MA8701 Advanced methods in statistical inference and learning

Part 5: Closing

Mette Langaas

4/13/23

Discussed 14.04.2023

Outline

- Learning outcomes and compulsory activity
- Take home messages from the Data analysis project
- Final evaluation
- Plan for the last three sessions!

Learning outcome

(The student should be able to)

1. Knowledge
   - Understand and explain the central theoretical aspects in statistical inference and learning.
   - Understand and explain how to use methods from statistical inference and learning to perform a sound data analysis.
   - Be able to evaluate strengths and weaknesses for the methods and choose between different methods in a given data analysis situation.

2. Skills

Be able to analyse a dataset using methods from statistical inference and learning in practice (using R or Python), and give a good presentation and discussion of the choices done and the results found.

3. Competence

- The students will be able to participate in scientific discussions, read research presented in statistical journals.
- They will be able to participate in applied projects, and analyse data using methods from statistical inference and learning.

Compulsory activity 2023

- Data analysis project (analyse, write report, review)
- Article presentation (present and discuss)

Take home messages from the Data analysis project

Short overview of the five data analysis projects

Team SuperGreat

- Data set: Framingham coronary heart disease (CHD), N=4238, p=15 (categorical, binary, continuous), binary response (15.2% cases).
- Aim: Understanding effects of covariates for prediction of CHD (10 years follow-up) and compare complete case and single imputation results.
- Missing: 13.7% in total (highest for glucose with 9%)
- Methods used: Single imputation vs complete case, boostrapping, lasso logistic and logistic regression, AIC.
- Result: important risk factors are age, male, systolic blood pressure, glucose (and for the imputed data also cigarettes pr day smoked).

Team CDF

- Data set: wine quality, N=6497, p=12 (binary, continuous), binary response (from dividing approximately in two).
- Aim: investigate how different physiochemical variables affect wine quality.
- Missing: 22.5 % in total (but not all imputed)
- Methods used: Single imputation vs complete case, lasso logistic and logistic regression, polyheder inference, train-test split for ROC-AUC.
- Result: important variables for wine quality was volatile acidity, residual suger, free sulfur dioxide, total sulfr dioxide, sulphates and alcohol.

Team Balance

- Data set: robotic arm kinematic data, N=17560 but reduce to N=176 to avoid time series correlations, p=28.
- Aim: A theoretical model for the movement of the robot arm exists, involving trigonometrical functions - giving background to considering a sum of second order polynomials of the covariates. The aim is then to arrive at an interpretable simplified model.
- Missing: no missing data.
- Methods used: ACF/PACF, train-test split, OLS, elastic net, multi-sample splitting (median) on training data.
- Result: Only one covariate "left" after multi-splitting, and this covariate did not give a sensible physical interpretation.

Team JAA

- Data set: superconductor critical temperature, N=21263, p=82 (very multicollinear), response: continuous critial temperature.
- Aim: To construct a prediction metod for the critical temperature of the superconductor so that the important factors influencing the critical temperature is understood.
  "Critical temperature (of a substance) can be defined as the highest possible temperature value at which the substance can exist as a liquid."
- Missing: no missing data (?)
- Methods: Forward selection with least squares, lasso, group lasso. Bootstrapping on single split data.
- Results: None of the models gave a small and interpretable model.

Team JKP

- Data set: Genome-wide association study, N=1796 p=183155 SNPs (for model selection) + 4 (23) clinical covariates, response=length of house sparrow wing at age 1 year.
- Aim: Which SNPs are associated with the reponse?
- Missing: present. Quality control defaults to removing SNPs and individuals with high missing rate. For the remaining missing data are imputed by single (mean) imputation (and often totally imputed SNPs are analysed).
- Method: Lasso regression (with snpnet and manual 10-fold CV for λ) and multi-sample splitting.
- Results: all SNPs had adjusted p-values of 1. No findings.

Group work

1) For all groups
   - a) What are you most proud of in your work?
   - b) What could have been done differently?
   - c) Choose one learning experience to share!
2) Specific questions for each group on paper hand-out
3) If you finish before we summarize: Discuss your study plan for the oral exam

Proud

a) The group work process

Implementing the multisplit

— " —————— and many qq plots

See the nuances on data leakage decisions

See a solution – indeed a reduced data set!

b) Differently

Explain better, try more methods

λ-tuning within multisplit

temporal analyses

multisplit

other data sets

c) learning experience

- sent folder
- "no results" is fine
- model the data 'correctly'
- don't underestimate multicoll.
- imputation model & response analysis

Common themes

Negative (or no) results!

