Assignment for DAT246/DIT246

Professor Richard Torkar

September 20, 2021

> This is an individual assignment, i.e., you must work alone. The assignment will be graded pass/fail, and to pass the course you need to pass this assignment.
>
> There are three deadlines for grading: Oct. 15, Nov. 12, and Dec. 17. After these three occasions we welcome new submissions during the autumn 2022, if one has not passed the assignment after the final deadline (Dec. 17).
>
> If you need help with the assignment then please talk to Richard during the open door sessions.

Note: Three chances to pass, with feedback given once per deadline.

THE DATA WE WILL USE in this analysis has partly been published before.[1] The data is from an experiment where 70 subjects participated. Of the 70, 46 subjects were categorized as less experienced (LE) and 24 were categorized as more experienced (ME). The experiment evaluated two software testing techniques, i.e., a new technique (NT) and an old technique (OT), used a small, noncritical system as the software under test, and had a 2×2 design to avoid learning bias. The effectiveness of each technique was measured through true positives (`tp`), i.e., the number of found faults classified as true faults, with more being better. Below we see the first rows of the data file that we will use (two observations/subject adds up to 140 rows):

[1] W. Afzal, A. N. Ghazi, J. Itkonen, R. Torkar, A. Andrews, and K. Bhatti. An experiment on the effectiveness and efficiency of exploratory testing. *Empirical Software Engineering*, 20(3): 844–878, 2015. ISSN 1573-7616. DOI: 10.1007/s10664-014-9301-4

```
> head(d)
  subject category technique tp
1       1       LE        NT  5
2       1       LE        OT  6
3       2       LE        NT  3
4       2       LE        OT  3
5       3       LE        NT  7
6       3       LE        OT  3
```

We would like to understand if the new technique (NT) is better, based on the given data, than the old technique (OT), and if there is a difference between less and more experienced subjects (LE/ME), using *Bayesian data analysis* as taught in the course. This way we will be able to decide if the technique should be used by the company and, hence, if there is a need to take experience levels into consideration. For further information regarding the experiment we refer you to the publication previously cited.

Downloading data

Instructions for installing all needed software can be found in Exercise 1. You can then download the data directly from GitHub.[2]

[2] https://raw.githubusercontent.com/torkar/dat321/master/data_autumn2020.csv

THE REPORT SHALL CONSIST OF not more than *four* (4) pages, font size *11*, including plots. The report shall be submitted as a PDF. The report shall be made individually, and submitted as such, i.e., there must be *one name only on the report*.[3]

[3] Not following these instructions will lead to an immediate fail

We wholeheartedly recommend students to use Overleaf for the writing of reports.[4] There is a LaTeX template that you can use for your report.[5] It includes examples of the most common things you might want to use in a report for any assignment. The link is a read-only version of the original template. To be able to actually write your report, you *might* have to create a copy of the project (`Menu` → `Copy Project`, in Overleaf). You can then edit the copy to your heart's content.

[4] https://www.overleaf.com

[5] https://www.overleaf.com/read/cyxdmrncrzpj

If you don't want to use Overleaf, you can instead download the source of the project (`Menu` → `Download` → `Source`).

For submissions, we will use the built in assignment system of Canvas. You should have an assignment entry in the content list, where you can upload your PDF.

IN ADDITION, we expect the following to be in the report at a minimum:

1. Description of the data, i.e., descriptive statistics, if there is a need to standardize/normalize, etc.
2. A defense of your likelihood(s). We assume you will develop at least two models with, perhaps, different likelihoods. Please use the same math notation to describe your model(s) as they do in the course book, e.g.,[6]

$$\begin{aligned}
\mathrm{L}_i &\sim \mathrm{Binomial}(n_i, p_i) \\
\mathrm{logit}(p_i) &= \alpha_{\mathrm{SUBJECT}[i]} + (\beta_P + \beta_{PC} C_i) P_i \\
\alpha_{\mathrm{SUBJECT}} &\sim \mathrm{Normal}(0, 10) \\
\beta_P &\sim \mathrm{Normal}(0, 10) \\
\beta_{PC} &\sim \mathrm{Normal}(0, 10)
\end{aligned}$$

[6] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2nd edition, 2020

This is only an example **not directly applicable** to this assignment!

3. A discussion regarding the priors you have chosen for your 'final' model. Defend them and show what happens if you change priors!
4. Results from running your 'final' model with `ulam()`, comparing it with other model(s) using WAIC or LOO,[7] and reason about what the results means.

[7] The rethinking package has a `compare()` function for this

5. (You can try to specify a model with an interaction effect. But it is **not mandatory** to pass the assignment.)

6. Adding a DAG is always nice (use the `ggdag` package in R). If you can explain direct and indirect causal effects without one then sure.[8]

7. Presentation of diagnostics from running Stan on the 'final' model, e.g., caterpillar traces (or trankplots),[9] $\hat{R}$ values, and/or effective sample size. There's no reason to show traceplots that take up a page!

8. Interpretation of what the results mean from a practical point of view, i.e, which technique is better, does experience influence the results, and how does the analysis support your argument?

Depending on the strength of your arguments, the wise use of visualization, and the style and clarity of your language, a pass grade can be given. A pass will not be possible unless one has appropriately answered items #1–4, and 6–8 above.

[8] https://ggdag.malco.io

[9] http://mc-stan.org/assets/img/bayesplot/mcmc_trace-rstan.png