Master Thesis

Multiple System Estimation Using Grouped Lasso

Examining the Effect of Sample Population and Registers

Author: Yanwen Zhang

Student number: 9087605

Project supervisor: Maarten Cruyff

Second Examiner: Peter van der Heijden

Word count: 3877

Applied Data Science
Utrecht University

Date: 01/07/2024

Abstract

This study evaluates the efficacy of Grouped Lasso regression models within the field of Multiple System Estimation (MSE) by simulating various data combinations using different parameters. Three sets of datasets were generated, starting with a baseline configuration of 3 registers 2 covariates with 3 levels each. The population size was fixed at 1000, with 25% sampling rate and standardization. To explore the impact of known population size in the sample, a second dataset utilized a 50% sampling rate. A third dataset increased the number of registers to 4, maintaining the same covariates and population size as baseline, to evaluate the effect of number of registers. Each dataset was analyzed under four setups to compare the effects of standardization versus non-standardization, and interaction versus factor grouping. The results revealed that the AIC/BIC methods generally outperformed the lasso methods, although AIC often introduced significant outliers. Optimal lambdas consistently exceeded the performance of the optimal lambda plus one standard deviation. While three contrasts did not achieve the target median as effectively, they exhibited much narrower interquartile ranges compared to AIC/BIC, indicating more consistent performance on precision.

This thesis project is completed with collaboration with group member: Jiajian Yan (student number 6763294), who examines the effect of lower percentage of population size in the sample and higher number of covariates, while keeping the baseline dataset the same to ensure the comparability.

Contents

1 Introduction

Nowadays, the methodology for estimating hidden populations has been significantly improved. Among these methods, Multiple System Estimation, abbreviated as MSE, has emerged as a crucial tool. The introduction of MSE has made up for incompleteness of the traditional sampling survey methodology where inaccurate survey questions, inappropriate sampling frames, and budget constraints (Zhang, 2012) has been improved. MSE is a statistical method used to estimate the size of undetected population through linking multiple incomplete population registers and analysing the overlap between them. The basic idea underlying is to fit statistical models on observed populations, identify potential interactions or relationships between populations and its covariates (if they exist), and estimate the total population size based on the information received above. MSE is commonly used in a range of academic fields such as public health, where it estimates the spread of diseases; wildlife conservation for assessing animal populations, especially endangered species; and humanitarian efforts, such as estimating the number of displaced persons in crisis situations. It is also stated in the work of UNODC (need to be cited) that the ideal data for the use of MSE is from three or more official and administrative sources.

As the number of potential models increases dramatically with the increasing registers and covariates (Cruyff et al., 2021), finding a suitable model for the data has been one of the biggest challenges for MSE. Thus, model selection is necessary but it still faces challenges when it comes to high dimensional and potential sparse data, as well as when complex interactions among variables need to be accommodated by models (Dahinden et al., 2017). Log-linear models are commonly used in MSE to estimate the unobserved population size based on the possible interactions between observed data sources. However, accuracy and reliability cannot be assured across different sets of parameters and covariates. Therefore, rigorous model assessment and selection are necessary to optimize the reliability of the estimates (IWDGMF, 1995). Traditional selection methods such as stepwise selection (forward selection and backward selection) are often used, where predictors are added or removed based on their

statistical significance and improvements they provide for the model fit, instead of fitting all possible models. However, this method is very time-consuming and possible to result in multiple models with an equivalent fit but different population size estimations (Binette & Steorts, 2022).

Like log-linear models, Poisson regression also uses a log-link function. Packages like glmnet and grplasso can do regularization for Poisson regression but the cross validation procedures cannot be directly applied to log-linear models. In the work of Dahinden et al. (2017), Logilasso package, which is uniquely designed for log-linear models, has been first introduced and the application of this package to MSE data is innovate within this field. The penalization approach involved helps in selecting significant variables through shrinking less informative coefficients to zero, which is very significant for managing the complexity of models required by MSE.

The Least Absolute Shrinkage and Selection Operator, commonly known as Lasso was originally introduced for linear models in the work of Tibshirani (1996). It is widely used to improve prediction accuracy and model interpretability, whilst achieving variable selection and regularization through penalizing and reducing the set of covariates used by the model. However, lasso cannot satisfy the hierarchical nature of log-linear models whilst Grouped Lasso is more likely to yield hierarchical models as an extension of the Lasso. Variables are divided into predefined groups based on prior knowledge or common characteristics in Grouped Lasso, in this case all interactions involved in a specific covariate can be seen as a single group. There are two types of grouping in Grouped lasso, "factor" and "interaction". Factor grouping, as proposed by Dahinden et al. (2017), means each level of the covariate or interaction is considered as a whole group and penalization is imposed on entire levels together. In addition to the factor grouping in Logilasso package, a new way of grouping is interaction grouping, where each group is defined based on interaction terms.

In recent years, human trafficking has been a severe problem throughout the world, statistical information is used to help governments and organizations to better monitor and prevent

the phenomenon from happening further (De Vries & Dettmeijer-Vermeulen, 2015). However, collecting such data has been problematic as the size of victims of human trafficking can be unstable and unreliable across and within countries over time due to different reasons, even the data from administrative agencies can vary to some extent. Therefore, better estimating true population sizes has been widely discussed by researchers and international organizations.

For the simulation study, three simulated data will be hold. The baseline setup is about 3 registers with 2 covariates at 3 levels under 25 percentage of population size in the sample. To evaluate the effect of different population size on the performance, the same setup but under 50 percentage of population size in the sample is also simulated. Different registers may also play a role, so another simulated dataset is 4 registers with 2 covariates at 3 levels under 25 percentage population size. Each data will be simulated for 10 times to assess the effectiveness of grouped lasso and stepwise function (backward selection) using AIC/BIC. For creation, treatment contrasts and non-standardization will also be added in this paper while most previous work is more focusing on polynomial and sum contrasts with standardization.

The remainder of this paper is structured as follows. Section 2 details on the relevant methods, including log-linear models, stepwise selection, and (Grouped) Lasso. Section 3 presents the analysis in detail including parameter selection and coding procedures. Results will be discussed in Section 4 with plots, and finally the paper will be ended by a discussion.

2 Methodology

2.1 Log-linear Models

In the field of Multiple System Estimation (MSE), log-linear models are statistical models commonly performed to analyze the relationship between categorical variables and/or their interactions in a contingency table, based on the assumption that the logarithm of expected frequencies can be described as a linear function of parameters. In a two-by-two contingency table, A and B represent two incomplete population registers. The log-linear model can be

expressed as:

$$\log(\mu_{ij}) = \lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB},$$

where:

- $\log(\mu_{ij})$ is the expected frequency for the cell corresponding to the i-th level of A and j-th level of B and $i, j \in \{0, 1\}$.
- λ_0 is the intercept.
- λ_i^A and λ_j^B are the main effects of registers A and B.
- λ_{ij}^{AB} is the interaction effect between registers. This cannot be estimated in MSE as the (0,0) cell is not observed.

The model effectively show how the log of expected frequencies changes with different categorical variable levels. Log-linear models are essential tools for analyzing data with multiple variables and offering insights into their interactions, which is advantageous in MSE with more than two registers and covariates. Log-linear models are commonly applied in several fields such as Epidemiology and Social Science.

