9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Rashomon Bound Regressions: a Reproducible and Explorable Method for Regression Problem

Keywords: Reproducibility crisis, Pre-registration, Rashomon set, Model reliance, Sociology

Extended Abstract[1]

Social sciences are experiencing a reproducibility crisis [1]. One reason for the reproducibility crisis is the Questionable Research Practice (QRP). Pre-registration limits researchers' freedom and prevents QRP such as p-hacking and harking. While pre-registration may be effective in studies that use experimental data or randomized controlled trials, these assumptions are not necessarily valid in fields such as sociology. Instead, regression analysis is commonly used to take advantage of heterogeneity in data. When applying regression analysis to observation data, problems such as multicollinearity arise, and the pre-registration occurs as the second type of error. In addition, research involving human subjects is prone to include noise. Therefore, it is necessary to search for data within an appropriate range. However, excessive data exploration reduces the reproducibility of the analysis. In this paper, we propose the Rashomon bound to solve the above trade-off between exploration and reproducibility in sociology.

Rashomon bound is based on the Rashomon set. Given a benchmark model f^*, $\varepsilon > 0$, and a class of models $\mathscr{F}$, the Rashomon set $\mathscr{R} \in \mathscr{F}$ can be defined as follows

$$\mathscr{R}(\varepsilon, f^*, \mathscr{F}) = \{f \in \mathscr{F} \mid L(f) \leq (1+\varepsilon)L(f^*)\}. \tag{1}$$

As is clear from the above definition, the Rashomon set represents a set of good performance models. Rashomon sets can be defined for various types of models, but in sociology, linear models are a common choice. In this paper, we focus on the linear model, especially for ridge regression, where overfitting is less likely to occur and is easy to analyze.

Next, we define Model Reliance (MR) [2] for jth variable mr_j as follows

$$mr_j(f) = L\left(f; [X_{\setminus j}, \bar{X}_j], Y\right) - L(f; X, Y) \tag{2}$$

where L represents the loss function and $\bar{X}$ refers to a random variable with the identical distribution as X but is sampled independently of X. These definitions were in line with [2]. From now on, we will show a more detailed analysis and robustness to noise.

First, ridge regression is a model that minimizes the following objective function where the model coefficients are β, and the ridge parameter is λ. Moreover, going forward, we will use the notations for the model f and its coefficients β interchangeably and consider β as the D-dimensional vector.

$$L_{ridge}(\beta) = \mathbb{E}\left[Y^2\right] - 2\mathbb{E}\left[YX^T\right]\beta + \beta^T\mathbb{E}\left[XX^T + \lambda I\right]\beta \tag{3}$$

Fortunately, since $L_{ridge}(f)$ is a convex function, the optimal solution β^* can be easily computed, and we can treat it as a benchmark model f^*. Thus, the Rashomon set of ridge models can be defined as follows

$$\mathbb{E}\left[Y^2\right] - 2\mathbb{E}\left[YX^T\right]\beta + \beta^T\mathbb{E}\left[XX^T + \lambda I\right]\beta \leq (1+\varepsilon)\, L_{ridge}(\beta^*). \tag{4}$$

[1] Proof, whole result, and source code will be added to the full paper or arXiv version.

$\mathbb{E}\left[XX^T + \lambda I\right]$ in (4) can be denoted by $\mathbb{E}\left[XX^T\right] + \lambda I$, and $\mathbb{E}\left[XX^T\right]$ and λI are obviously symmetric matrices. Moreover, $\mathbb{E}\left[XX^T\right]$ is always positive semi-definite, and it is easy to prove that $\mathbb{E}\left[XX^T\right]$ is positive definite in almost all data analysis settings, and $\mathbb{E}\left[XX^T + \lambda I\right]$ is also a positive definite matrix since it is represented by the sum of two positive definite matrices. Thus, if we ignore the second term on the right side of (4), the bound of (4) is represented as the surface of a D-dimensional ellipse in dimension D modified by the rotation matrix, and the set of β for which the LHS is smaller than the RHS is represented as the inner set of this rotated ellipse. Also, when the second term is considered, Bound can be represented as an affine transformation of the D-dimensional ellipse surface. In this study, from the viewpoint of clarity of analysis, we propose a method of observing elliptical surfaces, i.e., cases in which the equation hold, and call this surface Rashomon bound.

Our algorithm is extremely clear. First, we sample a D-dimension vector x from the normal distribution and convert its norm to 1. This allows us to sample points uniformly from the $(D-1)$-sphere [3]. Then, each point can be transferred to the space (which is beta space) by an Affine transform derived from (4). With these operations, we can obtain the set of Rashomon bounds and, moreover, we can examine the model reliance for each point (each model on Rashomon bound).

