Home assignments for the 2020 autumn semester – PSYP14-HT20

Regression models with fixed and random effects

Introduction

In this home assignment you are going to work with (simulated) data related to perioperative pain and its psychological and hormonal predictors. In assignment 1) you will assess the added benefit of including some psychological and hormonal predictors to the already established demographic predictors of pain; in assignment 2) you will contrast the performance of the theory based model with that of a model determined by an automated model selection approach; and in assignment 3) you will build the same model on data originating from different data collection sites.

Assignment 1

In this assignment you will gradually set up a hierarchical regression model to predict postoperative pain after wisdom tooth surgery.

Research problem

The amount of pain experienced around and after surgeries are highly variable between and within individuals. In order to improve surgical pain management regimens we need to understand what influences pain around surgical procedures and predict the amount of pain an individual will experience.

Your first study in this area is related to assessing the influence of trait and state psychological measures on pain, and to see whether taking into account these variables can improve our understanding of postoperative pain.

Procedures and measures

Data file 1 is called 'home_sample_1.csv', R can read this file from:

https://tinyurl.com/ha-dataset1

For example you can save this file to an R object by running:

```
data_sample_1 = read.csv("https://tinyurl.com/ha-dataset1")
```

You have collected data from 160 adults who were scheduled to undergo surgical extraction of the third mandibular molar (wisdom tooth surgery). Patients filled out a form in the waiting room before their surgery. The form contained questions about their sex, age, and weight, and psychological questionnaires assessing anxiety, pain catastrophizing, and mindfulness (see descriptions below). You also got blood samples and saliva samples from participants in the waiting room 5 minutes before their operations to determine the serum (a component of the blood) and salivary cortisol levels of participants. Participants were contacted 5 hours after the surgery to see how much pain they were experiencing. The level of pain at that moment was recorded using a numerical rating scale using a scale of 0 to 10, where 0 means "no pain" and 10 means "worst pain I can imagine".

The State Trait Anxiety Inventory - T: measures trait anxiety on a scale of 20 to 80, higher scores mean higher anxiety. Anxiety has been found in many studies to positively correlate with the level of pain experienced. This is variable STAI_trait in the dataset.

The Pain Catastrophizing Scale measures the extent of pain catastrophizing, which is characterized by a tendency to magnify the threat value of a pain stimulus and to feel helpless in the presence of pain, as well as by a relative inability to prevent or inhibit pain-related thoughts in anticipation of, during, or following a painful event. The total score on this scale ranges from 0 to 52, higher scores mean higher catastrophizing. Pain catastrophizing is one of the well-established predictors of clinical pain. This is variable pain_cat in the dataset.

The Mindful Attention Awareness Scale (MAAS) measures dispositional mindfulness, which may be described as a tendency to turn attention to present-moment experiences in an open, non-judgemental way. The MAAS total score ranges from 1 to 6 (an average of the item scores), with higher scores representing higher dispositional mindfulness. Trait mindfulness has been theorized to serve as a protective factor against pain, as the individual would be more objective about their pain experience and tend to associate less discomfort, despair, and hopelessness to the pain-related sensations. This is variable mindfulness in the dataset.

Cortisol is a stress hormone associated with acute and chronic stress. Cortisol levels are thought to be positively associated with pain experience. Cortisol can be measured from both blood and the saliva, although, serum cortisol is often regarded in medical research as more reliably related to stress (serum is a component of the blood plasma). These are variables cortisol_serum, and cortisol_saliva in the dataset.

The dataset also contains information about the participants' weight (in kilograms), IQ, and household income (in USD). However, these variables are not theoretically linked with pain reported in the perioperative period.

Research question 1

Previous studies and meta-analyses showed that age and sex are often predictors of pain (age is negatively associated with pain, while sex is a predictor more dependent on the type of the procedure). You would like to determine the extent to which taking into account psychological and hormonal variables aside from the already used demographic variables would improve our understanding of postoperative pain.

To answer this research question you will need to conduct a hierarchical regression, building a model containing age and sex as predictors of pain (model 1), then building a new model with the predictors: age, sex, STAI, pain catastrophizing, mindfulness, and cortisol measures (model 2). Notice that the predictors used in model 1 are a subset of the predictors used in model 2. Once you are done with the two models, you will have to do model comparison to assess whether substantial new information was gained about pain in model 2 compared to model 1.

What to report

As usual, before you can interpret your model, you will need to run data and model diagnostics. First, check the variables included in model 2 (age, sex, STAI, pain catastrophizing, mindfulness, and cortisol measures as predictors, and pain as an outcome)

for coding errors, and the model itself for influential outliers (for example using Cook's distance). Furthermore, check the final model to see if the assumptions of linear regression hold true, that is, normality (of the residuals), linearity (of the relationship), homogeneity of variance (also called homoscedasticity) and that there is no excess multicollinearity ("uncorrelated predictors" in Navarro's words). If you find anything amiss during these checks, make the appropriate decision or correction and report your findings and actions in your report. Remember, if you do any changes, such as exclude cases, or exclude predictors from the model, you will have to re-run the above checks for your final data and model.