2.2 Stepwise selection

As increasing the number of registers and covariates result in a significantly increased number of potential models to be performed, it is therefore important to perform model selection. It works by adding and removing predictors based on their statistical significance and improvement offered to the model fit, judged by changes in AIC and/or BIC.

Forward selection and backward selection are two types of stepwise selection, where one starts from easy to hard while another the other way around. Forward selection begins with no variables in the model, then adding the predictor one by one by choosing which one improves the most to the model until no additional significant improvement is observed. Backward

selection starts from the model with all potential variables included, then removing the least important predictor at a time until all remaining predictors have statistically significant contribution to the model. Stepwise selection can quickly identify a possible optimal model without fitting all potential combinations of predictors. However, it depends heavily on the order of predictors entered into the process and may introduce overfitting, especially in forward selection. In this paper, only backward selection will be used due to the limitations mentioned above and the time restriction.

BIC are AIC are two similar information criteria for adding and removing the variables. AIC tends to find the best model which explains the data with minimum number of predictors. The lower the AIC, the better the performance of the model is. BIC is similar to AIC but more strict with complex models by imposing a large penalty on the number of predictors. The performance of AIC/BIC models can be assessed using simulated data.

2.3 Lasso & Grouped Lasso

2.3.1 Lasso

In a log-linear model, the expected value μ_i of the observed counts y_i in a contingency table can also be expressed as:

$$\mu_i = e^{\beta_0 + \sum_j x_{ij}\beta_j}$$

where:

- $i \in \{1, n\}$ where n is the number of observed cell counts.
- $j \in \{0, p\}$ where p means the number of parameters.
- x_{ij} is a vector of predictors.
- β_0 is an intercept term.
- $\boldsymbol{\beta}$ is the parameter vector which can be estimated by maximizing the log-likelihood function:

$$\log \ell(\beta) = \sum_i \left(y_i(\beta_0 + x_{ij}\beta_j) - e^{\beta_0 + x_{ij}\beta_j} \right)$$

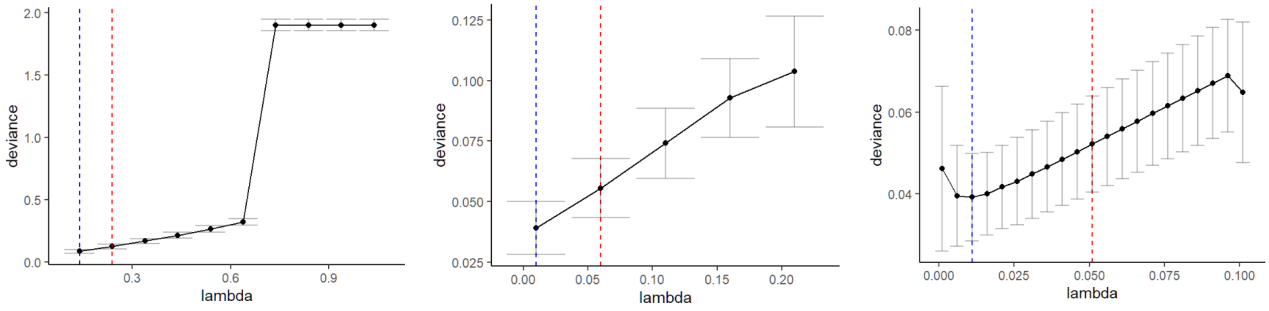
Lasso, as one regularization technique, works by penalizing on the sum of the absolute values of model coefficients, effectively improves the performance of model selection procedure. The aim of the lasso is to find the value that maximizes the lasso function, and the formula is displayed as following:

$$lasso = \log \ell(\beta) - \lambda \sum_{j=1}^{p} |\beta_j|$$

where:

- $|\beta_j|$ is the absolute value of the parameter β_j for $j \in \{1, p\}$. Noted that the intercept β_0 is excluded to ensure the baseline comparison. The larger the $|\beta_j|$, the higher the penalty, which is the larger value subtracted from the log-likelihood.

- λ is the regularization parameter which controls the amount of shrinkage and helps minimize the test error. After imposing the penalty on $\sum_{j=1}^{p} |\beta_j|$, the model is influenced by both shrinking the coefficients to zero and performing variable selection. Two optimal lambda values that result in best model performance will be selected through cross validation. Various values of lambda are tested until the optimal value is in the path, an example of this progress is shown below as plots. In the third plot, the optimal lambda is achieved at a very low value which may introduce overfitting. To test it, a slightly larger lambda value is applied, which is the red dot line.



2.3.2 Grouped Lasso

Grouped lasso is an extension of Lasso, but it imposes a penalty on the groups of coefficients instead of individual ones, which simplifies the model more effectively in terms of interpretability and manageability, and helps in preventing overfitting. Let g denotes a group and G is the total number of groups, the grouped lasso penalty can be expressed as:

$$grplasso = \log \ell(\beta) - \lambda \sum_{g=1}^{G} \sum_{j \in g} |\beta_j|$$

2.3.3 Grouping

As already mentioned previously, there are two grouping methods in Grouped Lasso, "factor" and "interactions" specification. Suppose there is a simulated dataset using simdat2 function, containing two register (R1 and R2), one covariate (X1) with two levels ('a' and 'b') and Freq denotes the frequency. K is set to 1, so only one dataset will be simulated. The table looks like in the following:

	R1	R2	X1	Freq
1	1	0	a	68
2	0	1	a	71
3	1	1	a	5
4	1	0	b	24
5	0	1	b	44
6	1	1	b	5

After applying the saturated model and treatment contrasts, the result is given as:

	(Intercept)	R11	R21	X1b	R11:R21	R11:X1b	R21:X1b	R11:R21:X1b
1	1	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	1	1	1	0	1	0	0	0
4	1	1	0	1	0	1	0	0
5	1	0	1	1	0	0	1	0
6	1	1	1	1	1	1	1	1

For this model matrix, $x_{i0} = (1, 1, 1, 1, 1, 1)$ is the intercept vector, $x_{i1} = (1, 0, 1, 1, 0, 1)$ is the coefficient vector for parameter $R11$, $x_{i2} = c(0, 1, 1, 0, 1, 1)$ is the vector for parameter

$R21$, etc.

