

Supplementary Materials: Response-Aided Score-Matching Representative Approaches for Big Data Analysis and Model Selection under Generalized Linear Models

Duo Zheng, Keren Li and Jie Yang

S1. Generating Data Blocks

As a stagewise optimization algorithm to conquer big data modeling problems, score-matching representative (SMR) approaches highly depend on the partition of the original data. In the original SMR framework [1], data are partitioned using K -means clustering, an algorithmic approach that clusters the data points based on their distance from each other. K -means clustering is an ideal way to bring informative data blocks to the SMR algorithm since the data points in those data blocks are more similar to each other than the points outside. Intuitively, the harmony inside the data blocks from K -means would make the corresponding representative more convincing since the data points are relatively homogeneous. While the benefit of K -means clustering is significant, there is a price to pay. K -means is computationally expensive especially when the data are massive and the dimension is high. In those cases, the K -means clustering itself would be computationally intractable. To make the proposed RASMR more practically useful, we recommend two alternative solutions to replace K -means clustering.

The first solution is called the *small K -means clustering*, which is essentially the subset clustering strategy recommended by [1]. This strategy starts with drawing a subset of a small proportion of the original data. A small proportion of data can be generated by simple random sampling or stratified sampling (see, for example, [2]) if the original data have some inherent natural structures. Instead of performing K -means clustering on the entire dataset, we implement K -means on a small subset. This subset could be 5%, 10%, or 20% of the original dataset in practice. Once the clusters are produced, the remaining data in the entire dataset can be clustered according to the produced K -means cluster centers. Apparently, the small K -means clustering consumes much less computational resources. Our simulation studies show that—under the seven distributional settings—the small K -means clustering serves the RASMR and SMR approaches almost equally well compared with the full data K -means clustering.

One disadvantage of the small K -means clustering is that it highly depends on the quality of the subset drawn from the entire dataset. If the subset represents the structure of the entire dataset reasonably well, we can expect that the performance of the small K -means would be comparable with the K -means clustering on the full dataset. Nevertheless, if the size of the subset is not large enough to reveal the structure of the entire dataset, or if the subset unfortunately contains some inherent bias, the performance of the RASMR algorithms may not be satisfactory.

The second solution to efficiently generate data blocks is the *correlation-based quantile split* strategy. This approach is similar to the decision tree type of algorithms. Instead of using loss functions, such as cross-entropy loss [3], to determine a variable, and a location to cut, the correlation-based quantile split ranks the covariates or predictors by their correlations with the response in descending order, and then cuts at the median starting from the top of the variable list. This is computationally much cheaper with the time complexity $O(kN \log(N))$, where k is the number of variables to cut and N is the number of data points in the entire dataset. It generates a sufficiently large number of data blocks much faster than the K -means clustering, sometimes almost instantly, which makes it very competitive with a massive data size or a moderate number of variables.

Although the clusters obtained from the correlation-based quantile split approach may not be as elegant or informative as those from K -means clustering, particularly with highly skewed or high-dimensional data, this is not a significant problem as long as the resulting

data blocks effectively serve the RASMR algorithms. Since RASMR typically has a very good convergence rate in most cases, any temporary disadvantages from clustering can be compensated for by increasing the number of RASMR iterations and configuring a suitable learning rate scheduling strategy (see Section 2.5).

Two possible modifications can be made to fulfill specific requirements in practice. First, when the number of variables is not large, the number of data blocks 2^k from a binary cut may not be enough to produce a sufficient amount of representatives. In this case, the binary cut may be replaced with a quantile split using the quartiles Q_1 , Q_2 , and Q_3 (see Section 5.3 for such an example). Another aspect is that, when the dataset contains a large number of variables, one does not need to exhaust the entire list of variables for binary cuts. We can simply go from the top of the variable list ranked by correlations and stop at the k th variable where 2^k is large enough to fit the model.

The advantage of the correlation-based quantile split is its cheap computational expense. It makes this strategy practically useful with a large number of variables. Based on our experience, performing binary cuts on 15 out of 100 total variables is almost instantaneous. In contrast, whether using full data or a small subset, K -means clustering becomes computationally very expensive or intractable when K exceeds 5000. Although the clusters obtained may not be as well separated as those from K -means clustering, this can be quickly compensated by increasing the number of iterations and carefully configuring the learning rate schedule.

These facts put us into a situation where we need to choose between the two model-fitting tactics. That is, whether we fit the model with fewer iterations but finer data blocks, or we fit the model with more iterations with cheaper data blocks. In practice, based on our experience, we suggest that, when the number of variables is small, we may use the small K -means as the clustering approach along with a small number of iterations to achieve a good estimation. As for the situations with a moderately large number of variables, we suggest using the correlation-based quantile split for faster data partitioning with more data blocks and then increasing the number of iterations to obtain a satisfactory result.

S2. More on the Delta Ratio Split

In this section, we provide more discussions and justifications for Algorithm 2 with the delta ratio split.

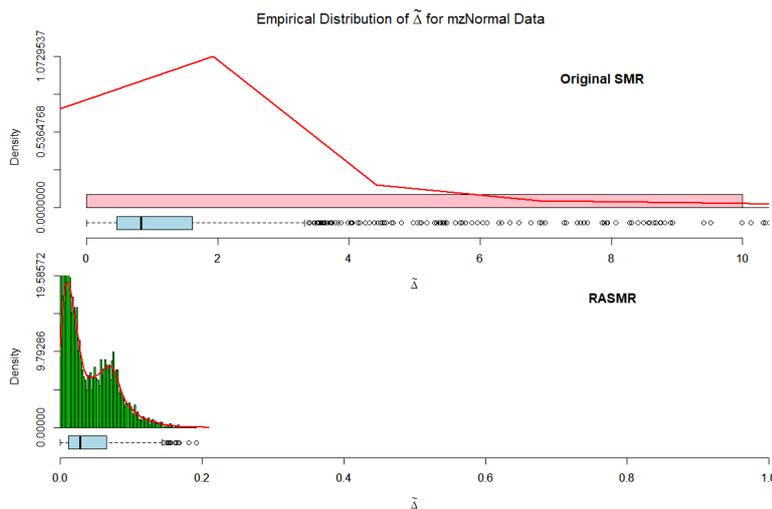


Figure S1. (a) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with mzNormal data.