Report the results of model 1 and model 2. For both models you should report the model test statistics (R^2, F, df, and p value). Also, report the statistics describing the coefficients of the predictors in a table format (unstandardized regression coefficients and 95% confidence intervals, standardized regression coefficients (B and Beta values), and p values).

Write up the regression equation of model 2 in the form of $Y = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n$, in which you use the actual regression coefficients of your models. (b0 stands for the intercept and b_1, b_2 … b_n stand for the model coefficients for each of the predictors, and X_1, X_2, … X_n denote the predictors).

Compare the two models in terms of how much variance they explain of pain's variability in the sample. Report Akaike information criterion (AIC) for both models and the F test statistic and p value of the model comparison returned by the anova() function.

What to discuss

In your discussion briefly interpret the results of the above analyses, and indicate whether you think that anything was gained by including the psychological and hormone measures in the model.

Assignment 2

Research question 2

Let us pretend that you have published your finding from Assignemt 1 in a scientific journal. Later that year a fellow researcher's commentary on your results is published. She indicated that instead of using your theory-based approach to determine the predictors of pain, you should have simply used all available variables in a starting model and submitted it to backwards regression, letting this stepwise model selection approach select the best model. She claims that she got better adjusted R^2 with this approach on your original data than you did with your theory-based model. You decided to pit your approaches against each other. Which one is more effective in predicting pain?

First, you will have to run a backward regression to confirm her claim. She used the following variables as predictors in the initial model (before stepwise exclusion): age, sex, STAI, pain catastrophizing, mindfulness, serum cortisol, weight, IQ, household income. She excluded salivary cortisol because, as she wrote, "it was essentially identical to serum cortisol". Run a backward regression using these predictors as an initial model. Use data file 1 to run this regression (the one called 'home_sample_1.csv', the same as the one used in assignment 1). If you have excluded any cases (participants) from analysis in assignment 1 for any reason, exclude them here as well. (Before you run the actual backward regression,

you will have to re-run the data and model diagnostics, as there are new variables in the model).

Run a new regression model now only using the predictors that were retained in the end of the backward regression, and save this model in a new R object. We will refer to this model as the "backward model". Run the full regression model you arrived at in the end of assignment 1 again, and save this model in another R object. We will refer to this model as the "theory-based model". Compare the backward model and the theory-based model based on AIC (and using the anova() function if appropriate).

After this, you decide to put the two models to the test on some new data. You collected new data from another 160 participants in the same way as you did in the first study described in Assignment 1.

Data file 2 is called 'home_sample_2.csv', R can read this file from:

https://tinyurl.com/ha-dataset2

On data file 2, make predictions on pain using the regression models or equations of the backward model and the theory-based model which were "trained" on data file 1. (IMPORTANT: do not fit the regression models on data file 2 (don't re-train your models), just use the regression equations that you derived based on data file 1. These regression equations should be applied on the new data (data file 2), to predict pain.) Compare the predicted values with the actual pain ratings. Which model was able to predict the actual pain ratings in data file 2 better?

What to report

Report the characteristics of the backward model when it was fit to data file 1, by reporting the model test statistics (R^2, F, df, and p value). Also, report the statistics describing the coefficients of the predictors in this backward model in a table format (unstandardized regression coefficients and 95% confidence intervals, standardized regression coefficients (B and Beta values), and p values). Compare the initial model (the model submitted to backward regression) and the backward model and report the AIC for both models. In a similar fashion, report the comparison of the backward model and the theory-based model.

Report the prediction performance of the backward model and the theory-based model on the new data (data file 2). This can be done with several measures, for example calculate the sum of squared differences between the predicted and the actual pain values (or the sum of absolute differences) for each model.

Write up the regression equation of the backward model just like instructed in assignment 1 ($Y = \ldots$).

What to discuss

Which model performed better on data file 1? Which model performed better when predicting pain in data file 2? Why? Which model would you choose to predict pain in an actual clinical context? Why?

Assignment 3

Research question 3

Your research paper on your original study was so successful, that you managed to secure research funding for a multi-site replication study. Here your collaborators collect data in the same way you did in the original study at 20 different hospital sites. The goal of the study is to increase the generalizability of your findings. You would like to assess the model coefficients and the overall predictive efficiency of the predictors in your model.

You will need two datasets for this assignment, Data file 3 and 4. These can be downloaded from the following links:

https://tinyurl.com/ha-dataset3

https://tinyurl.com/ha-dataset4

First, build a linear mixed model on data file 3, accounting for the clustering of the data at different hospital sites. We have no reason to assume that the effects of the different predictors would be different in the different hospitals, so fit a random intercept model including the random intercept of hospital-ID, and the fixed effect predictors you used in assignment 1. Once the model is built, note the model coefficients and the confidence intervals of the coefficients for all fixed effect predictors, and compare them to the ones obtained in assignment 1.

