Assignment 4

GV903 Advanced Methods – 2020/2021
University of Essex, Department of Government

Instructions

This assignment is due on Wednesday, 31 March at 9:45 am. Please submit on FASER. Instructions can be found in the module handbook on Moodle.

Please create a new document with your answers. Use LaTeX in conjunction with `knitr` via RStudio for document creation, and submit both a PDF file (the compiled document) and the source files necessary for compilation (e. g., with the extensions `.Rnw`, `.bib` etc). Up to ten points are given for proper use of technology and proper formatting, as part of the maximum of 100 attainable points.

Each task is worth a certain number of points, which is denoted below the tasks. The points add up to 100 (including the points for use of technology as described above) and determine your final mark for this assignment. The final mark you earn for this assignment is given by the equation

$$m = t + \sum_{i=1}^{n} p_i, \tag{1}$$

where p_i denotes the number of points you earn for Task i, n is the number of tasks in this assignment, and t is the number of points you earn for use of technology and formatting. All answers are evaluated on three criteria: how clearly the results are presented; how correct the results are; and how elegant, computationally efficient, or sophisticated the solution is (where applicable). Good luck!

Dataset: Publications by German political scientists, 2009–2013

Leifeld et al. (2017) analysed the co-authorship network among all 1,622 political scientists with a PhD at German universities and research institutes in 2014 for the years 2009 to 2013. In this assignment, you will re-analyse this dataset from a different angle. The file `plosone.csv` is provided with this assignment. It is an edited version of the original replication dataset, which can be found at `https://doi.org/10.7910/DVN/ZBRQU2`. The `plosone.csv` dataset contains nine variables for the 1,622 academics:

name Name of the academic.

uni University of research institute of the academic.

status Whether the academic is a professor or not.

birthyear The year of birth of the academic, as mentioned on the academic's CV or website.

phd The year in which the academic received a PhD.

gender The gender of the academic.

chair The name of the chair or research group in which the academic worked.

publications The total number of publications of the academic between 2009 and 2013.

articles The number of peer-reviewed journal articles of the academic between 2009 and 2013.

There are some duplicate entries in this dataset that did not matter much for the network analysis in the published article. While they may matter in minor ways for the analysis conducted here, we will ignore the issue of duplicate entries in order to keep things simple.

In this assignment, you will model the number of publications and the number of journal articles of each academic as a function of seniority status, gender, birth year, and PhD year. However, the dataset has some missing data, and this problem needs to be tackled first. The number of publications or articles are also count data, and you will need to select an appropriate model and interpret the results accordingly. Finally, as all of this involves several small decisions, you will also want to use matching to reduce model dependence. Each task below will give specific instructions on these steps.

Task 1: Modelling complete cases

In this task, you will ignore the problem with missing data and focus on complete cases. Your task is to model 1) the number of publications and 2) the number of peer-reviewed journal articles. Your independent variables are seniority status, gender, birth year, and PhD year. Your main research question of interest is whether gender makes a difference for explaining publication and article counts. As the two dependent variables are counts, you will need to estimate models that can deal with count data. But there are several candidate models available. Estimate these candidate models, and choose the most appropriate specification for interpreting the results. Diagnose possible problems with overdispersion, both using graphical tools and a statistical test. Show the results of the candidate models you have considered side by side in a `texreg` table that is embedded into the document using `knitr`. Are there any differences in results? Describe your steps and the results, and motivate your model choice, in up to 300 words. [20 points]

Task 2: Violated assumption

Which assumption is violated in this regression model? Describe the potential problem and its consequences in 150-300 words. (Just to be clear: While we would generally care about this in an empirical application, we will ignore the problem here in this assignment, apart from this task.) [6 points]

Task 3: Visualising missing data

Show several diagrams of the missingness patterns in the dataset. Interpret/describe them briefly. Discuss/speculate if the missingness is presumably MCAR, MAR, or MNAR and what that means for the choice of imputation measure. [6 points]