Figure S1 (a)~(g) show the empirical distributions of the delta ratio $\tilde{\delta}_k$ for simulated data under seven different distribution settings (see Section 4 and Section S5 in the Supplementary Materials for more detailed explanations of the seven distributions). For each subgraph labeled from (a) to (g), the top panel (original SMR) and the bottom panel

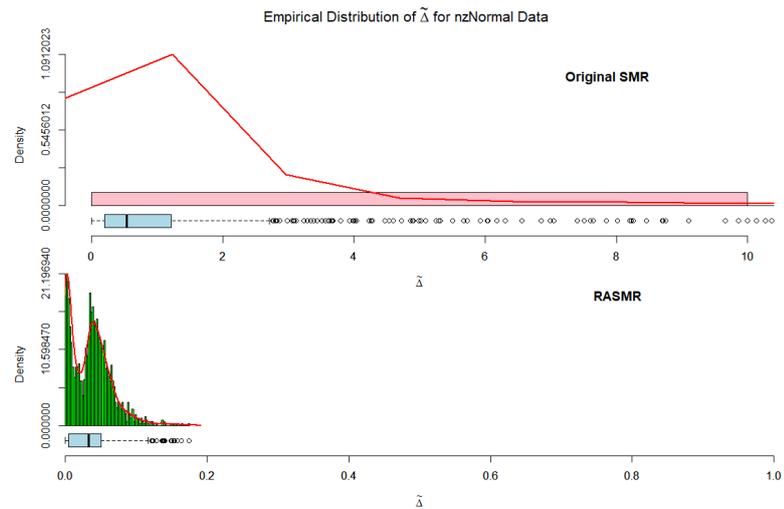


Figure S1. (b) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with nzNormal data

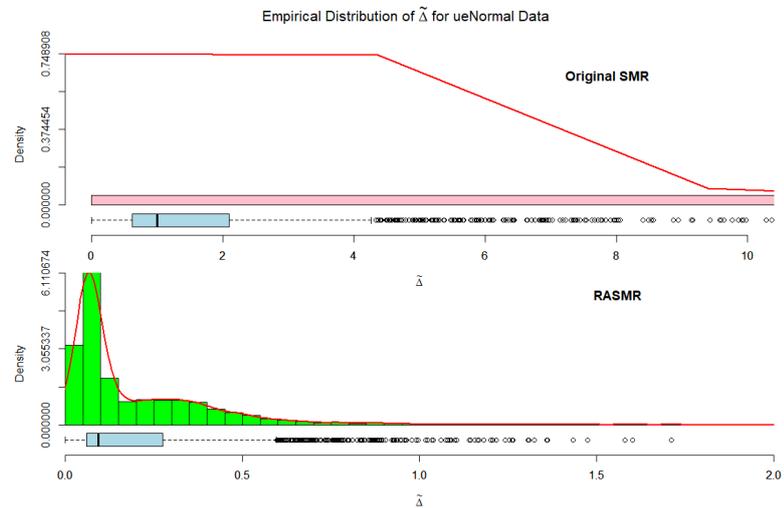


Figure S1. (c) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with ueNormal data

(RASMR) display the histogram, boxplot, and fitted density curves of the 1000 delta ratios based on SMR and RASMR, respectively, calculated from the 1000 data blocks. We observe that most SMR representatives are located outside of their data blocks (that is, $\tilde{\delta}_k > 1$). Moreover, in many cases, the delta ratio could be as large as several hundred, which means that those SMR representatives are dramatically distant from their data blocks. In contrast, RASMR produces representatives that make much more sense. The delta ratios of RASMR representatives are significantly smaller than those of SMR, which explains why RASMR performs better in estimating parameters and approximating the maximum likelihood. Another phenomenon is that the delta ratio tends to inflate when the distribution of variables is more extreme or complicated. For the ueNormal distribution—whose range is tremendously wide—the delta ratios of RASMR representatives spread out a lot, while in other settings, the delta ratios are typically much smaller than 1.

Figure S2 shows four examples of data blocks with large delta ratios from ueNormal data. These data blocks have delta ratios greater than 1, which means that the representatives stay outside their corresponding data blocks. We consider these data blocks underrepresented. From Figure S2, we can see a common feature of those blocks. This

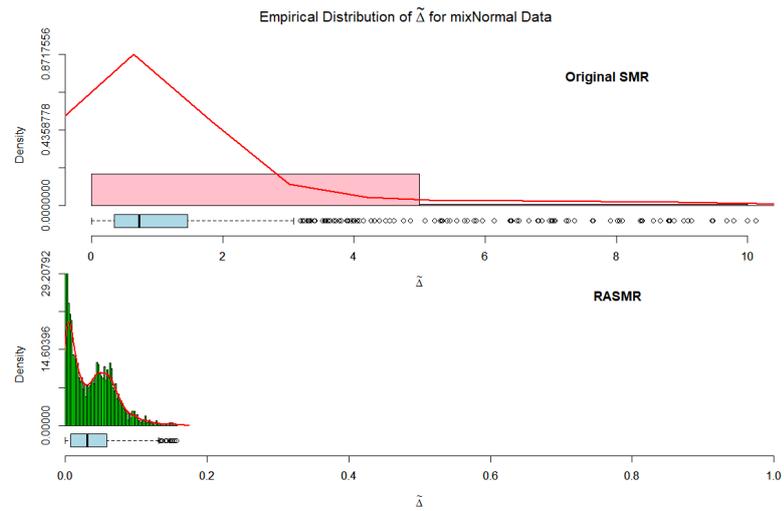


Figure S1. (d) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with mixNormal data

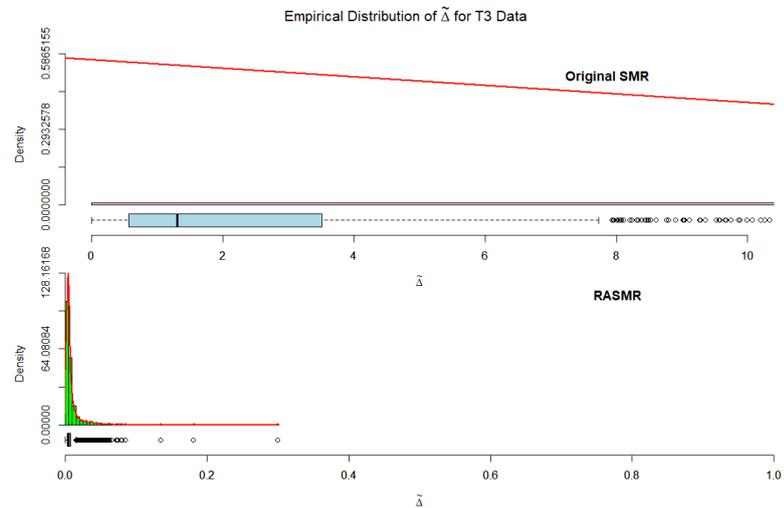


Figure S1. (e) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with T3 data

suggests that Equation (3) almost has a second solution for the data blocks, indicating the potential existence of another representative.