When grouping = "factor", each level of the covariate and interaction is seen as a group. In this case, the vector (0, 1, 2, 3, 4, 5, 6, 7) is given. Group 0 represents the intercept. R11 is group 1, R21 is group 2 and X1b is group 3. The interaction terms R11:R21, R11:X1b and R21:X1b are group 4, 5 and 6 respectively. Finally, R11:R21:X1b is group 7. Therefore,

$$\lambda \sum_{g=1}^{G} \sum_{j \in g} |\beta_j| = \lambda (\sum_{j=1}^{7} |\beta_j|)$$

"Interaction" grouping, where each group is defined by interaction terms, gives the vector (0, 1, 1, 1, 2, 2, 2, 3). Group 0 represents intercept, group 1 includes the main effect of parameters R11, R21 and X1b. Group 2 denotes the interaction terms of R11:R21, R11:X1b and R21:X1b. R11:R21:X1b is then group 3. The formula now becomes:

$$\lambda \sum_{g=1}^{G} \sum_{j \in g} |\beta_j| = \lambda (\sum_{j=1}^{3} |\beta_j| + \sum_{j=4}^{6} |\beta_j| + |\beta_7|)$$

For MSE data, the (Grouped) Lasso works the same but the parameters are estimated through removing the unobserved rows in the dataset, where data are not observed in any of the registers. Once the parameters are estimated and all x_{ij} in the model matrix are included, the population size estimate is obtained by:

$$\hat{y}_i = \sum_i e^{\hat{\beta}_0 + \sum_{j=1}^{p} x_{ij} \hat{\beta}}$$

2.4 Contrasts

There are three different types of contrasts within the MSE framework, which are treatment, polynomial and sum. Different contrasts can be specified based on the research question, hypothesises under the study and categorical features of the data. In the previous example, only treatment contrasts is used where the effect of another level (in this case is b) of categorical variable (e.g., X1) is measured against one level chosen as baseline (e.g., a). In this design

matrix, the baseline level is represented as a row of zeros. Polynomial contrasts assign polynomial terms to levels of categorical variables, which commonly used to work with ordinal data. It can be used to detect and model trends over ordered factor levels, capturing linear or high-order relationships between variables. Sum contrasts without making any reference points, compare each level to the overall mean of all levels, offering a comprehensive and balanced view of how each category relates to the average. Different contrasts yield different model matrix and in turn affect the interpretation of parameters. Thus, choosing proper contrast type is not negligible. In this paper, all three contrasts will be run to compare the performance of (Grouped) Lasso function.

2.5 Standardization

Standardization is an important preprocessing step to ensure the fair comparison and combination of data from different sources based on a common scale. In MSE, the standardization of a model matrix plays a crucial role especially when implementing regularization techniques like Lasso regression. Variables with large variances tend to have smaller unstandardized coefficients due to a small change in the feature can significantly impact the outcome. In this case, the absolute value of coefficients $|\beta_j|$ is relatively smaller, a lesser Lasso penalty will be applied on that variable as it is proportional to $|\beta_j|$, making these high-variance variables less likely to be reduced to zero. Conversely, a variable with small variance tend to have larger unstandardized coefficient to ensure comparable changes in the response variable, therefore a heavier Lasso penalty is expected to increase the probability of being shrunk towards zero. Without standardization, it can lead to a biased selection process where variables with larger variance are unfairly favored. Standardization addresses this issue through equalizing the scale of all variables, making the regularization process fairer and more effective in parameter selection, enhancing the overall accuracy and reliability of MSE analysis.

3 Analysis & Data Simulation

To evaluate how the performance of Grouped Lasso on MSE data is, simulated datasets with different combination of registers and covariates are conducted using simdat2 function in R. These two function share the similarities where both generate a sequence of frequencies from log linear parameters. Simdat2 generates a sequence of observed frequencies from different sets of randomly generated log linear parameters. In this paper, treatment contrasts and non-standardization will be added for creation while previous work is mostly focus on polynomial and sum contrasts with standardization. Therefore, simdat2 is a better choice as the default setting of this function fits the requirements. There is a range of arguments from simdat2 function that can be specified to form the desired data structure, some of which include (citation needed):

- *regs* which is the number of registers.
- *xlevs* is a vector with the number of levels of the covariates. If NULL, no covariates are made. E.g., (3, 3) means two covariates, each with three levels.
- N is the population size. N is set to 1000 in this paper.
- k is the number of simulated data sets, and set to 10.
- $mean, sd$ are arguments used to draw the log linear parameters. In this paper, only mean value is changed to achieve certain population size in the sample (E.g., 25 percent and 50 percent). *sd* is set to 0.5 as default for the main effects and to sd/sqrt(interaction order) for the interaction parameters.
- *seed* is the optional seed for reproducibility.

In total three simulated datasets are generated with different parameter combinations to keep comparability and interpretation. The baseline is 3 registers, 2 covariates with 3 levels and 25 percent population in the sample, with standardization. The default total population of 1000 is being used for all three datasets, with k being fixed at 10 for a total of 10 datasets

generated for each combination. To evaluate the effect of known population size in the sample, the same setup but with 50 percent population size is shown as second simulation data. In the third simulation dataset, an increased number of registers (4 registers) with same covariates and population size is performed. Noted that the parameters cannot be estimated are set to 0 for the simulated data. In order to compare the effect of standardization/non-standardization and show the difference between interaction and factor grouping, each simulated data is conducted under 4 setups, which are standardization with interaction grouping, standardization with factor grouping, non-standardization with interaction grouping and non-standardization with factor grouping.

4 Results

A good way to assess the outcomes of the various lasso contrasts and the AIC/BIC stepwise searches would be using box plots, where a red horizontal line is marked at y = 1000 representing a total simulated population of 1000. Box plots can display the medians for each methods, their interquartile ranges, as well as identifying any outliers. A perfect score for any method would be having its median as close to 1000 as possible, with its interquartile ranges as narrow as possible, thus achieving both high accuracy and high precision. "poly", "sum1" and "treat1" indicate the contrasts methods under a larger lambda value which is the optimal lambda plus one standard deviation in this case, while "poly2", "sum2" and "treat2" represent the results from the optimal lambda value.

4.1 Baseline

The baseline data, 3 registers 2 covariates with 3 levels under 25% population size will be simulated in this section.

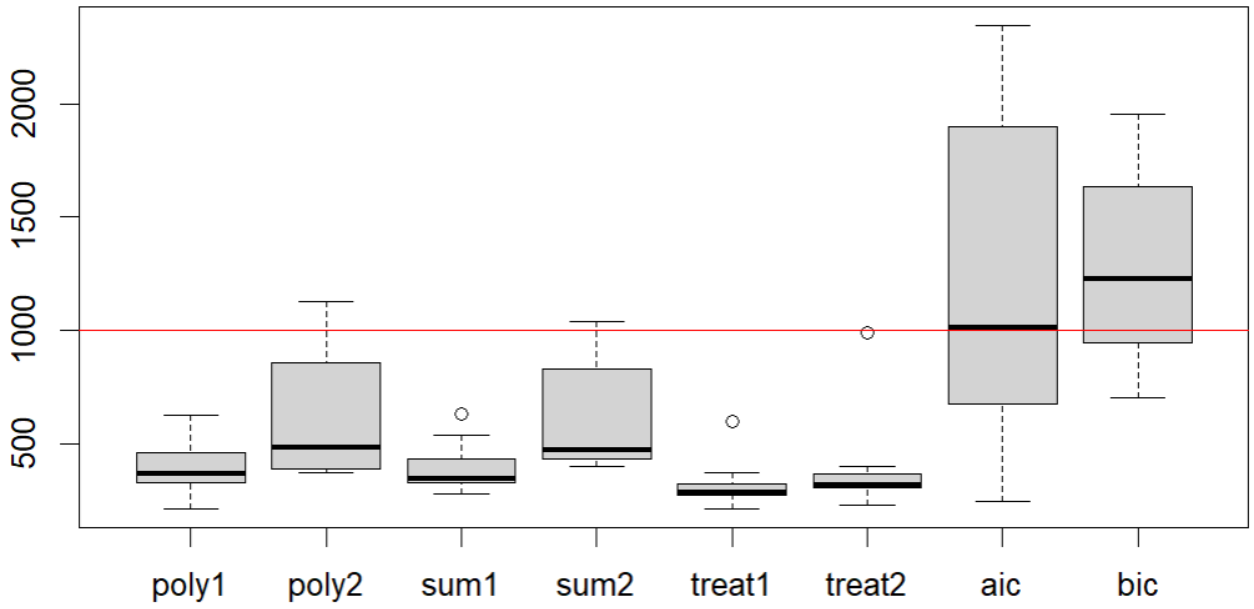


Figure 1: With standardization, interaction grouping, 50% of outliers were removed for AIC

