A model robust subsampling approach for Generalised Linear Models in big data settings

Amalan Mahendran, Helen Thompson, James McGree
School of Mathematical Sciences, Queensland University of Technology

Introduction

- **Subsampling -** subset of the big data is analysed and used as the basis for inference.
- **A key question** is how to select an informative subset based on the questions being asked of the data.
- **Recent approaches for regression modelling** are of obtaining a subsample based on subsampling probabilities that are determined for each data point.
- **A limitation** is that the appropriate subsampling probabilities rely on an assumed model for the big data.
- **To overcome this we propose a model robust approach where a set of models are considered, and the subsampling probabilities are evaluated based on the weighted average of probabilities.**

Literature review

- Current subsampling approaches that rely on a statistical model that is assumed to appropriately describe the big data and estimate the model parameters:
  1. softmax regression (Yao and Wang 2019)
  2. quantile regression (Ai, Wang, et al. 2021)
  3. Generalised Linear Models(GLMs) (Wang, Zhu, and Ma 2018; Ai, Yu, et al. 2021)
- Potential available solutions:
  1. For **linear regression** - select the best candidate model from a pool of models based on the Bayesian Information Criterion.
  2. For **GLMs** - using space-filling or orthogonal Latin hypercube designs so that a wide range of models could be considered (Shi and Tang 2021; Meng et al. 2021; Yi and Zhou 2023).

General subsampling algorithm for GLM

Algorithm 1 General subsampling algorithm (Ai et al. 2021b)

1: **Sampling:** Assign $\phi_i, i = 1, \ldots, N$, for F_N. Based on $\boldsymbol{\phi}$, draw an r size subsample with replacement from F_N to yield $S = \{\boldsymbol{x}_l^*, y_l^*, \phi_l^*\}_{l=1}^r = (\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{\phi}^*)$.

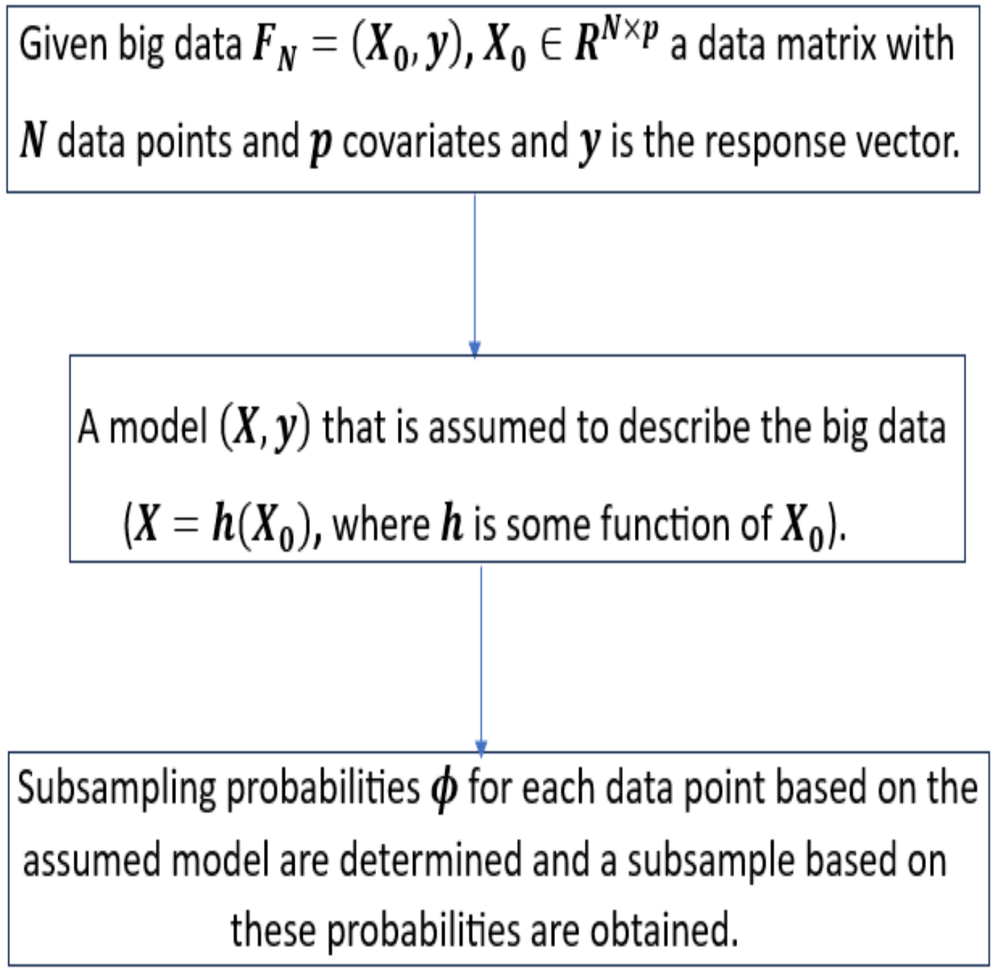
2: **Estimation:** Based on S, using $L(\boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{\phi}^*)$ the likelihood function, find:

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \ \log L(\boldsymbol{\theta}|\boldsymbol{X}^*, \boldsymbol{y}^*, \boldsymbol{\phi}^*) \\ &\equiv \arg\max_{\boldsymbol{\theta}} \ \frac{1}{r}\sum_{l=1}^{r} \frac{y_l^* u(\boldsymbol{\theta}^T \boldsymbol{x}_l^*) - \psi(u(\boldsymbol{\theta}^T \boldsymbol{x}_l^*))}{\phi_l^*}.\end{aligned}$$

3: **Output:** $\tilde{\boldsymbol{\theta}}$ and S.

- In the Algorithm 1 a subsample S is obtained using the subsampling probabilities ϕ that is determined based on a research problem. (For example prediction and/or parameter estimation).

Optimal subsampling approach for GLM (Ai, Yu, et al. 2021)

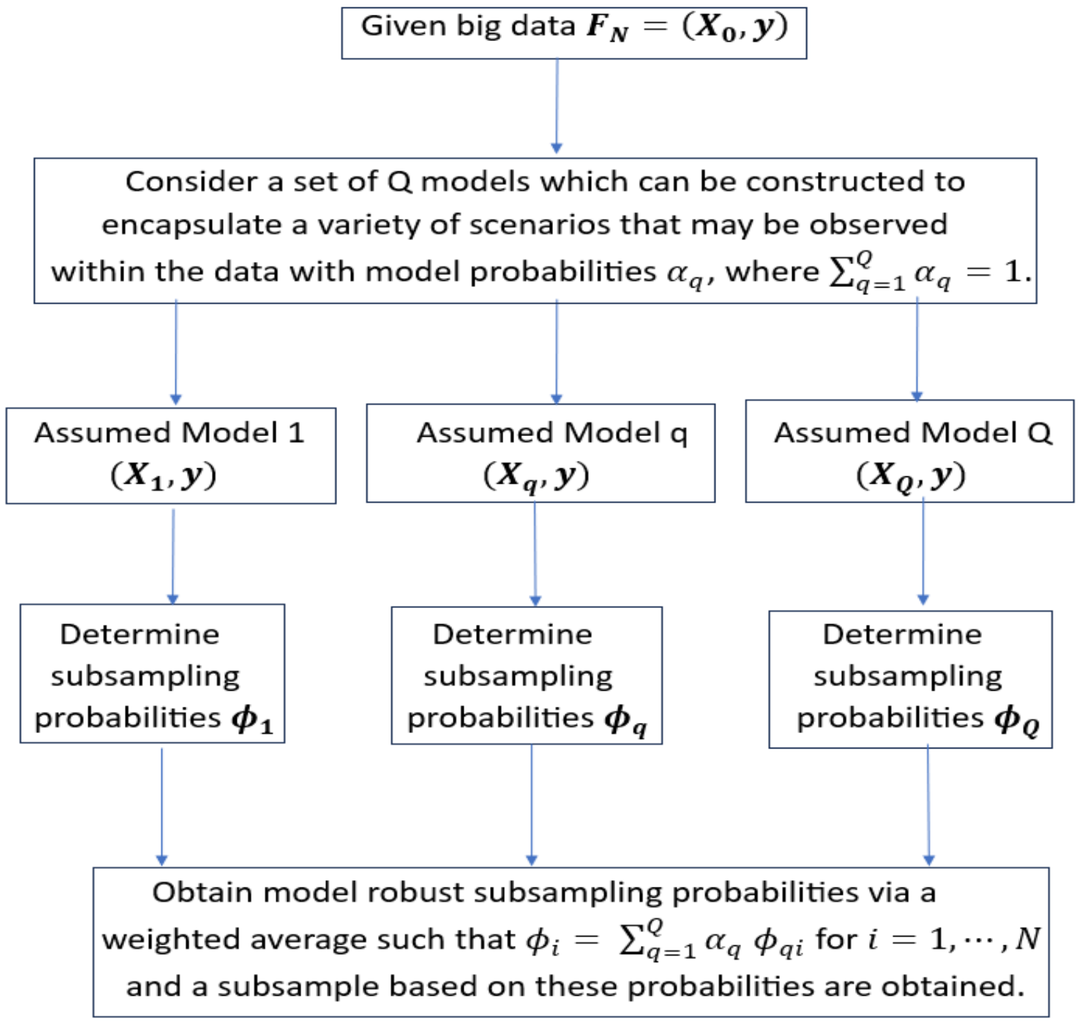


- $\phi = \phi^{mMSE}$ is obtained for A-optimality criterion (minimising the asymptotic mean squared error $\tilde{\theta}$). (Theorem 3)
- ϕ^{mMSE} cannot be determined directly as they depend on $\hat{\theta}_{MLE}$, the full data estimated model parameter.
- To address this a two-stage subsample strategy was proposed,
  - Stage 1- An initial random sample of the big data is used to estimate $\hat{\theta}_{MLE}$.
  - Stage 2- This estimate is then used to approximate ϕ^{mMSE} and obtain an optimal subsample.

Some major limitations and solutions of the optimal subsampling approach

1. **Computational expense in obtaining ϕ^{mMSE}** - for GLMs using the Johnson-Lindenstrauss Transform (JLT) and subsampled Randomised Hadamard Transform (SRHT) downsize matrix volume (Lee, Schifano, and Wang 2021).
2. **Approximation for the optimal subsampling probabilities can lead to some data points having zero probability of being selected** - resolve this by setting these subsampling probabilities to a small value to ensure such data points have some (non-zero) probability of being selected (Ai, Yu, et al. 2021).
3. **Subsampling probabilities are evaluated based on an assumed model and they are generally only optimal for this model** - We propose the development of methods that yield subsampling probabilities that are robust to the choice of model.

Model robust subsampling for GLM



- A set of Q models which can encapsulate a variety of scenarios observed within the data.
- For each model in this set, define our *a priori* model probabilities α_q for $q = 1, \ldots, Q$. For the q^{th} model $X_q = h_q(X_0)$.
- We form ϕ, that addresses the analysis aim regardless of which model is actually preferred for the big data, via a weighted average (based on α_q) of the subsampling probabilities that would be obtained for each model (singularly).

Two stage algorithm for the model robust subsampling method

Algorithm 2 Two-stage model robust subsampling algorithm for GLMs.