Figure S3 shows four well-represented examples of data blocks, whose delta ratios are small. In those examples, Equation (3) yields a unique solution.

Based on the definition of the delta ratio $\tilde{\delta}_k$, we may use it as a measure of the qualification of a representative to speak for its data block. If the delta ratio is greater than a predefined threshold, we consider the data block to be underrepresented and the representative to lack qualification. Our method for addressing underrepresented data blocks involves further splitting the block based on the assigned threshold. The threshold can be assigned based on the complexity of the distribution of variables. In general, the more complex the structure of the data, the smaller the threshold for delta ratios is preferred. Nevertheless, in practice, if the likelihood approximation is not a serious concern, we may set the threshold equal to one, which ensures the representative stays inside its data block. Once a representative is identified as unqualified, we split its data block at the mean of all data points in the block. That is, we calculate $\eta_i = \mathbf{X}_i \tilde{\beta}^{(t)}$ for each $i \in I_k$ and use the mean $\bar{\eta}_k$ of η_i 's as the cutoff point for the split. Here, we choose the mean of η_i 's to split the data

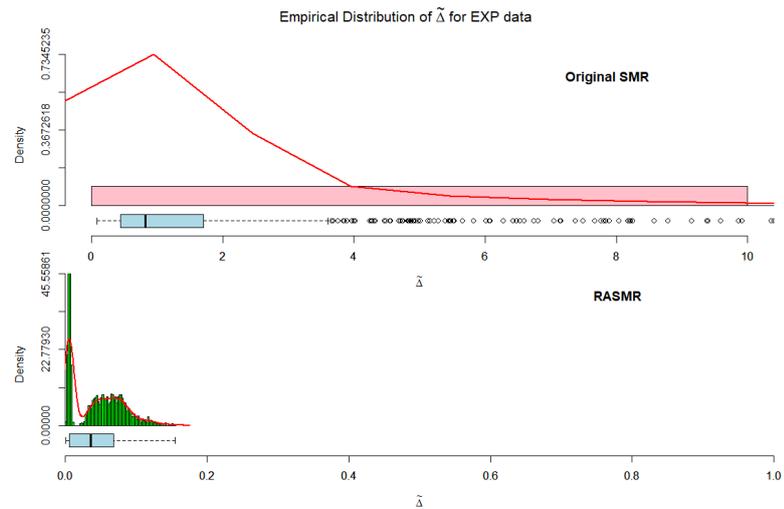


Figure S1. (f) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with EXP data

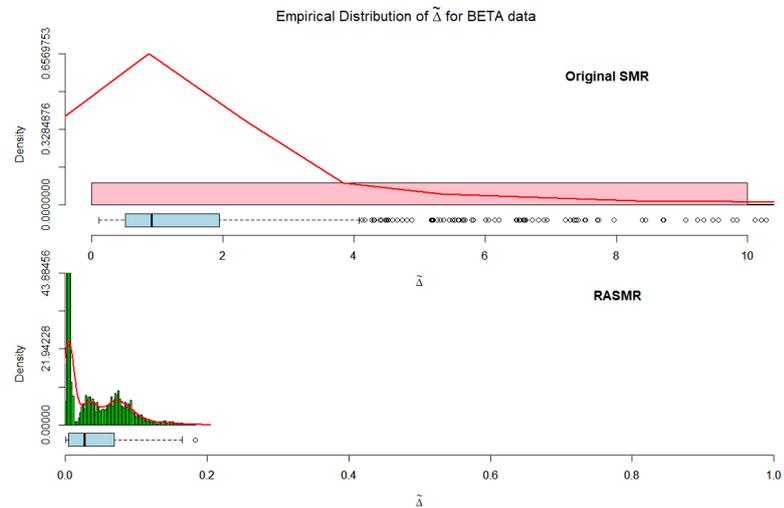


Figure S1. (g) Empirical distribution of the delta ratio $\tilde{\delta}_k$ for SMR and RASMR with BETA data

block over the median because the mean is more sensitive to outliers. Using $\tilde{\eta}_k$ as the split cutoff point enables us to separate the outliers from the majority points in the data block.

S3. More on Link Function Selection

In this section, we provide more technical details to support the discussion in Section 3.2. the objective is to select the most appropriate link function for a GLM given a massive dataset.

In our study, we utilize information-based criteria, such as *AIC* and *BIC*, for the link function selection. The key step is the precise estimation of the maximum likelihood. In Section 3.1, we develop two methods to estimate the maximum likelihood. Theorem1 guarantees the convergence of the estimate using the RASMR-estimated parameters and the full dataset. Theorem2 ensures the convergence of estimates using the RASMR-estimated parameters and representatives.

The first method of calculation involves plugging the parameters estimated from RASMR and the complete dataset into the explicit likelihood function. The \widetilde{AIC} and \widetilde{BIC} , which are estimates of their full-data counterparts, can be readily produced once the likelihood function value is obtained. The main advantage of this method is the outstanding

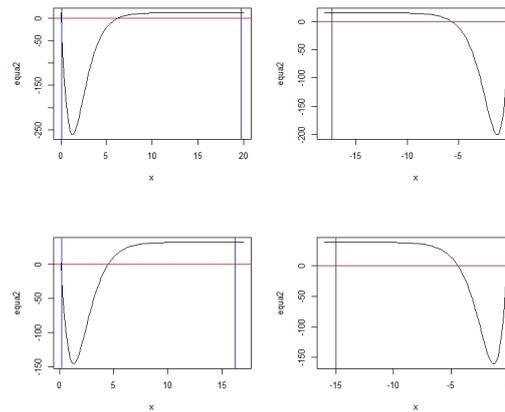


Figure S2. Equation (3) of four data blocks with large delta ratios from ueNormal data.

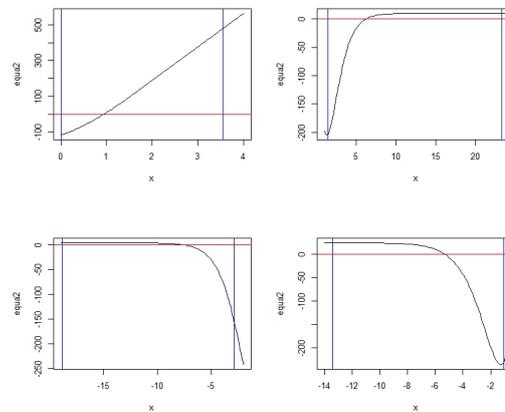


Figure S3. Equation (3) of well-represented data blocks with small delta ratios from ueNormal data.

accuracy of the RASMR estimates. Even in complex scenarios, the estimated parameters by RASMR can still be extremely accurate. However, this method requires an extra plug-in step using the complete dataset, which consumes additional computational resources to process the entire dataset. This can be fairly slow when the data size is ultra-large.