Also, compute the variance explained by the fixed effect predictors using marginal R^2, and the variance explained by the fixed and random effect terms combined using conditional R^2. Now use the regression equation obtained on data file 3 to predict pain in data file 4. (IMPORTANT: just like in assignment 2, do not fit the regression models on data file 4 (don't re-train your models), just use the regression equation you derived based on data file 3. The regression equation should be applied on the new data (data file 4), to predict pain.) Now compute the variance explained by the model on data file 4. You can do this by using the formula we learned in class: 1-(RSS/TSS). Compare this R^2 to the marginal and conditional R^2 values computed for the model on data file 3.

Build a new linear mixed effects model on dataset 3 predicting pain. However, instead of including all predictors, you should only include the most influential predictor from the previous model. Allow for both random intercept and random slope. Now visualize the fitted regression lines for each hospital separately.

What to report

Report the model coefficients and the confidence intervals of the coefficients for each fixed effect predictor obtained on data file 3 in a table. The table should also contain the same data obtained in the final model of assignment 1.

Report the variance components for the fixed effects, the random intercept, and the residuals (from the model on data file 3). Also report the marginal R^2 and the conditional R^2 squared obtained from the model on data file 3, and the R^2 of the model computed for data file 4.

Include the graph displaying the separate fitted regression lines for the hospitals from the mixed model including only the most influential predictor in the model.

What to discuss

Compare the model coefficients and the confidence intervals observed in assignment 3 and assignment 1 and discuss what you think the differences or similarities mean. Discuss whether the R^2 obtained in data file 4 was closer to the marginal or the conditional R^2 obtained on data file 3. Explain why! Based on the graph you created, discuss whether the random intercept or the random slope model is a better fit for the data, and why.

Style and submission requirements

The number of pages is restricted to 5, A4, pages, use Times New Roman 12 pt font, 1.5 line spacing, with 2.5 cm left and right margins. Your report (including all 3 home assignments) should not be more than 5 pages, not including figures and tables. Tables and figures should not be included in the main text, but should be included at the very end of the main text on the last pages of the document, and only referenced in the main text. For example: "the fit of the predictions of the random intercept and slope models to the actual data is shown in figure 1 and 2 respectively". For everything else, use APA formatting guidelines where possible, including in the reporting of statistical results (e.g. report $p < .001$ instead of $p = 0.0000231$). You are encouraged to present correctly labelled figures and tables.

The main text should include a very brief introduction, a comprehensive result section, and a short discussion *for each assignment*. You don't have to report every table or figure obtained when doing your analysis, only those that are relevant for your report. Report everything that is listed in the 'What to report' sections below, and use the 'What to discuss' sections for guidance about what should be included in the short discussion.

Upload your R code to github, and include a link to your code in your report!

E-mail your report in word or pdf format to zoltan.kekecs@psy.lu.se, no later than 24.00hr, 4th December, 2020.

Grading guidelines

E – The student understands the basic fundaments of the analysis and can perform the analysis using R (or similar program). The report includes a correct presentation of at least one analysis. Assumptions have been addressed, variables have been checked, and eventual problems have been corrected.

D – Everything in E. The student has a more extended knowledge of the analysis, and can correctly perform an analysis using appropriate methods and strategies, and interpret the results. The report includes a more advanced analysis than E. The results are correctly discussed. Statistical and mathematical copy in the text is reported following APA guidelines.

C – Everything in D. The student has a more thorough understanding of the analyses, suitable tables and or graphs (as appropriate) are provided for correct interpretation of the results, and the R code is shared in a public GitHub repository. The report includes comprehensive analyses that are presented clearly and concisely.

B – Everything in C. The student has a deeper understanding of the analyses is able to perform correct analyses and report this in a text typical of a research report. The report is

free from errors regarding the interpretation of models and discussion of results. The report shows a good understanding of the analyses.

A – Everything in B. The student has the ability to do and report the analyses in line with published research reports. The report is formulated in an outstanding way.

If only Part 1 of the assignment is completed the highest obtainable grade is D.

If Part 1 and Part 2 of the assignment are completed (but not Part 3) the highest obtainable grade is B.

If all 3 parts of the assignment are completed it is possible (but not guaranteed) to obtain grade A, but this does depend on the quality of the work.

Individual work

Assessment is individual. It is not allowed to write assignments together with someone else. If it is obvious from your assignments, for example because there are exactly the same errors, or exactly the same sentences in the text, then all with similar reports will fail. Do not forget that we are required to report attempts to cheat. Make sure that none of your classmates read your text before submitting it. It is never permissible to include text directly from other sources, including the Internet (i.e., plagiarism and citing material without crediting the author is not permissible), if this is done, the assignment will fail and it will be reported on.

It is OK to discuss different statistical approaches to solve the assignments with fellow classmates, but it is not OK to share materials like R code, written assignment reports wholly or partially, or figures/tables, and it is absolutely not OK to copy any of these materials from a fellow student's work. Every student should produce their report 100% on their own (writing the code, running the code, writing the report, producing the tables and figures).