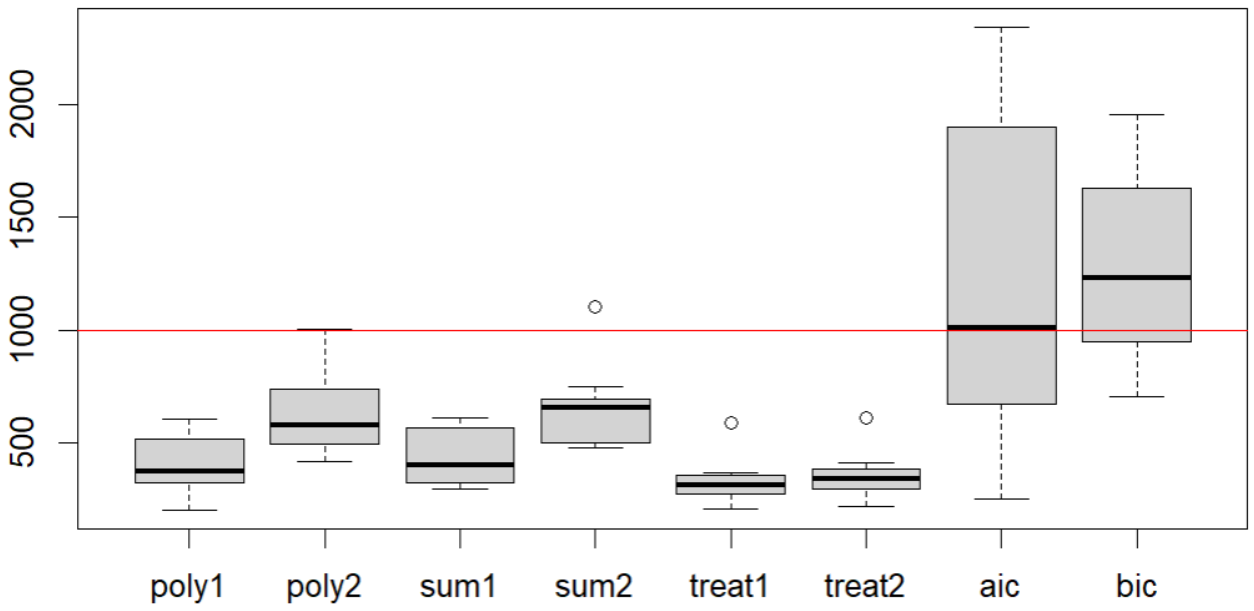


Figure 2: With standardization, factor grouping, 50% of outliers were removed for AIC

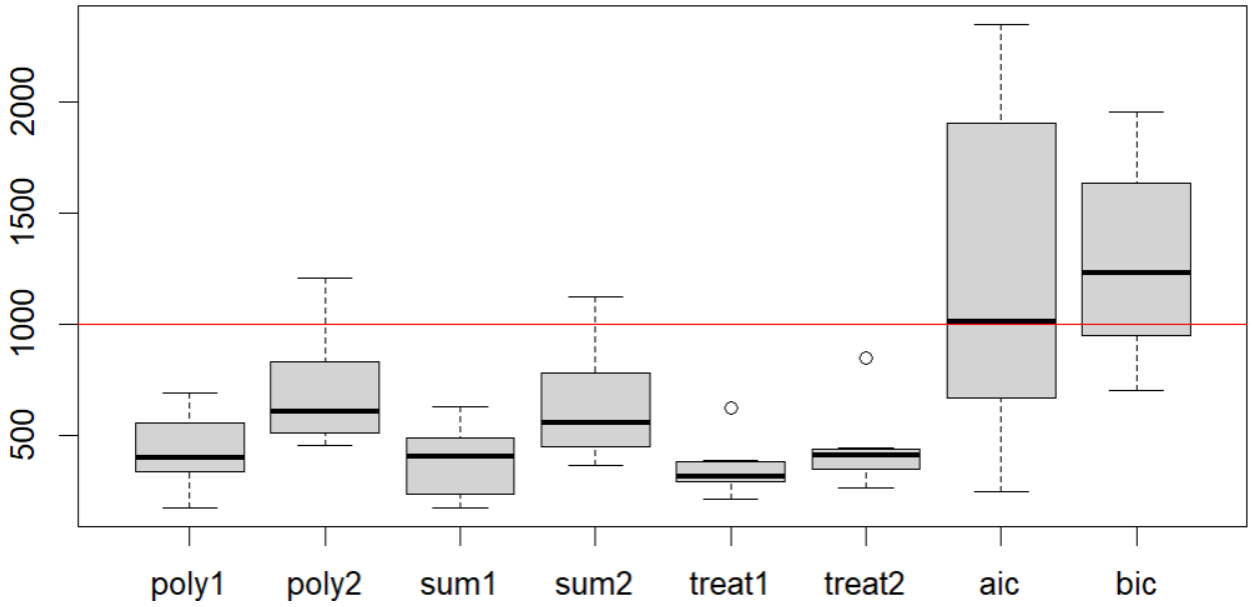


Figure 3: Non standardization, interaction grouping, 50% of outliers were removed for AIC