Stage 1

1: **Random Sampling :** Assign $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ for F_N. For example, in logistic regression $\phi_i = \phi_i^{prop} = c_0^{1-y_i} c_1^{y_i}$, where $c_0 = \left(2N - 2\sum_{i=1}^{N} y_i\right)^{-1}$ and $c_1 = \left(2\sum_{i=1}^{N} y_i\right)^{-1}$) are proportional subsampling probabilities that are based on the response composition, and $\phi_i = N^{-1}$ which is uniform subsampling probabilities for Poisson regression. According to $\boldsymbol{\phi}$ draw random subsamples of size r_0, such that $S_{r_0} = \{h_q(\boldsymbol{x}_{0l}^{r_0}), y_l^{r_0}, \phi_l^{r_0}\}_{l=1}^{r_0} = (h_q(\boldsymbol{X}_0^{r_0}), \boldsymbol{y}^{r_0}, \boldsymbol{\phi}^{r_0})$ for $q = 1, \ldots, Q$.

2: **Estimation:** For $q = 1, \ldots, Q$ and S_{r_0}, find:

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_q^{r_0} &= \arg\max_{\boldsymbol{\theta}_q} \ \log L(\boldsymbol{\theta}_q | \boldsymbol{X}_q^{r_0}, \boldsymbol{y}^{r_0}, \boldsymbol{\phi}^{r_0}) \\ &\equiv \arg\max_{\boldsymbol{\theta}_q} \ \frac{1}{r_0} \sum_{l=1}^{r_0} \left[\frac{y_l^{r_0} u(\boldsymbol{\theta}_q^T \boldsymbol{x}_{ql}^{r_0}) - \psi(u(\boldsymbol{\theta}_q^T \boldsymbol{x}_{ql}^{r_0})}{\phi_l^{r_0}} \right].\end{aligned}$$

Stage 2

1: **Optimal subsampling probability:** Estimate optimal subsampling probabilities $\boldsymbol{\phi}^{mMSE} = \sum_{q=1}^{Q} \alpha_q \boldsymbol{\phi}_q{}^{mMSE}$ or $\boldsymbol{\phi}^{mVc} = \sum_{q=1}^{Q} \alpha_q \boldsymbol{\phi}_q{}^{mVc}$ with $\hat{\boldsymbol{\theta}}_{qMLE}$ replaced by $\tilde{\boldsymbol{\theta}}_q^{r_0}$.

2: **Optimal subsampling and estimation:** Based on $\boldsymbol{\phi}^{mMSE}$ or $\boldsymbol{\phi}^{mVc}$, draw subsample of size r from F_N, such that $S_r = \{h_q(\boldsymbol{x}_0{}_l^r), \boldsymbol{y}_l^r, \boldsymbol{\phi}_l^r\}_{l=1}^{r} = (h_q(\boldsymbol{X}_0^r), \boldsymbol{y}^r, \boldsymbol{\phi}^r)$ for $q = 1, \ldots, Q$. Combine S_{r_0} and S_r to form $S_{(r_0+r)}$, and obtain :

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_q &= \arg\max_{\boldsymbol{\theta}_q} \ \log\left(L(\boldsymbol{\theta}_q | \boldsymbol{X}_q^{r_0}, \boldsymbol{y}^{r_0}, \boldsymbol{\phi}^{r_0}, \boldsymbol{X}_q^r, \boldsymbol{y}^r, \boldsymbol{\phi}^r)\right) \\ &\equiv \arg\max_{\boldsymbol{\theta}_q} \ \frac{1}{r_0 + r} \left[\sum_{k \in \{r_0, r\}} \sum_{l=1}^{k} \frac{y_l^k u(\boldsymbol{\theta}_q^T \boldsymbol{x}_{ql}^k) - \psi(u(\boldsymbol{\theta}_q^T \boldsymbol{x}_{ql}^k))}{\phi_l^k} \right].\end{aligned}$$

3: **Output:** $\tilde{\boldsymbol{\theta}}_q$ and $S_{(r_0+r)}$ for $q = 1, \ldots, Q$.