The second method of calculation more elegantly inherits the spirit of representative approaches. It utilizes not only the estimated parameters from RASMR but also the representatives of data blocks. The maximum likelihood can be calculated via an explicit expression of the likelihood function using the parameter estimates and the representatives from RASMR. The advantage of this approach is that the likelihood estimation is integrated with the model-fitting process, which means that the estimation of the full-data information criteria is immediately available once the model-fitting process ends. No additional steps need to be taken to obtain \widetilde{AIC} and \widetilde{BIC} . This saves a considerable amount of time with massive datasets. Despite the computational benefits, \widetilde{AIC} and \widetilde{BIC} may suffer from higher bias and variance than \widehat{AIC} and \widehat{BIC} because uncertainty is introduced into the likelihood calculation from both the parameter estimates and the representatives.

In our numerical experiments, we mainly focus on testing the performances of \widetilde{AIC} and \widetilde{BIC} , as these approaches are more practically useful in big data analysis. Similarly to Example 1, we consider seven extensively used distributions to generate the covariates. The underlying model is a logistic regression model. For each of the 100 simulations, $N = 10^6$ data points are generated and clustered into $K = 1,000$ data blocks by K -means clustering.

Table S1 shows the performance of link function selection using AIC. Selection based on the full data directly delivers perfectly correct rates without any doubt, with the true link

Table S1. Correct rate for selecting the true link function using AIC over 100 simulations.

Data Block Method	K-means (K = 1000)								
	Full data			MR			SMR		
Setting up	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit
mzNormal	100%	100%	100%	100%	100%	100%	100%	100%	0%
nzNormal	100%	100%	100%	100%	100%	100%	74%	77%	100%
ueNormal	100%	100%	100%	100%	0%	100%	100%	8%	0%
mixNormal	100%	100%	100%	100%	100%	100%	100%	81%	100%
T_3	100%	100%	100%	100%	100%	100%	81%	72%	0%
EXP	100%	100%	100%	100%	100%	100%	0%	0%	0%
BETA	100%	100%	100%	100%	99%	100%	0%	6%	0%

Table S2. Correct rate for selecting the link function based on RASMR data blocks

Data Block Representatives	K-means (K = 1000) plus Delta Ratio Split with $\delta_0 = 0.05$								
	Full data			MR			RASMR		
Setting up	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit
mzNormal	100%	100%	100%	100%	100%	100%	100%	100%	100%
nzNormal	100%	100%	100%	100%	100%	100%	100%	100%	100%
ueNormal	100%	100%	100%	100%	100%	100%	100%	100%	100%
mixNormal	100%	100%	100%	100%	100%	100%	100%	100%	100%
T_3	100%	100%	100%	100%	99%	100%	100%	100%	100%
EXP	100%	100%	100%	100%	100%	100%	100%	100%	100%
BETA	100%	100%	100%	100%	100%	100%	100%	100%	100%

function, logit, selected 100% under all circumstances. As a relatively naive representative approach, \widetilde{AIC} with the mean representative (MR) performs surprisingly well, except in the case of Logit vs. Probit with ueNormal distribution, where the covariates exhibit increasingly wild variances. This fact makes MR less favorable when the data structure is complex. For \widetilde{AIC} with the original score-matching representative (SMR) approach, the performance in link function selection is rather disappointing. Under various conditions, SMR fails to choose the logit link function over its competitors. Despite its accurate parameter estimation, SMR is much less reliable for model selection. This unreliability also motivates our proposal of the response-aided score-matching representative (RASMR) approach.

As explained in Sections 2.4 and S2, we recommend using RASMR with the delta ratio split for the link function selection, especially when the data exhibit complex structures. The delta ratio split imposes an additional rule on splitting data blocks to ensure high-quality representatives. As the threshold of the delta ratio split decreases, the estimated maximum likelihood using RASMR parameter estimates and representatives approaches the true maximum likelihood, making \widetilde{AIC} more reliable.

In our numerical experiment, we split the K-means data blocks further based on RASMR with the delta ratio split $\delta_0 = 0.05$. The results are shown in Table S2. We can see that MR achieves a considerable improvement in the correct rate of link function selection based on the RASMR data blocks, while \widetilde{AIC} with RASMR achieves a 100% correct rate for all cases.

For comparison purposes, we also apply a divide-and-conquer (DC) approach with majority voting to conduct the link function selection. For each of the 100 simulations, $K = 1000$ randomly generated data blocks are fed to the DC algorithm for modeling and the link function is selected by majority voting. The results are shown in Table S3, which implies that the DC with majority voting manages to select the true logit link in the first four settings but fails in T_3 , EXP, and BETA.

Table S3. Percentage of candidate link functions chosen by DC and majority voting based on AIC.

Simulation	K = 1000 Randomly Generated Data Blocks			
	Logit	cloglog	Probit	Cauchit
mzNormal	100%	0%	0%	0%
nzNormal	100%	0%	0%	0%
ueNormal	100%	0%	0%	0%
mixNormal	100%	0%	0%	0%
T_3	0%	55%	45%	0%
EXP	0%	100%	0%	0%
BETA	0%	100%	0%	0%

S4. More on Variable Selection Using RASMR

In this section, we provide more technical details to support the discussion in Section 3.3. The objective is to apply RASMR for subset selection and stepwise variable selection for massive data analysis.

When analyzing massive data, traditional statistical methods may not be feasible, including variable selection. The variable selection process requires not only reliable model-fitting capabilities but also highly accurate statistical inference. RASMR provides both extremely accurate estimates of parameters and likelihood function values, making it ideal for facilitating the variable selection task. Nevertheless, we still need to consider several important factors to make the algorithms more applicable and time-efficient in practice.

First, the subset selection process requires fitting candidate models multiple times, which is very time-consuming if each model fitting takes a noticeable amount of time. This fact presents a paradox in the application of RASMR. To make better decisions on variables, more accurate results are needed from RASMR. However, as an iterative optimization algorithm, RASMR requires more iterations to deliver better parameter estimations and high-quality representatives, which can considerably slow down the variable selection process. So, we need to find a balance between processing time and the performance of variable selection. We address this problem by introducing a fast variable screening process with the mean representatives (MRs). That is, the MR is first implemented in the forward selection process (see, for example, [3]) to obtain an initial result of the variable selection. Denoted by AIC and BIC , the information criteria are estimated by using the mean representative and the corresponding estimates of parameters. We continue the variable selection process until there is no further reduction of \overline{AIC} or \overline{BIC} is detected. Recall the experiment results listed in Table S1. MR performs sufficiently well in the link function selection using information-based criteria. Based on our experiences, MR preserves the order of likelihood and makes correct decisions based on information criteria in many cases. For this reason, we recommend MR for a quick initial screening in variable selection. Then, we may implement a forward stepwise selection process using RASMR for a finer selection.