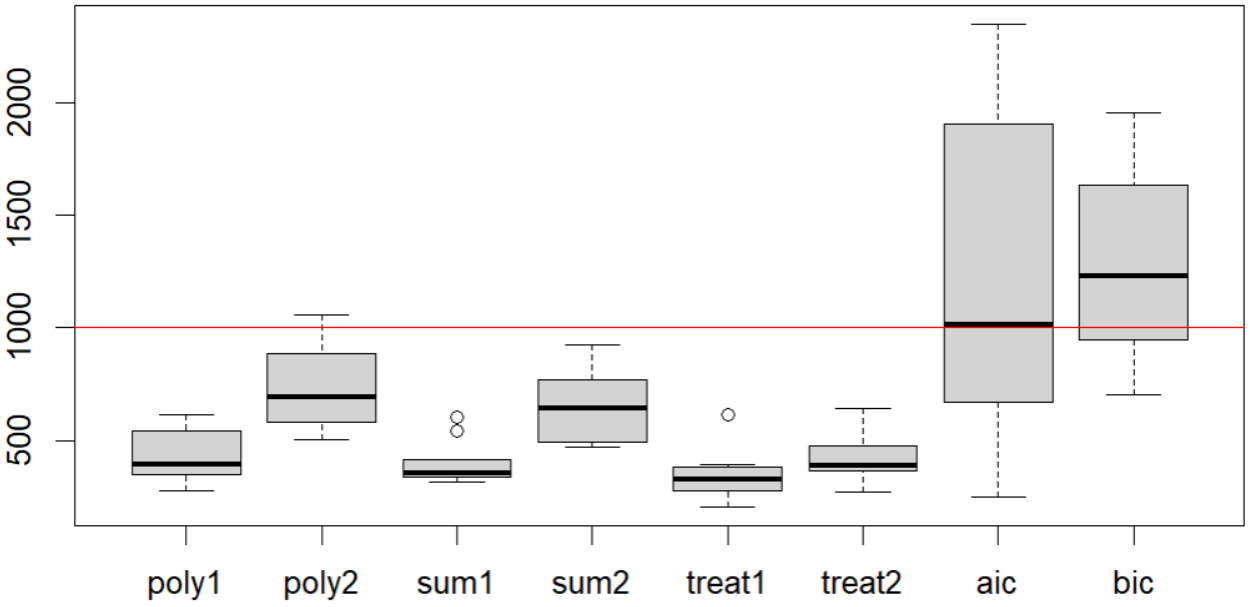


Figure 4: Non standardization, factor grouping, 50% of outliers were removed for AIC

Looking at Figure 1 to 4, it is surprisingly to see that under all 4 setups, AIC has provided the best performing results with its median marked almost exactly on the true population line. BIC has its median at a higher value but relatively works better than grouped lasso methods where those significantly undershot the true population. However, AIC always obtains the highest number of outliers, which somewhat distorts the box plot for AIC. The optimal polynomial and sum contrasts (e.g., poly2 and sum2) have both outperformed the results obtained with 1 lambda higher (e.g., poly1 and sum1). Treatment contrasts have performed the worst with a lowest median. Even though lasso methods perform not very well, they all have much narrower interquartile range and no outliers compared to AIC/BIC, indicating that their predicitons are more precise, especially treatment contrasts.

4.2 Higher Percentage of Population Size

Compared to the baseline data, the population size is increased from 25% to 50% in the sample while other parameter remain unchanged. Therefore, this simulated dataset is 3 registers 2 covariates with 3 levels under 50% population size.

Analyzing the performance from figure 5 to 8, consistent patterns in the behavior of MSE models can be observed under all setups. Both polynomial and sum contrasts, specifically "poly1" and "sum1", consistently underperform than AIC and BIC. However, the difference between optimal lambda and 1 lambda higher is not as significant as shown in figure 1 to 4, compared with the baseline. In this case, AIC no longer has its median marked on the 1000 line but the whole interquartile range has been above while BIC has its whole interquatile range below. AIC has removed slightly lower number of outliers, in this case 4 out of 10. Lasso methods overall perform better than baseline as their median all increase, especially treatment contrasts with much higher median mark but wider interquartile range. Judging the difference of AIC and BIC between two different population size, it can be seen that changing the percentage of population to a higher amount in the sample did not have any significant effect on the estimations of the true population.

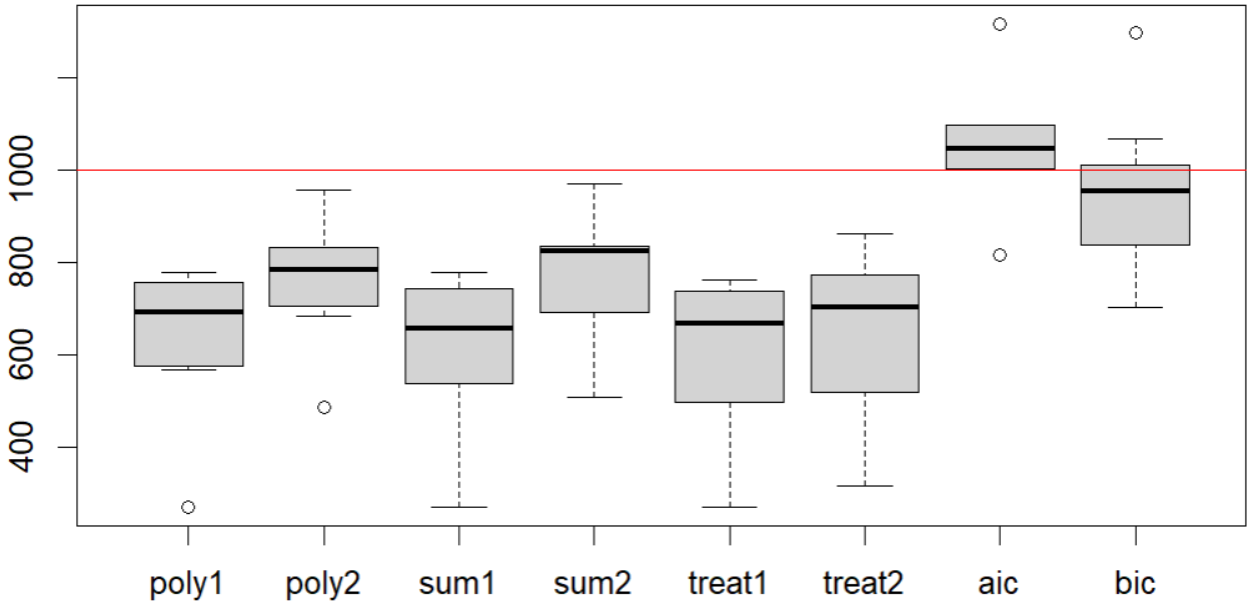


Figure 5: With standardization, interaction grouping, 40% outliers were removed for AIC