- Our proposed model robust subsampling approach is assessed via simulation and in real-world scenarios under poisson and logistic regression.
- In the simulation study and real-world scenario for poisson regression we compare our proposed model robust subsampling algorithm with the optimal subsampling approach of Ai, Yu, et al. (2021) and random sampling.
- We used the R statistical programming language to code the simulation study and the real-world application.

Simulation study

Table 1: Model set assumed fo the simulation study

Model set
$\theta_1 + \theta_2 x_1 + \theta_3 x_2$
$\theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1^2$
$\theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_2^2$
$\theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2$

Data Model: $\boldsymbol{X_1}, \boldsymbol{X_2}$

1. Random sampling to estimate the parameters of the data model.
2. Optimal subsampling under the data model.
3. to 5. Optimal subsampling under alternative models (All models in Table 1 except the data model).
6. Model robust subsampling of the data model under the assumption that each of the q models are equally likely a priori.

Using the estimated model parameters from the subsamples evaluate the simulated mean squared error, $\text{SMSE}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{p+t} (\tilde{\theta}_{nm} - \boldsymbol{\theta}_n)$, where $\tilde{\boldsymbol{\theta}}$ is an $M \times (p+t)$ matrix, $p + t$ is the number of parameters in the data model.

- For poisson regression we generate the covariate data X from the uniform distribution.
- Set $N = 10000, Q = 4, \alpha_q = \frac{1}{4}$, $r_0 = 100, r = 100, 200, \ldots, 1400$ and $M = 1000$.
- For a selected data model under the six scenarios evaluate the SMSE to compare θ with $\tilde{\theta}$.
- R code for the simulation study -

Results

- In Figure 1, SMSE values for various r sizes show that:
  - On average, random sampling performs worst.
  - The proposed model robust approach and the optimal subsampling method based on the data generating model perform the best.
  - Model robust approach is preferable overall even though the complex model performs well in some instances.