Another important aspect is the data clustering strategy. Prior to the implementation of RASMR, the data blocks need to be provided. Two data clustering techniques can be used to generate the data blocks. They are K -means clustering and correlation-based binary cut (see also Section S1). The K -means clustering generates more elegant data blocks but also consumes more computational resources. The correlation-based binary cut (that is, cut at the mean value of each variable) is computationally much cheaper but would be less informative given the same number of clusters. However, this shortcoming can be accommodated by increasing the number of clusters. According to their unique advantages, respectively, we propose two solutions to generate the data blocks for implementing RASMR in variable selection.

Solution one: We perform a one-time partition on the entire dataset using K -means clustering. All the candidate models in the variable selection process are fitted with RASMR using this universal partition.

Solution two: In each step of the forward stepwise selection using RASMR, we perform the correlation-based binary cut on the current set of active variables. For example, at the t th step, given the currently active set V_t of variables, we perform the correlation-based binary cut on $V_t \cup \{i\}$ for some $i \notin V_t$ to move forward, or on $V_t \setminus \{j\}$ for some $j \in V_t$ to move backward.

Solution one is ideal for handling variable selection tasks for big data with a small number of variables. A well-tuned group of data blocks from K -means using the complete set of variables would be sufficient to facilitate the RASMR model-fitting process by providing a good global view of the data structure. Solution two is recommended to handle variable selection tasks for big data with a moderately large number of variables.

Let the total number of covariates or predictors under selection be p . Denote the index set of all variables under selection by \mathcal{V} and denote the index set of the selected variables

at the t th step by V_t . For illustration purposes, the practical algorithm using the global K -means clusters for variable selection with RASMR is described as follows:

1. Generate data blocks using K -means based on the complete set \mathcal{V} of variables.
2. Do variable screening using the forward selection with the mean representative.
 - (a) Generate the mean representative for each data block.
 - (b) Initiate $V_0 = \emptyset$. For each variable $x_i, i \in \mathcal{V}$, evaluate the \overline{AIC} or \overline{BIC} for the corresponding univariate model based on x_i with the mean representative and the estimates of parameters. Suppose the univariate model with x_{i_*} attains the smallest AIC or BIC . Let $V_1 = \{i_*\}$.
 - (c) At the $(t + 1)$ th step, for each $i \notin V_t$, evaluate the \overline{AIC} or \overline{BIC} of the model based on variables $\{x_j, j \in V_t \cup \{i\}\}$ with the mean representative and the estimated parameters.
 - (d) Suppose $x_{i_*}, i_* \notin V_t$ attains the smallest \overline{AIC} or \overline{BIC} in the $(t + 1)$ th step. If the smallest \overline{AIC} or \overline{BIC} in the $(t + 1)$ th step is strictly less than the \overline{AIC} or \overline{BIC} at the t th step, let $V_{t+1} = V_t \cup \{i_*\}$. Otherwise, stop the selection process and settle with $\{x_i, i \in V_t\}$.
3. Do stepwise variable selection using RASMR.
 - (a) Initiate V'_0 to be the index set of the selected variables from the variable screening step.
 - (b) At the $(t + 1)$ th step, given that $V'_t \neq \mathcal{V}$, to move forward, find the best variable $x_{i_*}, i_* \notin V'_t$ that minimizes \widetilde{AIC} or \widetilde{BIC} of the model based on variables $\{x_i, i \in V'_t \cup \{i_*\}\}$ and RASMR.
 - (c) At the $(t + 1)$ th step, given that $V'_t \neq \emptyset$, to move backward, find the best variable $x_{j_*}, j_* \in V'_t$ that minimizes \widetilde{AIC} or \widetilde{BIC} of the model based on variables $\{x_j, j \in V'_t \setminus \{j_*\}\}$ and RASMR.
 - (d) Compare the smallest \widetilde{AIC} or \widetilde{BIC} obtained in (b) and (c) with their counterparts based on V'_t . If a strict reduction in \widetilde{AIC} or \widetilde{BIC} is detected, update V'_{t+1} accordingly and go to the $(t + 2)$ th step. Otherwise, stop the selection process and settle with the variable index set V'_t .

S5. More Simulation Studies

As described in Section 4, the feature variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ are generated from one of the following seven distributions (see also [1]):

- (1) **mzNormal** $N_d(0, \Sigma)$, where Σ is a $d \times d$ matrix with diagonal 1 and off-diagonal 0.5.
- (2) **nzNormal** $N_d(1.5, \Sigma)$, which leads to imbalanced responses.
- (3) **ueNormal** $N_d(0, \Sigma_u)$, where Σ_u is a $d \times d$ matrix with diagonals $1^2, \dots, d^2$ and off-diagonal 0.5.
- (4) **mixNormal** $0.5N_d(-1, \Sigma) + 0.5N_d(1, \Sigma)$, representing a bimodal case.
- (5) **T₃** multivariate t with 3 degrees of freedom $t_3(0, \Sigma)/10$, which is heavy-tailed.
- (6) **EXP** IID exponential distribution with a rate parameter $\lambda = 2$, which has a heavy tail on the right.
- (7) **BETA** IID $Beta(\alpha = 0.5, \beta = 0.5)$, which is a bounded and U-shaped.

Note that in distributions (1) mzNormal, (3) ueNormal, (4) mixNormal, and (5) T_3 , the distribution of each component of \mathbf{x}_i is symmetric about zero. If $\beta_0 = 0$, then the linear predictor η_i defined in Model (8) (see Section 4.1) also has a distribution that is symmetric about zero.

S5.1. Comparing Different Variable Selection Strategies

In this section, we first test the performances of variable selection strategies using the mean representative (MR) approach only. Consider a logistic regression model with its linear predictor, as follows:

$$\eta_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i,20}\beta_{20}.$$

Suppose the true parameter values are $\beta_i = 0.5$, for $i = 1, 2, \dots, 7$, and $\beta_j = 0$ for $j = 8, \dots, 20$. This configuration matches the true model described in Section 4.

The predictors x_i are generated from one of the seven distributions mentioned earlier. The average number of variables selected, the average false positive rate (FPR), the average true positive rate (TPR), and their corresponding standard deviations are recorded and shown in Table S4.

