Analiza danych niepewnych: Introduction to fuzzy statistics

Final project

Deadline: 6th of June, 2025

The file `data_caseStudy.dat` contains a dataset from a large-scale survey conducted in Flanders in 2021, which investigated four main factors related to sexual intimacy [7]. For this project, we extracted a smaller subsample of participants and focused on the following variables only:

- age (in years, `age`)
- relationship duration (in years, `rel_length`)
- desire (composite score, `sex_desire`)
- partner's gender (binary, `gender_partner`)
- partner responsiveness (composite score, `respo_partner`)
- perceived intimacy (self-reported, `intimacy`)

The analytical sample consisted of $n = 318$ participants, including 232 women (mean age: 34.15; SD: 11.95) and 95 men (mean age: 31.07; SD: 9.58). The mean relationship duration was 7.74 years (SD: 8.91 years) for women and 7.71 years (SD: 8.33 years) for men. Since `intimacy` is the outcome variable in this study, the original Likert-scale ratings were fuzzified using the fuzzy-IRTree methodology [1], resulting in a triangular fuzzy variable defined over the $(1, 5)$ range.[1]

Import the file into R and respond to the following questions:

[1]The fuzzy variable is represented using the suffixes `_lb` (left bound), `_ub` (right bound), and `_m` (mode).

1. Create a new variable containing the defuzzification of the `intimacy` variable using Delgado's Expected Value (EV) method [3], as implemented in the `FuzzyNumbers` package.

2. Create a new variable measuring the fuzziness of the `intimacy` variable. You may use Delgado's Ambiguity measure [3], available in the `FuzzyNumbers` package.

3. Under a 5-Fold Cross-Validation scheme ($K = 5$), evaluate the predictive performance of the following regression models, using all available predictors in the dataset:
   (a) A multiple linear regression model to predict the defuzzified outcome (see Question 1).
   (b) A multiple linear regression model allowing for heteroscedasticity, by incorporating the fuzziness of the response variable (see Question 2) as the `weight` in the fitting procedure.
   (c) A multiple possibilistic linear regression model for the triangular fuzzy response. Use the `fuzzyreg` package.[2]
   (d) A multiple fuzzy least squares regression model [5], under the interactive assumption between modes and left/right spreads.
   (e) A multiple linear regression model fitted via fuzzy maximum likelihood estimation [4].

4. For each of the above models, compute the following prediction error measures:
   (i) Root Mean Square Error (RMSE) between observed defuzzified $\bar{\mathbf{y}}_{\text{obs}}$ and predicted defuzzified values $\hat{\bar{\mathbf{y}}}$:
   $$\text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\bar{y}_{\text{obs}_i} - \hat{\bar{y}}_i \right)^2 \right)^{1/2}$$
   (ii) Mean Absolute Error (MAE) between observed defuzzified $\bar{\mathbf{y}}_{\text{obs}}$ and predicted defuzzified values $\hat{\bar{\mathbf{y}}}$:
   $$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| \bar{y}_{\text{obs}_i} - \hat{\bar{y}}_i \right|$$
   (iii) Bertoluzza's distance [2] between observed $\tilde{\mathbf{y}}_{\text{obs}}$ and predicted fuzzy numbers $\hat{\tilde{\mathbf{y}}}_{\text{obs}}$:
   $$D_\xi^\lambda(\hat{\tilde{y}}, \tilde{y}) = \left(\int_0^1 \left(\text{mid } \hat{\tilde{y}}(\alpha) - \text{mid } \tilde{y}(\alpha) \right)^2 + \xi \left(\text{spr } \hat{\tilde{y}}(\alpha) - \text{spr } \tilde{y}(\alpha) \right)^2 \, d\lambda(\alpha) \right)^{1/2}$$
   Here, mid $\tilde{z}$ denotes the midpoint of the α-cut interval of fuzzy set $\tilde{z}$, and spr $\tilde{z}$ is its length. Parameter $\xi > 0$ (e.g., $\xi = 1/3$) is a weight, and λ is a weighting function over α (commonly, the identity function). You may use the implementation available in the `FuzzyResampling` package [6].

5. Identify the model with the lowest prediction error. Comment on the results, with particular attention to comparisons among structurally similar methods.

Note that while RMSE and MAE are computed on defuzzified response values, Bertoluzza's distance requires fuzzy sets as input. For regression methods that do not yield fuzzy outputs, this distance can either be omitted or computed using *degenerate* fuzzy sets as predictions.[3]

[2]Note: The fuzzy response variable should be made symmetric, as this package does not support asymmetric fuzzy variables.

[3]In practice, this can be done by passing the mode value for all parameters required by the R function `BertoluzzaDistance(...)`, thus representing the predicted response as a crisp (non-fuzzy) value.

References

[1] Antonio Calcagnì, Niccolò Cao, Enrico Rubaltelli, and Luigi Lombardi. A psychometric modeling approach to fuzzy rating data. *Fuzzy Sets and Systems*, 447:76–99, 2022.

[2] María Rosa Casals Varela, Norberto Octavio Corral Blanco, María Ángeles Gil Álvarez, María Teresa López García, María Asunción Lubiano Gómez, Manuel Francisco Montenegro Hermida, María Gloria Naval Alegre, Antonia Josefina Salas Riesgo, et al. Bertoluzza et al.'s metric as a basis for analyzing fuzzy data. *METRON: International Journal of Statistics, 71 (3)*, 2013.

[3] Miguel Delgado, Maria-Amparo Vila, and William Voxman. On a canonical representation of fuzzy numbers. *Fuzzy sets and systems*, 93(1):125–135, 1998.

[4] Thierry Denœux. Maximum likelihood estimation from fuzzy data using the em algorithm. *Fuzzy sets and systems*, 183(1):72–91, 2011.

[5] Pierpaolo D'Urso. Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. *Computational Statistics & Data Analysis*, 42(1-2):47–72, 2003.

[6] Maciej Romaniuk and Przemyslaw Grzegorzewski. Resampling fuzzy numbers with statistical applications: FuzzyResampling package. *The R Journal*, 15(1):271–283, 2023.

[7] Jacques Van Lankveld, Nele Jacobs, Viviane Thewissen, Marieke Dewitte, and Peter Verboon. The associations of intimacy and sexuality in daily life: Temporal dynamics and gender effects within romantic relationships. *Journal of social and personal relationships*, 35(4):557–576, 2018.