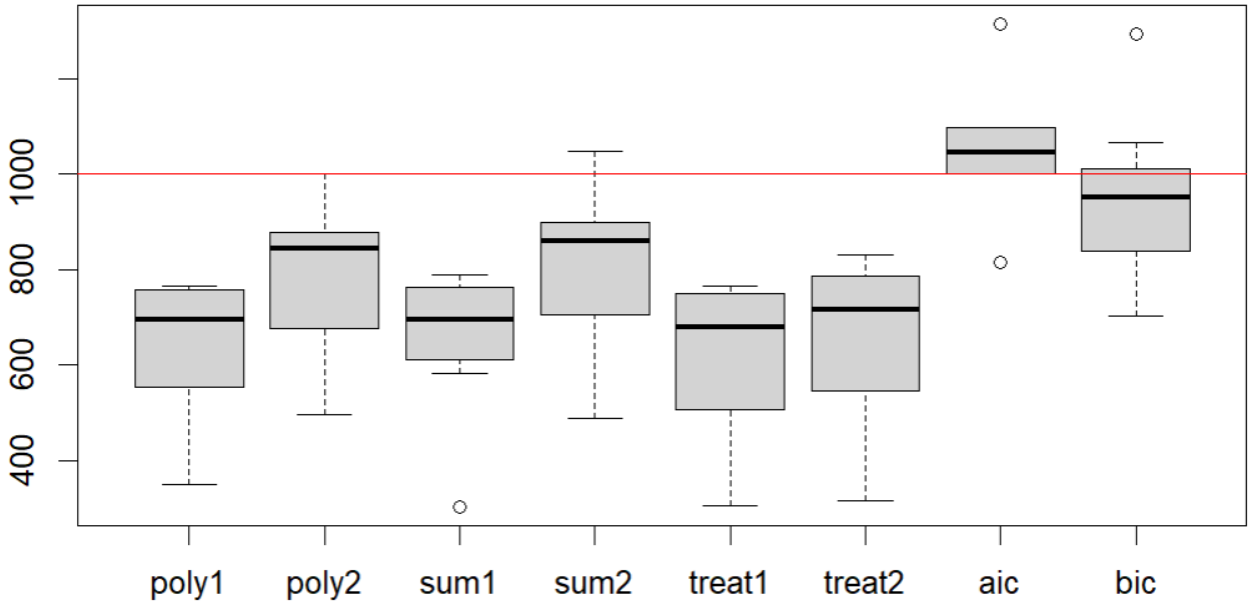


Figure 6: With standardization, factor grouping, 40% outliers were removed for AIC

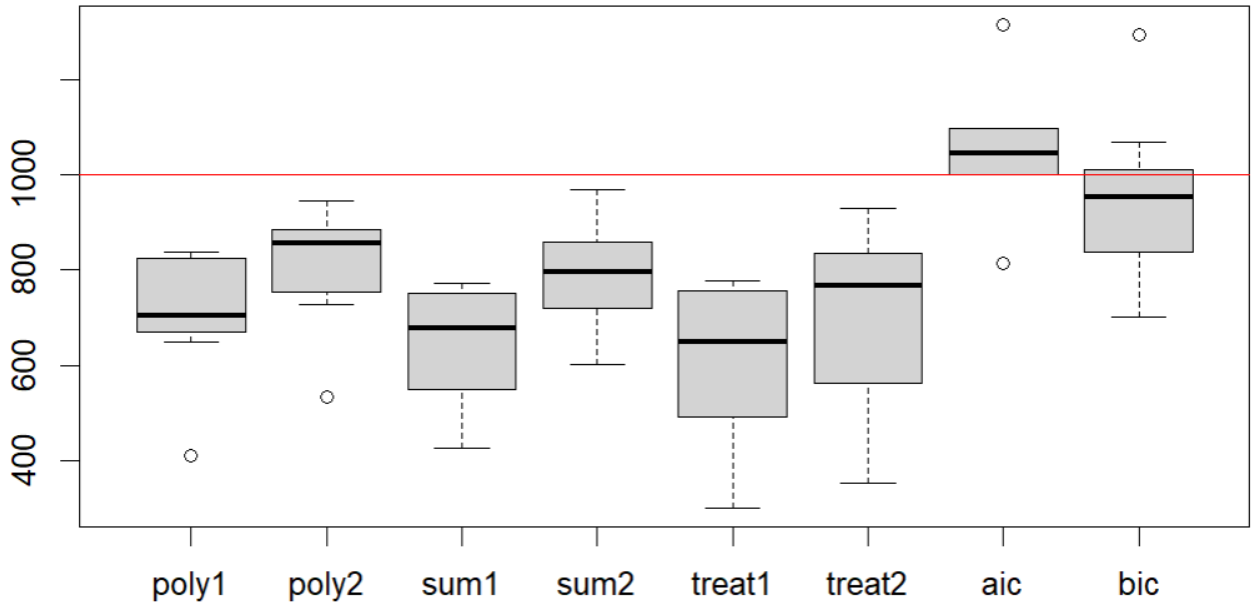


Figure 7: Non standardization, interaction grouping, 40% outliers were removed for AIC

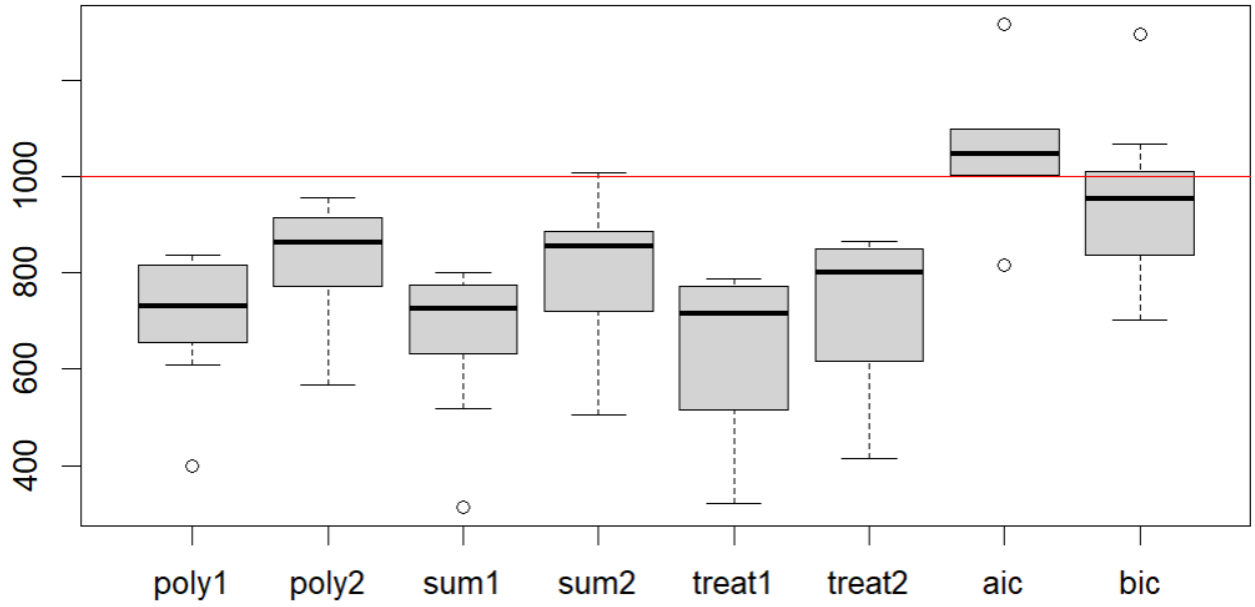


Figure 8: Non standardization, factor grouping, 40% outliers were removed for AIC

4.3 Higher Number of Registers

This section increases the number of registers from 3 to 4, while other parameters hold still. So, this simulated dataset is 4 registers 2 covariates with 3 levels under 25% population size in the sample.