Regression model

A linear regression model (or the linear predictor in the GLM) is linear in the regression parameters,
not necessarily in the covariates. In addition interaction term may be needed for a good model.
For tree-based methods any non-linearity in the covariates and interactions between covariates are *easily* picked up, but for methods like the lasso, we need to specify the linear predictor ourselves.
How to make sure the "right" linear predictor is used?

Missing imputation

- Specification of the imputation model in missing imputation
- Should the analysis model response be a covariate in the imputation model?

ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 1092–1101

Journal of Clinical Epidemiology

SPECIAL SERIES: ORIGINAL ARTICLES

Using the outcome for imputation of missing predictor values was preferred

Karel G.M. Moons[a,d,*], Rogier A.R.T. Donders[a,b], Theo Stijnen[c], Frank E. Harrell, Jr[d]

[a]*Julius Center for Health Sciences and General Practice, University Medical Center, Utrecht, P.O. Box 80035, 3508 GA Utrecht, The Netherlands*
[b]*Department of Innovation Studies, Copernicus Institute, Utrecht University, P.O. Box 80125, 3508 Tc Utrecht, The Netherlands*
[c]*Department of Epidemiology & Biostatistics, Erasmus University Medical Center, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands*
[d]*Department of Biostatistics, Vanderbilt University Medical Center, S-2323 Medical Center North, Nashville, TN 37232-2158, USA*

Accepted 10 January 2006

Abstract

Background and Objective: Epidemiologic studies commonly estimate associations between predictors (risk factors) and outcome. Most software automatically exclude subjects with missing values. This commonly causes bias because missing values seldom occur completely at random (MCAR) but rather selectively based on other (observed) variables, missing at random (MAR). Multiple imputation (MI) of missing predictor values using all observed information including outcome is advocated to deal with selective missing values. This seems a self-fulfilling prophecy.

Methods: We tested this hypothesis using data from a study on diagnosis of pulmonary embolism. We selected five predictors of pulmonary embolism without missing values. Their regression coefficients and standard errors (SEs) estimated from the original sample were considered as ''true'' values. We assigned missing values to these predictors—both MCAR and MAR—and repeated this 1,000 times using simulations. Per simulation we multiple imputed the missing values without and with the outcome, and compared the regression coefficients and SEs to the truth.

Results: Regression coefficients based on MI including outcome were close to the truth. MI without outcome yielded very biased—underestimated—coefficients. SEs and coverage of the 90% confidence intervals were not different between MI with and without outcome. Results were the same for MCAR and MAR.

Conclusion: For all types of missing values, imputation of missing predictor values using the outcome is preferred over imputation without outcome and is no self-fulfilling prophecy.

Keywords: Bias; Imputation; Missing predictors; Precision; Prediction

The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data

Angela M. Wood[*,1], **Patrick Royston**[2], and **Ian R. White**[3]

[1] Department of Public Health and Primary Care, Strangeways Research Laboratory, University of Cambridge, Worts Causeway, Cambridge CB1 8RN, UK
[2] MRC Clinical Trials Unit at UCL, Aviation House, 125 Kingsway, London WC2B 6NH, UK
[3] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK

Received 7 January 2014; revised 2 October 2014; accepted 13 October 2014

from the observed outcome requires consideration. Previous work (Schafer, 1997; Moons et al., 2006; Sterne et al., 2009; White et al., 2011) emphasizes the importance of including the outcome in imputation models in order to maintain observed relationships between covariates and the outcome as required for regression modeling. However, since model performance measures are defined as functions of the observed outcomes and predictions, we hypothesize that such measures may be optimistic when the predictions are constructed from a set of imputed covariates partly derived from the observed outcomes.

Evaluation: Oral exam

May 10, 15 and 22.

Pass/fail, with B as pass limit.

- On the last lecture (April 24) a list of five possible topics (questions) will be available at https://wiki.math.ntnu.no/ma8701/2021v/exam.
- If you want you may prepare a 5-10 minutes presentation of one of the topics (bring notes, but no slides, talk and write by hand) to be held in the start of the oral exam.
- The rest of the exam is general questions from the reading list (no notes)

Total duration $<$ 30 minutes.

Plan ahead

- Next week: https://wiki.math.ntnu.no/ma8701/2023v/assignmentsap#presentation_schedule
- Monday April 24: Discussion on central topics for each part of the course, and present the five possible topice to prepare for the first part of the oral exam.