Table S4. Variable selection comparison between representative approaches and others

Simulation Method	MR for Variable Selection with K-means Clustering and AIC								
	Forward			Backward			Stepwise		
	No. Selected	FPR	TPR	No. Selected	FPR	TPR	No. Selected	FPR	TPR
mzNormal	7.4(0.70)	2.9%(0.05)	100%(0)	9.3(1.33)	16.4%(0.096)	100%(0)	7.4(0.70)	2.9%(0.05)	100%(0)
nzNormal	8.1(0.57)	7.9%(0.041)	100%(0)	10.1(1.52)	22.1%(0.109)	100%(0)	8.1(0.57)	7.9%(0.041)	100%(0)
ueNormal	8.5(1.08)	10.7%(0.077)	100%(0)	10.9(1.73)	27.8%(0.123)	100%(0)	8.5(1.08)	10.7%(0.077)	100%(0)
mixNormal	8.1(0.88)	7.9%(0.063)	100%(0)	9.8(0.92)	20.0%(0.066)	100%(0)	8.1(0.88)	7.9%(0.063)	100%(0)
T_3	7.2(0.98)	1.4%(0.03)	100%(0)	9.4(1.17)	17.1%(0.084)	100%(0)	7.2(0.98)	1.4%(0.03)	100%(0)
EXP	13.0(1.63)	42.8%(0.117)	100%(0)	14.4(1.64)	52.9%(0.118)	100%(0)	13.0(1.63)	42.8%(0.12)	100%(0)
BETA	9.2(1.03)	15.7%(0.074)	100%(0)	8.9(1.52)	13.6%(0.109)	100%(0)	9.2(1.03)	15.7%(0.074)	100%(0)

(a) Ave (Std) of FPR and TPR of variable selection using MR and AIC.

Simulation Method	MR for Variable Selection with K-means Clustering and BIC								
	Forward			Backward			Stepwise		
	No. Selected	FPR	TPR	No. Selected	FPR	TPR	No. Selected	FPR	TPR
mzNormal	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)
nzNormal	7.1(0.32)	0.7%(0.023)	100%(0)	7.3(0.48)	2.1%(0.035)	100%(0)	7.1(0.32)	0.7%(0.023)	100%(0)
ueNormal	7.2(0.42)	1.4%(0.030)	100%(0)	7.2(0.42)	1.4%(0.030)	100%(0)	7.2(0.42)	1.4%(0.030)	100%(0)
mixNormal	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)	7.0(0.14)	0%(0)	100%(0)
T_3	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)
EXP	8.1(0.32)	7.9%(0.023)	100%(0)	7.1(0.32)	0.71%(0.023)	100%(0)	7.1(0.32)	0.7%(0.023)	100%(0)
BETA	8.0(0)	7.1%(0)	100%(0)	7.0(0)	0%(0)	100%(0)	7.0(0)	0%(0)	100%(0)

(b) Ave (Std) of FPR and TPR of Variable Selection Using MR and BIC

Simulation Method	MR+RASMR for variable selection with K-means clustering and AIC								
	Forward MR Screening + Stepwise RASMR			Backward MR Screening + Stepwise RASMR			Stepwise MR Screening + Stepwise RASMR		
	No. Selected	FPR	TPR	No. Selected	FPR	TPR	No. Selected	FPR	TPR
mzNormal	7.0(0)	0%(0)	100%(0)	9.0(1.33)	14.3%(0.095)	100%(0)	7.0(0)	0%(0)	100%(0)
nzNormal	7.4(0.84)	2.9%(0.060)	100%(0)	9.1(1.66)	15.0%(0.119)	100%(0)	7.4(0.84)	2.8%(0.060)	100%(0)
ueNormal	8.0(0.82)	7.1%(0.058)	100%(0)	9.1(1.37)	15.0%(0.098)	100%(0)	8.0(0.82)	7.1%(0.058)	100%(0)
mixNormal	8.0(0.94)	7.1%(0.067)	100%(0)	9.7(1.06)	19.2%(0.076)	100%(0)	8.0(0.94)	7.1%(0.067)	100%(0)
T_3	7.1(0.32)	0.71%(0.023)	100%(0)	8.9(1.29)	13.6%(0.092)	100%(0)	7.0(0)	0%(0)	100%(0)
EXP	10.8(2.53)	27.1%(0.181)	100%(0)	11.0(1.94)	28.6%(0.139)	100%(0)	10.8(2.53)	27.1%(0.181)	100%(0)
BETA	8.7(1.06)	12.1%(0.076)	100%(0)	7.9(0.99)	6.4%(0.071)	100%(0)	8.0(0)	7.1%(0.067)	100%(0)

(c) Ave (Std) of FPR and TPR of variable selection using MR + RASMR and AIC.

Simulation Configuration	Penalized Logistic Regression with Majority Vote					
	$K=1000, w=500$			$K=20, w=10$		
	No. Selected	FPR	TPR	No. Selected	FPR	TPR
mzNormal	7(0)	0%(0)	100%(0)	7(0)	0%(0)	100%(0)
nzNormal	0(0)	0%(0)	0%(0)	7(0)	0%(0.02)	100%(0)
ueNormal	7(0)	0%(0)	100%(0)	9.67(1.64)	19.1%(0.117)	100%(0)
mixNormal	6.43(1.58)	0%(0)	91.2%(0.23)	20(0)	100%(0)	100%(0)
T_3	0(0)	0%(0)	0(0)	0.02(0.14)	0%(0)	0.3%(0.02)
EXP	0(0)	0%(0)	0(0)	0(0)	0%(0)	0%(0)
BETA	0(0)	0%(0)	0(0)	0(0)	0%(0)	0%(0)

(d) Ave (Std) of FPR and TPR of penalized logistic regression with majority vote, 1000 data blocks.

The forward selection strategy with the mean representative is implemented to select informative variables from the 20 variables in total. Estimated information criteria AIC and BIC based on MR are tested separately. The results are shown in Tables S4a and S4b. We can see that MR performs extremely well in controlling the true positive rate (TPR). MR does not miss any informative variables under all circumstances in the numerical experiments.

Table S7. Correct rate over 100 simulations based on RASMR for cross-validation.