We have conducted two experiments to demonstrate the effectiveness of the proposed method, one on toy data with $D = 2$ and the other on real-world data, Boston Housing dataset. For each experiment, we prepared several datasets with noise added to y (X was left unchanged). For these data, we compared simple ridge regression with an analysis using the Rashomon bound.

Fig. 1 shows the result for $D = 2$ toy datasets. As shown in Fig. 1, when the ridge regression is simply applied, both the absolute values of the coefficients and the model reliance vary significantly according to the noise, and almost no pattern can be observed. In contrast, the Rashomon bound shows a common trend for all data, even though the coefficients and model reliance scales may differ. Fig 2 shows the result for the Boston Housing dataset. Consistent with the previous results, the results for the Rashomon bound are more robust to noise than in the case of simple ridge regression. It can also be seen that our results in the Rashomon bound for both cases sufficiently represent the characteristics of the data. For example, we can see the tendency for certain β to be not essential regardless of the value of other β and the trade-off between specific pairs of *beta*. In addition, such holistic views prevent the researcher from presenting only the results that he or she desires and increases the explainability and reproducibility of the results, and also make it possible to explore the dataset features.

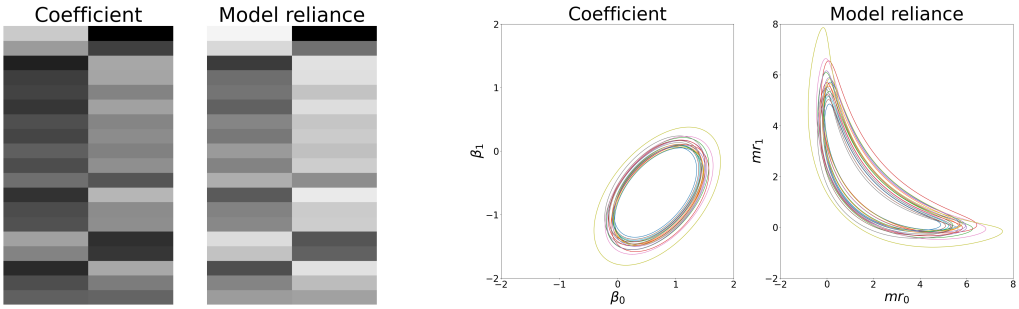
In conclusion, we found that Rashomon bound allows for reproducible and exploratory data analysis, and furthermore, its robustness against noise makes it an effective method for fields such as sociology The limitation of this paper is that our analysis is qualitative, and future research will require quantitative analysis.

References

[1] BAKER, M. 1,500 scientists lift the lid on reproducibility. *Nature 533*, 7604 (2016), 452–454.

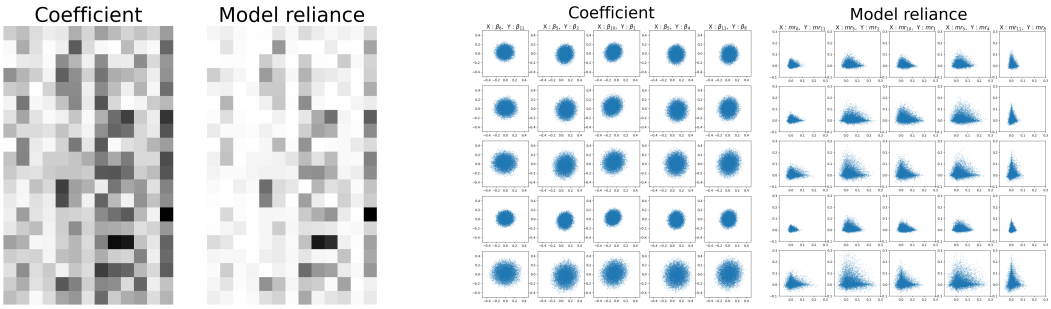
[2] DONG, J., AND RUDIN, C. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence 2*, 12 (2020), 810–824.

[3] MARSAGLIA, G. Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics 43* (1972), 645–646.



(a) (b)

Figure 1: Results for a two-variable data set. (a) Ridge regression results with noise added to the original y. The model for each dataset is shown on the vertical axis. The blacker the color, the larger the absolute value of the coefficient, and the whiter the color, the smaller the absolute value of the coefficient. (b) Result of Rashomon bound with noise added to the original y. Each contour line represents the result for each dataset.



(a) (b)

Figure 2: Results for a Boston Housing set. (a) Ridge regression results with noise added to the original y. The model for each dataset is shown on the vertical axis. The blacker the color, the larger the absolute value of the coefficient, and the whiter the color, the smaller the absolute value of the coefficient. (b) Result of Rashomon bound with noise added to the original y. Each point represents a Rashomon bound projected onto the plane of two randomly selected coefficients β_i and β_j. The horizontal axis represents the coefficients for the different β_i and β_j pairs, the model reliance, and the vertical axis represents the results for each data set.