Analysis of Continuous Data and Statistical Modelling

Project 24/25 Semester 1

Table of contents

Intro 1

Practical Assignment 2

Theoretical Exercise 4

Practical Information 5

References 5

Intro

SUPPORT was a prospective cohort study, designed to develop and validate a prognostic model for 6 month survival of seriously ill hospitalized patients (Knaus 1995).

A subset of the corresponding data is obtained from http://hbiostat.org/data courtesy of the Vanderbilt University Department of Biostatistics. The data is available as `support.dta` and `support.txt` on Ufora. An overview of the variables and their meaning is available in `info_support.xlsx`.

Practical Assignment

Our continuous outcome of interest is the total hospital cost (`totcst`). The primary set of predictors contains age, sex, disease class (`dzclass`), number of comorbidities and years of education (`edu`).

For the part with binary outcomes (for Statistical Modelling only), In-Hospital-Death (`hospdead` - yes/no) is the endpoint of interest.

1. **Define a relevant question**.

    a. This should be on the association between the continuous outcome and one of the predictors of interest and how this association may depend on other predictors.
    b. For the binary endpoint, prediction is the aim.
    c. For both parts, consider at least three models: 1) a model only containing the key predictor of interest, 2)the model containing the key predictor and the selected predictors - main effects only and 3) a model containing these same predictors, but applying model building techniques and potentially adding interactions.

2. **Protocol**:

    - Carefully state your research question (broad and specific - in words and formulae) and a brief (non-statistical) explanation on why one may be interested in this specific question.
    - Describe how you will approach and perform the initial descriptive statistics, the linear regression(s) and any further analyses you envisage. Don't forget the diagnostics.
    - For the binary endpoint, clearly define how prediction-performance of the model will be evaluated
    - On a separate page: make up a preliminary distribution of tasks for the team members with time line.

Part 1

3. Randomly split the dataset into 80% training and 20% test-data.

4. Perform the analyses on the training-data as set out in the protocol and decide on a 'final' model.

5. For the final model, perform model diagnostics and outlier detection.

6. Interpret key parameters in the model.

7. a. Use the final model to estimate the average cost over all the people in the sample
    b. We are now 30 years after the study ran. Could you use the model to similarly estimate the average cost for patients in hospital now? Explain why (not).

8. Now fit the final model to the test-data.

    a. Compare the estimates and their standard errors with those on the training set. How are the SEs in the test- and training set related?
    b. You can derive prediction intervals for the final model, from the training-dataset. Use the test-data to assess their coverage. Evaluate.

9. Go back to the training data and add Length of Stay (`slos`) to the final model. Again interpret the key parameters. Is including Length of Stay a good idea, given your research question? Explain.

Part 2

10. Apply the models and the model building to the training set.
11. Interpret the key parameters in all models
12. Evaluate the goodness of fit and the predictive performance of the models
13. Evaluate sensitivity and specificity. Add confidence intervals for these measures?
14. Add Length of Stay (`slos`) to the final model. Again interpret the key parameters and assess model performance. Is including Length of Stay a good idea, given the research aim? Explain.

Report

15. Write a concise report, explaining what has been done and how to interpret the results:

    a. Research question and methods
    b. Overview of the results
    c. A discussion of your main findings, with interpretation of key covariates and critical assessment. What have you learned and what are the limitations of your method and the obtained information? What would you do differently next time. What do you plan/recommend for further research?
    d. Conclusions

16. Some questions that deserve special attention / that **must be discussed**:

- Which question(s) would a uni-variable (key-predictor only) and a multivariable (selected other predictors as well) regression answer.
- What are the potential reasons for / is the potential impact of missingness in outcome and predictors
- Discuss (observed and to-be-expected) deviations from assumptions and their potential impact
- Discuss confounding and effect-modification

Theoretical Exercise

Additionally, there is a theoretical assignment based on the article on *Sugar-Sweetened Beverages and Genetic Risk of Obesity* by Qibin et al. (2012).

In groups of 4, one should take the lead on questions 1 and 2, the others lead on one question each.

First in Table 1:

1. In Table 1, we have considered a possible simple linear relationship between the categorical variable (with 4 levels: 0,1,2,3) *Servings of sugar sweetened beverages' and the continuous variable* Physical activity — MET-hr/wk'. Now look at these same data and write down a multivariate linear model that allows a separate mean outcome at every level of the categorical variable. Compare it with the simple linear model from before.
2. Of all the assumptions imbedded in this multivariate model: which are still the same assumptions on these data as for the earlier simple linear model and which have changed?
3. Calculate the standard errors for the derived estimated mean outcomes first for category 0, then for category 2 (or at each of the levels of the categorical variable).
    a. Do they differ from the simple standard error of these means based on just the data in each separate category? How or (intuitively) why (not)?
    b. Do they differ from the standard error of the estimated mean outcomes based on the simple linear mean model? How or (intuitively) why (not)?
4. Perform a test of whether the simple linear regression model fits the data, or instead the new multivariate linear model is what is needed (is the latter fitting the data significantly better?). State your conclusion.

On Table 2:

5. Consider the data described in the first line of table 2. Write down a linear regression model that will produce the reported results as a regression coefficient in the model (or a value derived thereof)

Practical Information

- Enroll to the Project-groups on Ufora. 4 people per group. A mix of people from both the statistical and the computational track is required.
- The **protocol** is due **Thursday 14 November 23:59**. Post it on the Assignment-page on Ufora.
- The **report** is due **Thursday 12 December 23:59**. Post it on the Assignment-page on Ufora
- The page limit for the body of the report (the practical assignment) is **5 pages**. You can reference some less important figures or tables in the appendix.
- The theoretical exercise should be presented in a separate document and does not contribute to the 'maximum number of pages'.
- Add an appendix with the set code that can be executed to get the material you present in the main text. The set of commands should be complete and commented, so that the purpose of each block of code is clear. Important comment: every student should contribute some of the code. Let us know who programmed what
- Take care to properly reference any source materials that you use. Add the reference list to your report. Also write who did what in a small paragraph here. This does not count in the page limit

For the students of the Continuous Data Analysis course, an **oral presentation** is scheduled on **Thursday 19 December**. A detailed schedule will be set up shortly after the project reports have been submitted. Make sure to upload your slides on Ufora on the day before the oral presentation. After the final submission, you will be asked to provide some peer evaluation of your fellow group members.

Name the document Group*Nr*_Name1_Name2_Name3_Name4

Enjoy the discovery and good luck!

References

Knaus, William A. 1995. "The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults." *Annals of Internal Medicine* 122 (3): 191. https://doi.org/10.7326/0003-4819-122-3-199502010-00007.