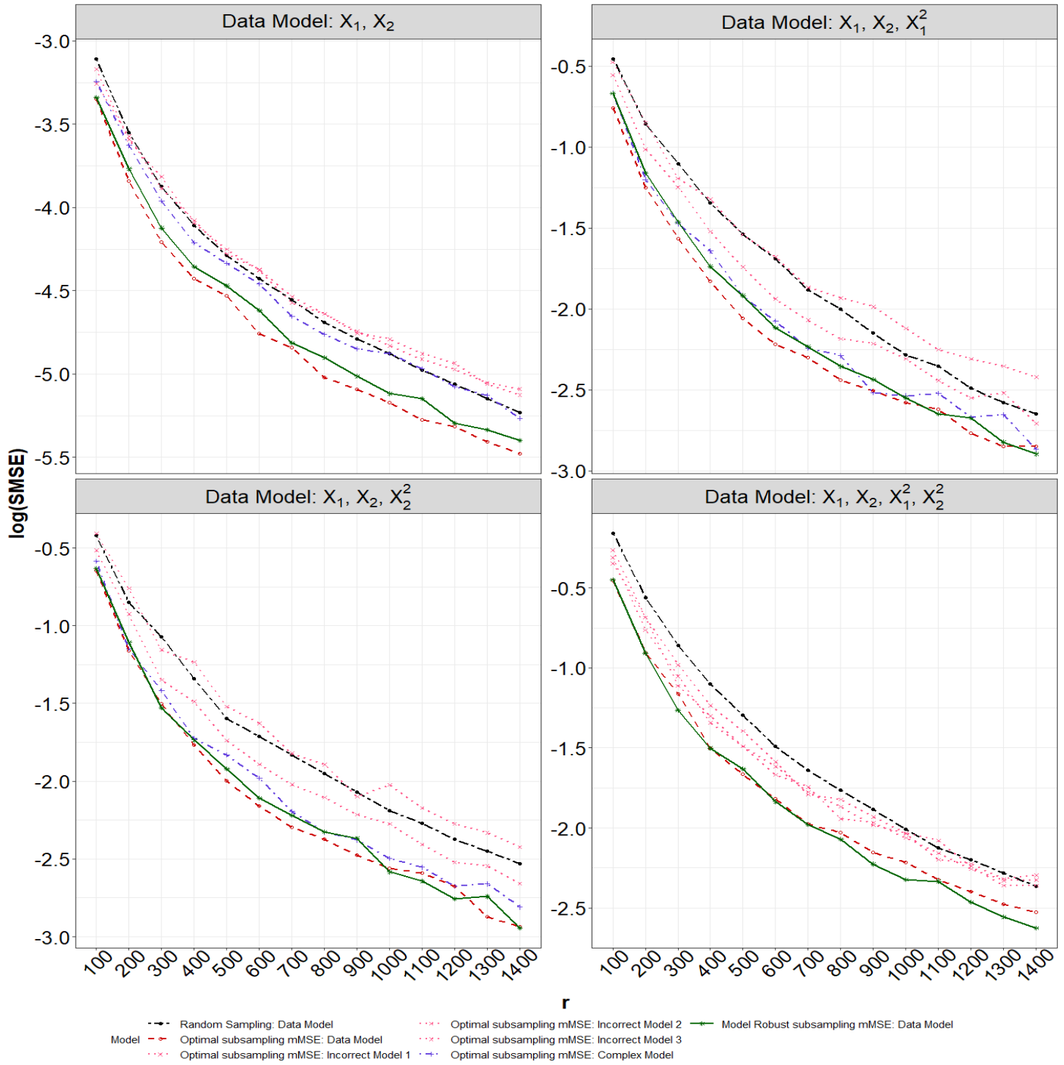


Figure 1: Logarithm of SMSE for the subsampling methods for the poisson regression model ϕ^{mMSE}. Covariate data were generated from the uniform distribution.

Real-world scenario

- In the simulation study, the parameters were specified for the data generating model.
- However, in real-world applications these are unknown.
- Therefore to compare the subsampling methods, for the Q model set, the Summed SMSE (SSMSE) under each subsampling methods can be evaluated as follows:

$$SSMSE(\hat{\theta}_{MLE}) = \frac{1}{M}\sum_{i=1}^{Q}\alpha_q\sum_{m=1}^{M}\sum_{n=1}^{p+t}(\tilde{\theta}_{q_{nm}} - \hat{\theta}_{q,MLE_n}),$$

  where $\tilde{\theta}_q$ is a matrix with the estimated parameters for the q^{th} model over M simulations and $\hat{\theta}_{q,MLE}$ denotes $\hat{\theta}_{MLE}$ for the q^{th} model.
- The set of Q models includes the main effects model, with intercept, and all possible combinations of quadratic terms for continuous covariates with α_q set to $1/Q$.

"New-York City taxi fare" data

- New York City (NYC) taxi trip and fare information of year 2013 consisting of over 170 million records. (Donovan and Work 2016)
- We are interested in how taxi usage varies with day of the week (weekday/weekend), season (winter/spring/summer/autumn), fare amount, and cash payment.
- Each data point includes
  1. y- poisson response variable - the number of rides recorded against the medallion number (a license number provided for a motor vehicle to operate as a taxi),
  2. x_1 - weekday or not,
  3. x_2 - winter or not,
  4. x_3 - spring or not,
  5. x_4 - summer or not,
  6. x_5 - summed fare amount in dollars, and
  7. x_6 - the ratio of cash payment trips in terms of all trips.
- Poisson regression was used to model the relationship between the number of rides per medallion and these covariates.
- Optimal and model robust subsampling probabilities based on A- and L- optimality criterion, that is ϕ^{mMSE} and ϕ^{mV_c} are evaluated, respectively.
- Assign $M = 1000, r_0 = 100, r = 100, 200, \ldots, 1900$ for the subsamples.
- The model set leads to $Q = 4$, consisting of
  1. the main effects model $(x_1, x_2, x_3, x_4, x_5, x_6)$
  2. the main effects model $+ x_5^2$
  3. the main effects model $+ x_6^2$
  4. the main effects model $+ x_5^2 + x_6^2$
- Each of these models were considered equally likely a priori leading to $\alpha_q = 1/4$.
- R-code for the New-York City taxi fare data -

