Home assignment ("take-home exam"): Stats III Mixed-Models 2016

Deadline: April 4, 2016, 1 minute before midnight; upload your materials to the respective turnitin assignment on BlackBoard.
(If, but only if, you encounter technical problems, send your materials via email to Bernd Figner: b.figner@psych.ru.nl.)

First of all: As you will see, this document is rather long as it contains a lot of explanations and concrete instructions what you have to do (and also sometimes what you should NOT do). Please read it very carefully. And you should also read it several times throughout the time that you're working on this assignment, to make sure you don't forget or overlook anything!

(A) What you have to hand in

1. Word or pdf document that (i) describes the data analysis, results, and brief conclusions/interpretation of the results, thereby (ii) answers the research questions below, and (iii) includes figures (don't forget to include figure captions).
Maximum length: Try to be no longer than 10 A4 pages maximally (not counting the title page and references).

2. Commented R script
The script has to be commented in enough detail so that a person with R knowledge but no intimate knowledge of the data set/task can understand what you did in this script.
Also: The script has to work as it is, i.e., besides adjusting the path to the working directory, the R script needs to be able to run without adjustments.

NOTE: Everybody has to hand in their own work! If two (or more) people hand in the same work, they both will fail the exam.

Unless you have specific *clarification* questions, do NOT post your questions related to doing the take-home exam on BlackBoard. For clarification questions, I added a forum to the BlackBoard Discussion page.

(B) General Instructions

Detailed instructions and questions are below, but the more general instruction is that you will do several mixed-models, including all the preparatory steps (such as loading the data, merging data frames, preparing the variables, etc etc), diagnostics before running the model(s), running the model(s), diagnostics after running the model(s), obtain *p* values, create figures, and report your models and results. I.e., all the usual steps that we covered in the course.
You'll report the model and the results as if writing the results section for a paper that you would submit to a scientific journal. Follow APA style (i.e., also include a brief reference section). There is no need for a separate introduction or methods section because you all use the same data set (which happens to come from a study I know...).

The **goal** is that you demonstrate that you can do all the steps involved in a mixed-effects models analysis and report the results in text and figure. Thus, the goal is less that you present some specific significant results, but that you demonstrate your understanding of both the practical and theoretical aspects, and can evaluate the various steps involved in mixed modeling; this also means, for example, that you are able comment on the model residuals

and make conclusions from this about what method to get *p* values might be a good idea to use.

BTW, if some of your analysis steps use *bootstrapping*, you can use relatively low numbers of simulations (say 500 or so; not thousands or tens of thousands), to avoid that your models run for hours and hours.

(C) Specific instructions and questions that you need to answer

You will work with a data set that comes from the following study:
(This is not really relevant for your take-home assignment, but if you're curious: these data are mostly real data from a real study, although some aspects I have changed somewhat.)

Each participant was presented with 12 investment scenarios, across which the following variables were varied according to a full factorial design:

- Sustainability of the investment: A continuous variable that ranges from 1 (low sustainability) to 4 (high sustainability)
- Expected payoff: A continuous variable that ranges from 1 (1 Mio) to 3 (3 Mio)

In each of these scenarios, participants were asked to rate their willingness to invest in that specific investment opportunity by indicating their willingness on a continuous visual analogue rating scale ranging from -50 (not willing at all) to +50 (absolutely willing).

Data were collected in a country far away and participants were recruited from 3 groups (i.e., 3 different major programs): business students, ecology students, and physics students.

I.e., in this study, we were interested whether and how willingness to invest in different business opportunities is influenced both by considerations about sustainability and by considerations about expected profits. Further, we were interested whether these things differ between different study majors.

More concretely, your task is it to investigate and answer the following research questions:

- Does the willingness to invest differ significantly across majors?
- What is the relationship between sustainability and willingness to invest:
    - Is there evidence for a significant linear relationship?
    - Is there evidence for a significant quadratic relationship?
    - Is there evidence that the linear relationship differs across majors?
    - Is there evidence that the quadratic relationship differs across majors?
- What is the relationship between expected profit and willingness to invest:
    - Is there evidence for a significant linear relationship?
    - Is there evidence for a significant quadratic relationship?
    - Is there evidence that the linear relationship differs across majors?
    - Is there evidence that the quadratic relationship differs across majors?
- Is there evidence that the linear relationship between sustainability and willingness to invest is moderated by the linear predictor of expected profit?
    - Is there evidence that this (potential) moderation effect differs across majors?

