Aalto-yliopisto
Perustieteiden
korkeakoulu

A Robust Optimisation Approach for Stable Linear Regression (topic presentation)

Veera Wilkki

6.3.2023

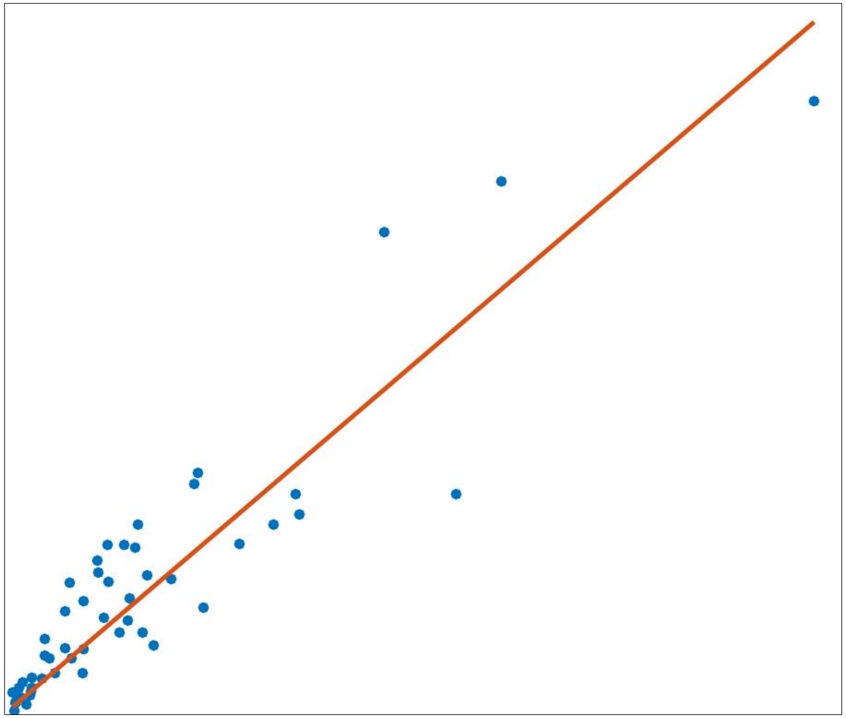
Instructor: *Paula Weller*

Supervisor: *Fabricio Oliveira*

Systeemianalyysin
laboratorio

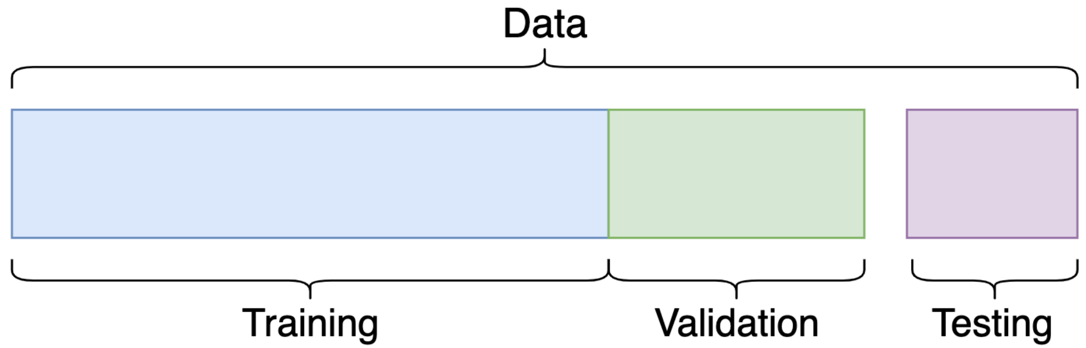
Background: training linear regression models

- Linear regression: method used to model the linear relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data



Background: training linear regression models

- Usually trained by assigning data to training and validation sets **randomly**
    1. Choose a random subset as the testing set
    2. Split the rest of the data into a training set and a validation set
    3. Test the final accuracy of the model using the held out testing set



Background: training linear regression models

- This randomised approach has some issues: models can vary significantly based on the choice of training/validation splits
- This causes problems of interpretability and accuracy
- A method called k-fold cross validation is often used to avoid this to a certain extent

Training linear regression models using robust optimisation

- Instead of splitting data randomly, this step can be integrated into the optimisation problem directly as presented in Bertsimas and Paskov (2020):

$$\min_{\beta} \max_{z \in \mathcal{Z}} \sum_{i=1}^{n} z_i |y_i - x_i^T \beta| + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{with} \quad \mathcal{Z} = \left\{ z : \sum_{i=1}^{n} z_i = k, \quad z_i \in \{0, 1\} \right\}$$

where z_i is an indicator variable, indicating which point (x_i, y_i) belongs to the training set and which to the validation set, λ is a regularisation parameter, $\Gamma(\cdot)$ is the regularisation function and k represents the desired proportion between the size of the training and validation sets

Training linear regression models using robust optimisation

- By taking the linear optimisation dual of the inner maximisation problem, we get this final optimisation problem:

$$\min_{\beta,\theta,u_i} k\theta + \sum_{i=1}^{n} u_i + \lambda \sum_{i=1}^{p} \Gamma(\beta_i)$$

$$\text{subject to} \quad \theta + u_i \geq y_i - x_i^T\beta, \quad \theta + u_i \geq -(y_i - x_i^T\beta), \quad u_i \geq 0.$$

Training linear regression models using robust optimisation

- The advantages of this optimisation approach:
  - Better performance in terms of prediction error
  - More stable
  - Allows the recovery of true support
  - Identification of the “hardest subpopulation”
- The limitations of this approach:
  - The regularisation parameters cannot be set optimally but need to be tuned by scanning through prefixed values
  - Slightly slower than the randomised approach

Aims

- Goal is to create a function that performs this robust optimisation approach to training linear models using Julia
- The function will be tested on a number of data sets acquired from the internet, e.g. UCI Machine Learning Repository and Kaggle
- The results will be examined and compared to the randomisation approach

Schedule

- 10/2022 Getting the topic and starting the project
- 3/2023 Topic presentation
- 4/2023 Finishing the programming part
- 5/2023 Writing
- 6/2023 Presenting the results in the seminar

References

Bertsimas, D., & Paskov, I. (2020). Stable regression: On the power of optimization over randomization in training regression problems. *The Journal of Machine Learning Research*, *21*(1), 9374-9398.