Results

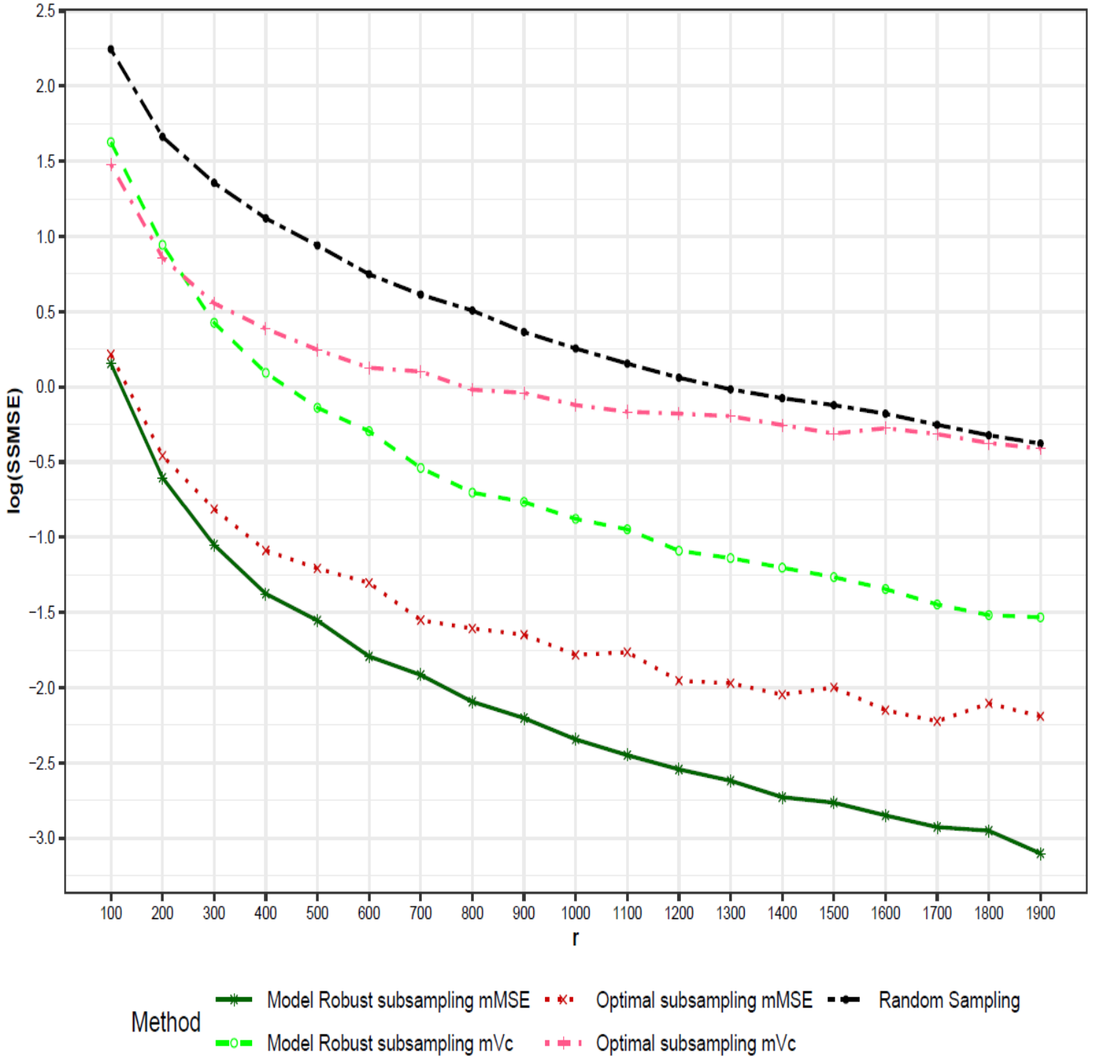


Fig 2: Logarithm of SSMSE over the available models for Poisson regression applied on the 2013 NYC taxi usage data.

- In Figure 2, SSMSE over the four models is shown, where .
  - Random sampling performs the worst.
  - Our proposed model robust approach outperforms the optimal subsampling method for almost all sample sizes for both ϕ^{mMSE} and ϕ^{mVc}.
  - Under ϕ^{mV_c}, the model robust approach actually initially performs worse than the optimal subsampling approach but this is quickly reserved as r is increased.

Discussion

- We proposed a model robust subsampling approach for GLMs.
- This new approach extends the current optimal subsampling approach by considering a set of models (rather than a single model) when determining the subsampling probabilities.
- The robustness properties of this proposed approach were demonstrated in a simulation study, and real-world analysis problem, where it outperformed optimal subsampling and random sampling.

Future research

- Main limitation of our proposed approach is that the specified model set could become quite large as the number of covariates increases.
- One approach to address this issue is by extending the specification of the model set to a flexible class of models.
- Another avenue of interest that could also be explored is determining α_q and/or reducing the model set based on the initial subsample, that is models that clearly do not support the data could be dropped.

References

Ai, Mingyao, Fei Wang, Jun Yu, and Huiming Zhang. 2021. “Optimal subsampling for large-scale quantile regression.” *Journal of Complexity* 62: 101512. https://doi.org/10.1016/j.jco.2020.101512.