Important: Specify your model so that you can answer all of these research questions. But do not make your model unnecessarily complex (e.g., do not include interaction terms for which there is no corresponding research question).

Follow-up models
If you find a significant effect (i.e., $p < .05$) involving the factor "major" (main effect or interaction), you need to conduct follow-up models to understand the main effect and/or interaction(s):

- If you find a significant *main effect* of major, you need to do follow-up models to do pairwise comparisons, to be able to answer which majors differ from each other and which don't.
- If you find a significant *interaction* involving major, you also need to do follow-up models:
    - The first set of follow-up models is to do pairwise comparisons to be able to say which majors do and don't differ from each other in the relationship between the continuous predictor(s) and the dependent variable.
    - After that, you need to do further follow-up models to investigate each major separately, to be able to say for each major whether there is or is not a significant relationship between the continuous predictor(s) and the dependent variable.

Please note: Do NOT use a post-hoc function like lsmeans or glht to do these analyses! The task is to conduct the follow-up models as described above.

General comments and tips:

- You obtain 2 data sets, not just one (one is the main data file; the other file contains the information regarding the majors). You will need to merge these data files.
- For factors, use sum-to-zero coding throughout.
- Use poly() to create the predictors necessary to investigate the questions about linear and quadratic relationships.
- To determine *p* values, use Type 3 tests, and use one of the different approaches to test "effects" that we discussed in class (except Wald Chisquare tests: you are not allowed to use those). Please justify why you chose the method you chose to use. Report exact *p* values whenever possible (except when the *p* value is smaller than .001; in those cases write $p < .001$).
- When doing your analysis, specify the random effects as it is appropriate, following the advice by Barr et al. (2013) and Barr (2013) to "keep it maximal" whenever possible.
- If you encounter convergence problems, follow the steps discussed in Bernd's lecture and the advice by Barr et al. (2013). HINT: poly() can sometimes lead to very small numbers in the variable, and it sometimes helps with convergence if they are not that small...
- Make sure that the model(s) you report **did** converge (i.e., do not report the results of models that did not converge).

Your report should include at a minimum the following things:

Basic things

- Briefly describe the sample (how many participants in each major).
- Describe the model setup (including if you did some preparatory steps such as transforming the data, how you created your predictors (centered, standardized, something else, ...), what kind of contrast settings you used, etc).

- Include the syntax of the main model that you ran (in a usual scientific paper you wouldn't do that, but here please do it).
- If you did some model-selection procedure or had to adjust your model (e.g., due to convergence problems), explain the rationale and the involved steps.

Figures

- At least 3 plots to understand the raw data (i.e., plots that you typically create **before** you run any model), for example a figure that shows the distribution of the DV (for example separately for some levels in a factor), and/or a figure that shows individual differences and variability among participants, and/or of course figures that are relevant to some (or all) of your research questions.
- At least 4 model diagnostic plots (create them only for your main model; you don't have to create them for your follow-up models), namely: distribution of the scaled residuals; qqPlot of the scaled residuals; fitted vs. residuals with a smoothed line; scatterplot of the fitted versus observed data. Discuss the diagnostic plots in the respective figure caption. In addition, compute the proportion of scaled residuals for the usual +/-2, 2.5, 3 ranges, and discuss them in the text.
- Result-related figures
    - Show plots that visualize the significant main effects and/or interactions. You want to show each significant effect/interaction in a figure (but you may show several effects combined in 1 figure, if you prefer, working, e.g., with multipanel plots).
    - What type of figure you use for which result is your choice. But to show that you are able to create different types of plots, among your results figures should be at least each one of the following (note: Note: for an article in a scientific journal, we wouldn't typically mix those types of figures, but the goal is here that you show off your skills...)
        - 1 figure visualizing a significant main effect and/or interaction using the raw data (if possible and appropriate, show not only the means but also the SEs or CIs)
        - 1 figure visualizing one (or more) significant (and/or non-significant) main effect(s) and/or interaction(s) by showing the estimated fixed effect coefficient(s)
        - 1 figure visualizing a significant main effect and/or interaction by showing the model-based means (and indicators of the uncertainty such as CIs or SEs)

Results

- Include both the statistical significances (including, where appropriate, the coefficients, their standard errors, test statistics, *p* values). Include also some descriptives, for example, when analyzing the effect of a factor, report the means (in addition, you can refer to a figure(s) if there is one that shows these means).
- Include brief statements that explain the results ("This result shows that willingness to invest increased as a function of .....").

Best of luck!