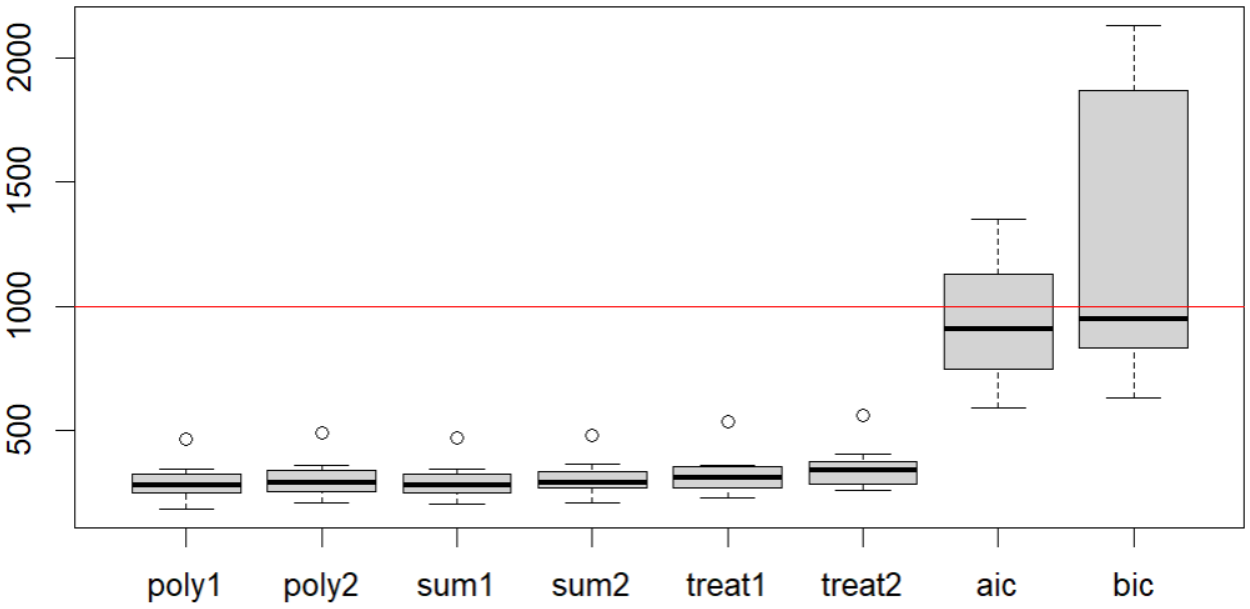


Figure 9: With standardization, interaction grouping, 70% outliers were removed for AIC

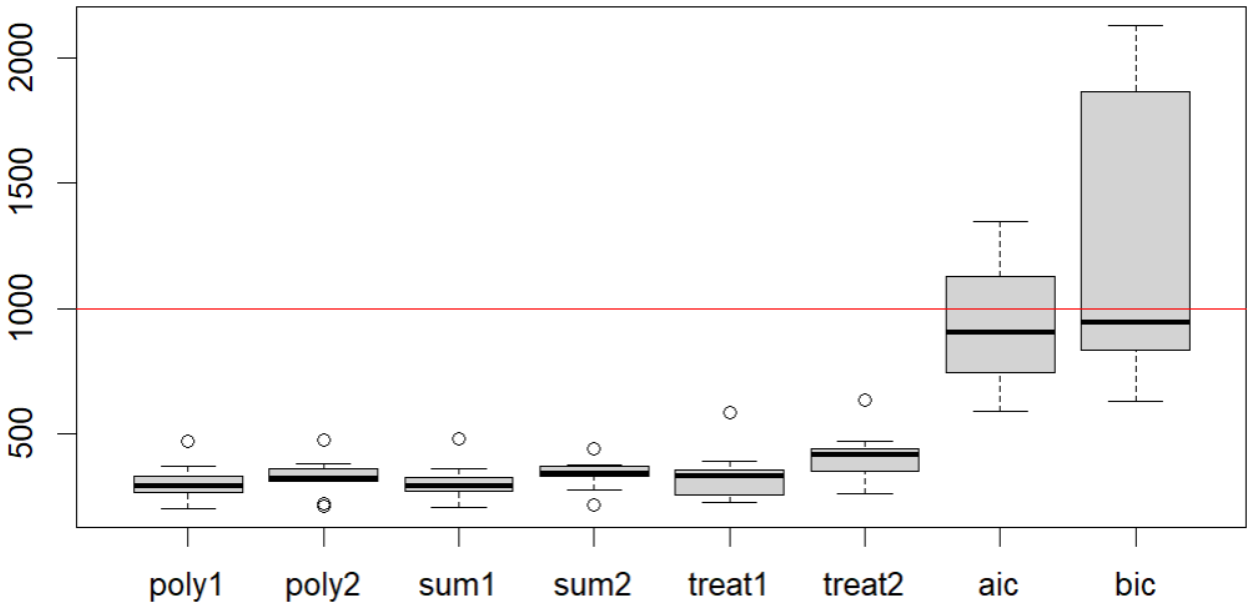


Figure 10: With standardization, factor grouping, 70% outliers were removed for AIC

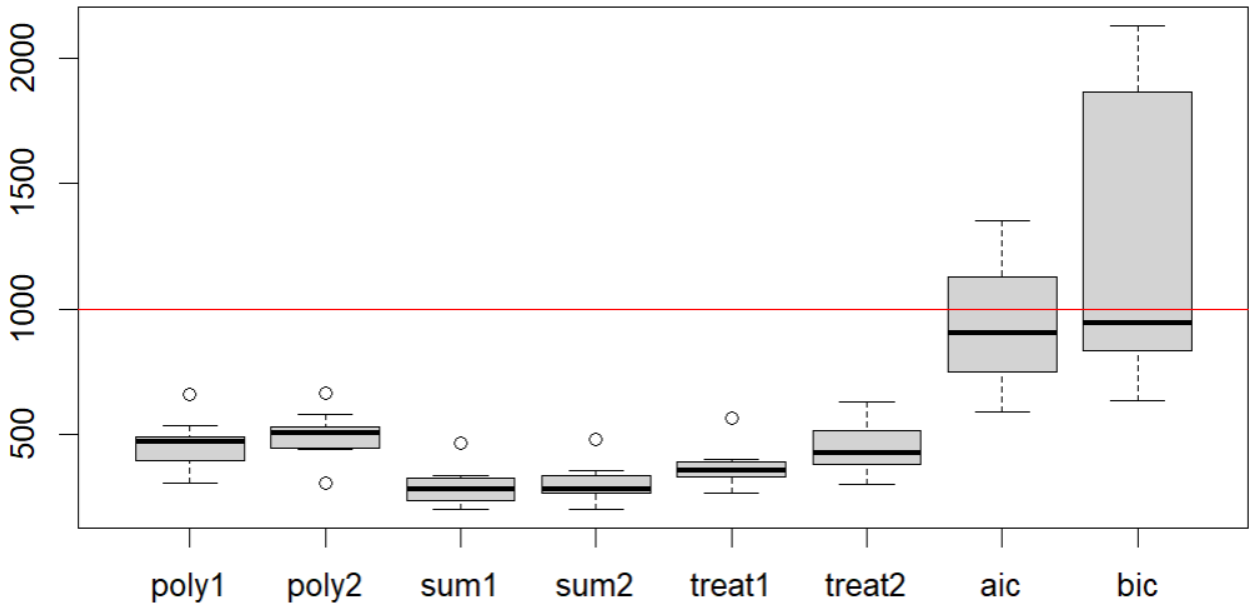


Figure 11: Non standardization, interaction grouping, 70% outliers were removed for AIC