Data Blocks Representatives	K-means ($K = 1000$) plus RASMR		
	RASMR		
Setting up	Logit VS Cloglog	Logit VS Probit	Logit VS Cauchit
mzNormal	100%	100%	100%
nzNormal	100%	100%	100%
ueNormal	100%	100%	100%
mixNormal	100%	100%	100%
T_3	100%	100%	100%
EXP	100%	100%	100%
BETA	100%	100%	100%

S6. Proofs and Relevant Lemmas

Proof of Theorem 1: The major difference between a RASMR algorithm (Algorithm 1 or 2) and the original SMR algorithm is the carefully designed response-aided partition scheme, making the RASMR algorithm stabler, more accurate, and practical. But fundamentally they are derived to solve the same equation, the score-matching equation of η . Theorem 3.2 in [1] has established the convergence property of the SMR algorithm, which is essentially held for RASMR as well. That is, $\|\tilde{\beta} - \hat{\beta}\| = O(\tilde{\Delta}^{1/2})$. Since $l(\beta; \mathbf{y}, \mathbf{X})$ is twice differentiable in C , then $\|\nabla l(\beta; \mathbf{y}, \mathbf{X})\| \leq M$ for all $\beta \in C$, where $M > 0$ is some constant. Since $l(\tilde{\beta}; \mathbf{y}, \mathbf{X}) - l(\hat{\beta}; \mathbf{y}, \mathbf{X}) = \nabla l(c\tilde{\beta} + (1-c)\hat{\beta}; \mathbf{y}, \mathbf{X}) \cdot (\tilde{\beta} - \hat{\beta})$ for some $c \in (0, 1)$, according to the mean value theorem with several variables, the convergence rate of the likelihood function $l(\tilde{\beta}; \mathbf{y}, \mathbf{X})$ to $l(\hat{\beta}; \mathbf{y}, \mathbf{X})$ is the same to the convergence rate of $\tilde{\beta}$ to $\hat{\beta}$ only up to a constant M . So, we have $l(\tilde{\beta}; \mathbf{y}, \mathbf{X}) - l(\hat{\beta}; \mathbf{y}, \mathbf{X}) = O(\tilde{\Delta}^{1/2})$ as $\tilde{\Delta} \rightarrow 0$. \square

Proof of Theorem 2: Follow the notation in Section 2.5 of [5], a general form of the log-likelihood function of a generalized linear model can be expressed as follows:

$$l(\beta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^N \left[\frac{y_i \theta(\mathbf{X}_i^T \beta) - b(\theta(\mathbf{X}_i^T \beta))}{a(\phi)} + c(y_i, \phi) \right],$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are known functions, ϕ is the dispersion parameter, and $\theta(\cdot) = (b')^{-1}(g^{-1}(\cdot))$. For the k th data block individually, the exact log-likelihood contribution based on the full data set is as follows:

$$l_k(\beta) = \sum_{i \in I_k} a(\phi)^{-1} \left[y_i \theta(\mathbf{X}_i^T \beta) - b(\theta(\mathbf{X}_i^T \beta)) \right],$$

while the contribution of the k th representative $(n_k, \tilde{\mathbf{X}}_k^T, \tilde{y}_k)$ to the log-likelihood $l(\tilde{\beta}; \tilde{y}, \tilde{\mathbf{X}})$ based on the representative data points is as follows:

$$\begin{aligned} \tilde{l}_k(\beta) &= n_k a(\phi)^{-1} \left[\tilde{y}_k \theta(\tilde{\mathbf{X}}_k^T \beta) - b(\theta(\tilde{\mathbf{X}}_k^T \beta)) \right] \\ &= a(\phi)^{-1} \left(n_k \tilde{y}_k - \sum_{i \in I_k} y_i \right) \theta(\tilde{\eta}_k) + \sum_{i \in I_k} a(\phi)^{-1} \left[y_i \theta(\tilde{\mathbf{X}}_k^T \beta) - b(\theta(\tilde{\mathbf{X}}_k^T \beta)) \right] \\ &= \sum_{i \in I_k} a(\phi)^{-1} \left[y_i \theta(\tilde{\mathbf{X}}_k^T \beta) - b(\theta(\tilde{\mathbf{X}}_k^T \beta)) \right]. \end{aligned}$$

since $\tilde{y}_k = n_k^{-1} \sum_{i \in I_k} y_i$.

We compare the absolute difference between $\tilde{l}_k(\boldsymbol{\beta})$ and $l_k(\boldsymbol{\beta})$ using the Cauchy–Schwarz inequality. The \tilde{l}_k is approximated by its first-order Taylor expansion about $\tilde{\mathbf{X}}_k$ at \mathbf{X}_i .

$$\begin{aligned} |\tilde{l}_k(\boldsymbol{\beta}) - l_k(\boldsymbol{\beta})| &= \left| \sum_{i \in I_k} \left\{ \left[(y_i - G(\mathbf{X}_i^T \boldsymbol{\beta})) \nu(\mathbf{X}_i^T \boldsymbol{\beta}) \right] (\tilde{\mathbf{X}}_k - \mathbf{X}_i)^T \boldsymbol{\beta} + o(\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\|) \right\} \right| \\ &\leq \left(\sum_{i \in I_k} (y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))^2 \nu(\mathbf{X}_i^T \boldsymbol{\beta})^2 \cdot \sum_{i \in I_k} \|\tilde{\mathbf{X}}_k - \mathbf{X}_i\|^2 \cdot \|\boldsymbol{\beta}\|^2 \right)^{1/2} + \sum_{i \in I_k} o(\|\tilde{\mathbf{X}}_k - \mathbf{X}_i\|) \\ &\leq n_k \tilde{\Delta} \|\boldsymbol{\beta}\| \left(n_k^{-1} \sum_{i \in I_k} (y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))^2 \nu(\mathbf{X}_i^T \boldsymbol{\beta})^2 \right)^{1/2} + \sum_{i \in I_k} o(\tilde{\Delta}). \end{aligned}$$

Denote $F_k = (n_k^{-1} \sum_{i \in I_k} (y_i - G(\mathbf{X}_i^T \boldsymbol{\beta}))^2 \nu(\mathbf{X}_i^T \boldsymbol{\beta})^2)^{1/2}$. For all $\boldsymbol{\beta} \in C$ and $\tilde{\Delta}$ sufficiently small, we obtain the following:

$$\begin{aligned} |\tilde{l}(\boldsymbol{\beta}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}) - l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})| &\leq \sum_{k=1}^K n_k \tilde{\Delta} \|\boldsymbol{\beta}\| F_k + \sum_{i=1}^N o(\tilde{\Delta}) \\ &\leq N \tilde{\Delta} \|\boldsymbol{\beta}\| \cdot \max_k F_k + N \cdot o(\tilde{\Delta}) \\ &\leq M \tilde{\Delta} \end{aligned} \tag{S6.1}$$

for some number $M > 0$, which is not related to the representatives or $\tilde{\Delta}$ but relies on the data. In conclusion, for $\boldsymbol{\beta}$, in the compact set C , $\tilde{l}(\boldsymbol{\beta}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ converges to the full log-likelihood function $l(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X})$ uniformly, as $\tilde{\Delta}$ goes to 0. Specifically, for a small enough $\tilde{\Delta}$,

$$|\tilde{l}(\tilde{\boldsymbol{\beta}}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}) - l(\tilde{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X})| \leq M \tilde{\Delta}.$$