Ai, Mingyao, Jun Yu, Huiming Zhang, and HaiYing Wang. 2021. “Optimal subsampling algorithms for big data regressions.” *Statistica Sinica* 31: 749–72. https://doi.org/10.5705/ss.202018.0439.

Donovan, Brian, and Dan Work. 2016. “New york city taxi trip data (2010-2013).” University of Illinois at Urbana-Champaign. https://doi.org/10.13012/J8PN93H8.

Lee, JooChul, Elizabeth D Schifano, and HaiYing Wang. 2021. “Fast optimal subsampling probability approximation for generalized linear models.” *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.02.007.

Meng, Cheng, Rui Xie, Abhyuday Mandal, Xinlian Zhang, Wenxuan Zhong, and Ping Ma. 2021. “LowCon: A design-based subsampling approach in a misspecified linear model.” *Journal of Computational and Graphical Statistics* 30 (3): 694–708. https://doi.org/10.1080/10618600.2020.1844215.

Shi, Chenlu, and Boxin Tang. 2021. “Model-Robust subdata selection for big data.” *Journal of Statistical Theory and Practice* 15 (4): 1–17. https://doi.org/10.1007/s42519-021-00217-9.

Wang, HaiYing, Rong Zhu, and Ping Ma. 2018. “Optimal subsampling for large sample logistic regression.” *Journal of the American Statistical Association* 113 (522): 829–44. https://doi.org/10.1080/01621459.2017.1292914.

Yao, Yaqiong, and HaiYing Wang. 2019. “Optimal subsampling for softmax regression.” *Statistical Papers* 60 (2): 585–99. https://doi.org/10.1007/s00362-018-01068-6.

Yi, Si-Yu, and Yong-Dao Zhou. 2023. “Model-Free Global Likelihood Subsampling for Massive Data.” *Statistics and Computing* 33 (1): 1–16. https://doi.org/10.1007/s11222-022-10185-0.

THANK YOU

Link to the paper.