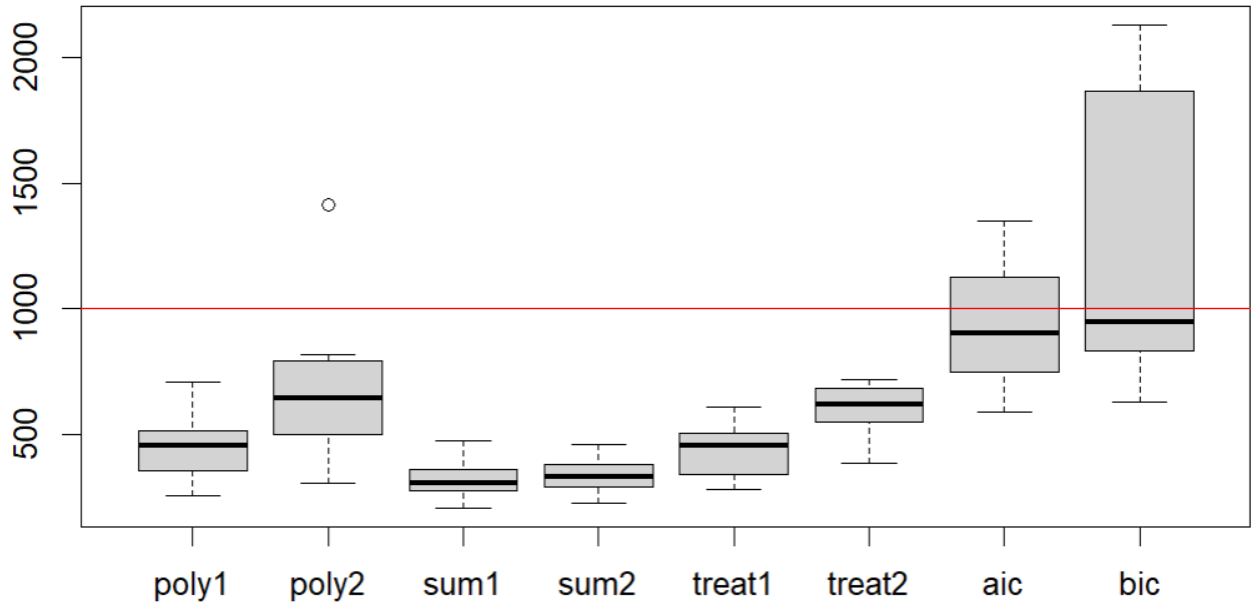


Figure 12: Non standardization, factor grouping, 70% outliers were removed for AIC

Compared to the baseline data, Figure 9 to 12 displays a much narrower interquartile range for polynimal and sum contrasts with their median and overall predictions have further lowered over the baseline. In the two figures of non standardization, "poly2" seems to be outperforming sum and treatment contrasts. This is not the case when standardization is applied as different contrast methods have performed very similarly. However, the various

contrast methods still lack significantly behind BIC and AIC, although the latter has as many as seven out of ten outliers, some of which are at extreme magnitudes. Compared to the baseline plot, narrower interquartile range for AIC but wider interquartile range for BIC, with both median relatively lower than 1000 line. In this case, lasso methods provide precise but inaccurate predictions, which is the opposite of AIC and BIC.

5 Discussion

To test the efficacy of Grouped Lasso regression models in the field of MSE, this study simulates different combinations of data using various parameters. Three sets of simulated datasets are generated, maintaining a baseline configuration of three registers and two covariates with three levels each, ensuring a standard total population of 1000 and 25% population sampling with standardization. To assess the impact of known population size on the sample, the same setup with a 50% population size is used for the second dataset. The third dataset increases the number of registers to four, keeping the covariates and population size the same. Each dataset is processed under four setups to compare effects of standardization/non-standardization and interaction/factor grouping, providing a detailed examination of the different modeling strategies within MSE.

The majority of the combinations did not perform as well, with AIC/BIC often outperforming any contrast methods even though AIC often introduces severe outliers. The optimal lambdas consistently outperformed the optimal lambda plus one standard deviation, which aligns with expectations. Sum and poly contrasts, while typically underperforming in achieving the target median, showed much narrower interquartile ranges compared to their AIC/BIC counterparts.

There are several limitations to this study, mostly due to computational and time constraints, as well as limitations from the lasso package. Choosing different standard deviation value where it is not divided by the interaction order for the interaction parameters to avoid the potential of favoring the AIC/BIC. Moreover, setting the inestimable parameters to 0 may

have also favored the AIC/BIC. The number of simulations was limited to ten for each combination, which is arguably sufficient to show an overall trend for all combinations and setups. The number of parameter combinations could be increased, and the range of parameters that provide the greatest advantage for lasso methods could be explored systematically. Additionally, the study used a softened stop criterion for finding the optimal lambda, changing from the default value of 1e-08 to 1e-06 to prevent errors caused by the algorithm not finding the optimal lambda. The total population and the standard deviation used to generate simulated data were not varied throughout the simulations, which could be areas for further exploration in future studies. This paper provides insights into the potential application of group lasso methods in MSE, though it does not conclude why group lasso methods do not offer an advantage in many parameter combinations, those instances where they do provide a promising direction for future research. Using other algorithms such as ridge and elastic net can also be interesting for the future research.

6 Reference

Binette, O., & Steorts, R. C. (2022). On the Reliability of Multiple Systems Estimation for the Quantification of Modern Slavery. Journal of the Royal Statistical Society. Series a. Statistics in Society/Journal of the Royal Statistical Society. Series a, Statistics in Society, 185(2), 640–676. https://doi.org/10.1111/rssa.12803

Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases. (1995). American Journal of Epidemiology, 142(10), 1059–1068. https://doi.org/10.1093/oxfordjournals.aje.a117559

Cruyff, M., Overstall, A. M., Papathomas, M., & McCrea, R. S. (2021). Multiple system estimation of victims of Human trafficking: model assessment and selection. Crime & Delinquency/Crime and Delinquency, 67(13–14), 2237–2253. https://doi.org/10.1177/00111 28720981908

Dahinden, C., Parmigiani, G., Emerick, M. C., & Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. BMC Bioinformatics, 8(1). https://doi.org/10.1186/1471-2105-8-476

De Vries, I. & C. Dettmeijer-Vermeulen, Extremely wanted: human trafficking statistics-what to do with the hodgepodge of numbers? In Kangaspunta, K. (ed.) Forum on Crime and Society, Special Issue: Researching Hidden Populations: Approaches to and Methodologies for Generating Data on Trafficking in Persons, 8, pp. 15-37, 2015.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288.

Zhang, S. X., Trafficking of Migrant Laborers in San Diego County: Looking for a Hidden Population, San Diego, CA: San Diego State University, 2012.

7 Appendix

The codes used in this thesis can be found in the following Google Drive link: `https://drive.google.com/drive/folders/1mRYoBaZSMvyKo1aRjXkED7QIHsE_9Jqo?usp=drive_link`

Baseline_323_k10_25_simdat2: This is the codes for the baseline dataset, where 3 registers 2 covariates with 3 levels, under 25% population size using simdat2 function, run for 10 times.

323_k10_50_simdat2: Higher percentage of population size in the sample, which increases to 50% population size compared to the baseline dataset.

423_k10_25_simdat2: Higher number of registers, which increases to 4 registers compared to the baseline dataset.