Task 4: Single imputation with chained equations

Use chained equations, also known as iterative regression, to impute the missing data in all variables. Use single imputation in this task. You would normally use the `mice` package for chained equations, but the package does not support count data very well. Write up your own chained equations single imputation algorithm, and apply it to the dataset to impute all missing values. You can ignore the nominal variables (the name, university, and chair name) in the imputation here and in later tasks. You will require a mix of count and non-count models as part of your implementation. When you use count models, make sure you round the predictions to integer counts each time because count models don't deal well with decimal places. You can run the algorithm for a fixed duration of 100 iterations. Plot the trace of two model coefficients of your choice and two imputed values of your choice. Conduct a few plausibility checks by discussing a few data points. Describe your solution in up to 400 words or less, including your implementation of the algorithm, the trace plots, and plausibility checks. [14 points]

Task 5: Visualisation with `tidyverse` functions

Using the single imputation from Task 4, show the distribution of the imputed values and the distribution of the observed values as density curves, for each of the four variables affected by missingness; are the distributions similar for the imputed and the observed values? Try to use exclusively the `tidyverse` to do this (and possibly its extension packages by third parties, such as `gridExtra` if necessary), including data preparation and plotting, by making use of the pipe (`%>%`) and `ggplot2` and oher functionality from the tidyverse. The only function you are allowed to use from outside the tidyverse and its extensions is `c`. Ideally, create two diagrams with two panels or facets each. The first diagram should show the count variables: articles and publications in the two panels. The second diagram should show the year variables: birth year and PhD year in the two panels. In each panel, show a density curve for the observed values and another density curve in a different colour for the imputed values. You should limit the range on the x axes to reasonable parts of the distribution to make the main highest-density areas of the distribution most visible (i. e,. you can discard the tails of the distribution while plotting to make the plots easier to understand). If you have not been able to complete Task 4, use only the observed values in this task, and a few points will be deducted. Summarise your findings briefly. [7 points]

Task 6: Re-estimation with imputed data

Use the imputed dataset from Task 4 to re-estimate the candidate models from Task 1 and show the results side by side again in a table. Compare your new results with the old results: Do the results and your conclusions change? Can we trust the results? Discuss briefly. [5 points]

Task 7: Multiple imputation with chained equations

In Task 4, you wrote up an algorithm for chained equations with a mix of count and non-count data. But it was only single imputation. Now take that algorithm and turn it into a multiple imputation algorithm. If you have a sufficiently fast computer, aim to use 200 iterations this time. If not, some lower number is fine, perhaps 50—the results should not differ much. Hint 1: Calculating predicted values was covered in the first semester and again in Assignment 3. Hint 2: The inverse link function of both Poisson and negative binomial models is $f(x) = e^x$.

Once you have created multiple imputed datasets, apply the model you chose in Tasks 1 and 6 to the imputed datasets and combine the results using Rubin's rules. You can use `R` packages for combining the results and do not have to apply Rubin's rules manually here. Do your results change compared to the single imputation? Interpret and describe in up to 150 words. [12 points]

Task 8: Matching

Let's say your main variable of interest is gender: What is the gender premium for the number of journal articles—how many more or fewer publications do men get compared to women? As you have seen by now, there are several model choices available, and the specification may matter for the results. It may be a good idea to make our modelling as immune as possible to such decisions by using matching. Use the `MatchIt` package to create a dataset that is as balanced as possible, and then compare the article counts for men and women. What do you find? Does it differ in any way from your model? Please use your single-imputed dataset from Task 4 for this purpose. (Using an imputed dataset with chained equations kind of defeats the purpose of matching because lots of modelling assumptions went into the imputation, but we will ignore this here for training purposes.) If you did not manage to complete Task 4, you can use the complete-case dataset instead. Hint 1: There are fewer women than men in the dataset, so you should use female as the treatment; otherwise you may run out of suitable matches. Hint 2: If you want to complete this matching task before trying your luck with the imputation, you can use the complete observations from the `plosone.csv` dataset first and then later update your analysis once you have the imputed dataset. [20 points]

References

Leifeld, P., Wankmüller, S., Berger, V. T. Z., Ingold, K., and Steiner, C. (2017). Collaboration patterns in the German political science co-authorship network. *PloS ONE*, 12(4):e0174671. `http://dx.doi.org/10.1371/journal.pone.0174671`.