According to Theorem 1, $l(\tilde{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}) - l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}) = O(\tilde{\Delta}^{1/2})$ as $\tilde{\Delta} \rightarrow 0$, then we have the following:

$$\begin{aligned} &|\tilde{l}(\tilde{\boldsymbol{\beta}}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}) - l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X})| \\ &\leq |\tilde{l}(\tilde{\boldsymbol{\beta}}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}) - l(\tilde{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X})| + |l(\tilde{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}) - l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X})| \\ &\leq M \tilde{\Delta} + O(\tilde{\Delta}^{1/2}) \\ &= O(\tilde{\Delta}^{1/2}). \end{aligned}$$

That is, $\tilde{l}(\tilde{\boldsymbol{\beta}}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}) - l(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}) = O(\tilde{\Delta}^{1/2})$ as $\tilde{\Delta}$ goes to 0. \square

Proof of Theorem A1: According to the arguments in the context of Equation (5) in Section 2.2, (3) is equivalent to (5). Let $\bar{S} = n_k^{-1} \sum_{i \in I_k} S(\eta_i)$ be the right hand of (5). Since $S(\eta)$ is strictly monotone on $[\eta_k^\wedge, \eta_k^\vee]$, then \bar{S} is between $S(\eta_k^\wedge)$ and $S(\eta_k^\vee)$. The continuity of $\nu(\eta)$ and $G(\eta)$ implies that $S(\eta)$ is continuous on $[\eta_k^\wedge, \eta_k^\vee]$. By the intermediate value theorem, there must exist an $\eta_* \in [\eta_k^\wedge, \eta_k^\vee]$ that solves (5). Since $S(\eta)$ is also strictly monotone, such an η_* is unique. \square

Proof of Theorem A2: For a GLM with $Y_i \sim \text{Normal}(\mu_i, \sigma^2)$ with the identity link $g(\mu_i) = \mu_i$, which is essentially a linear regression model, we have $G(\eta) = \eta$ and

$$S'(\eta) = \tilde{y}_k - 2\eta \begin{cases} > 0, & \eta < \frac{1}{2}\tilde{y}_k; \\ = 0, & \eta = \frac{1}{2}\tilde{y}_k; \\ < 0, & \eta > \frac{1}{2}\tilde{y}_k. \end{cases}$$

Apparently, $S(\eta)$ is strictly monotone on $[\eta_k^\wedge, \eta_k^\vee]$ if $\frac{1}{2}\tilde{y}_k \notin (\eta_k^\wedge, \eta_k^\vee)$. That is, $\tilde{\eta}_k$ is unique if $\frac{1}{2}\tilde{y}_k \notin (\eta_k^\wedge, \eta_k^\vee)$. Otherwise, there might be one $\tilde{\eta}_k \in [\eta_k^\wedge, \frac{1}{2}\tilde{y}_k]$ and one $\tilde{\eta}_k \in [\frac{1}{2}\tilde{y}_k, \eta_k^\vee]$ solving (3).

Suppose the index block I_k is obtained after the first split ($\eta_i < 0$ or $\eta_i > 0$) and the second split ($y_i > G(\eta_i)$ or $y_i < G(\eta_i)$). Recall that $\bar{\eta}_k = n_k^{-1} \sum_{i \in I_k} \eta_i$ and $\bar{S} = n_k^{-1} \sum_{i \in I_k} S(\eta_i)$. Then, there are up to two solutions solving $S(\eta) = \bar{S}$, or equivalently, (3):

$$\bar{\eta}_{k,1} = \frac{1}{2} \left(\bar{y}_k - \sqrt{\bar{y}_k^2 - 4\bar{S}} \right), \quad \bar{\eta}_{k,2} = \frac{1}{2} \left(\bar{y}_k + \sqrt{\bar{y}_k^2 - 4\bar{S}} \right).$$

Note that we always have $\bar{\eta}_{k,1} \leq \bar{y}_k/2 \leq \bar{\eta}_{k,2}$.

Case (i): $\eta_i > 0$ and $y_i < G(\eta_i)$ for all $i \in I_k$. We claim that the solution to (5) is unique in this case. Moreover, note that $G(\eta_i) = \eta_i$ in this model. Then, $0 < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < \infty$,

$$\bar{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} < \frac{\sum_{i \in I_k} \eta_i^2}{\sum_{i \in I_k} \eta_i} = \frac{\sum_{i \in I_k} \eta_i^2}{n_k \bar{\eta}_k},$$

and

$$\bar{S} = \frac{1}{n_k} \sum_{i \in I_k} S(\eta_i) = \frac{1}{n_k} \sum_{i \in I_k} (\bar{y}_k - \eta_i) \eta_i = \bar{y}_k \cdot \bar{\eta}_k - \frac{1}{n_k} \sum_{i \in I_k} \eta_i^2 < 0.$$

If $\bar{y}_k \leq 0$, then $S(\eta)$ is strictly monotone decreasing on $[\eta_k^\wedge, \eta_k^\vee]$ and $\bar{\eta}_k = \bar{\eta}_{k,2} \in [\eta_k^\wedge, \eta_k^\vee]$ is the only solution. If $\bar{y}_k > 0$, then $S(\eta) \geq 0$ when $\eta \in [0, \frac{1}{2}\bar{y}_k]$, there is no solution of $\bar{\eta}_k \in [\eta_k^\wedge, \frac{1}{2}\bar{y}_k]$ solving $S(\bar{\eta}_k) = \bar{S}$, or equivalently, (3). Therefore, $\bar{\eta}_k \in [\eta_k^\wedge, \eta_k^\vee]$ solving (3) is unique, which is $\bar{\eta}_{k,2} \in (\frac{1}{2}\bar{y}_k, \eta_k^\vee]$.

In summary, in this case, $\bar{\eta}_k = \bar{\eta}_{k,2}$ is the only solution.

Case (ii): $\eta_i > 0$ and $y_i > G(\eta_i)$ for all $i \in I_k$. In this case, (5) may not be unique in $[\eta_k^\wedge, \eta_k^\vee]$. Moreover, we have $0 < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < \infty$,

$$\bar{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} > \frac{\sum_{i \in I_k} \eta_i^2}{\sum_{i \in I_k} \eta_i} \geq \frac{n_k \bar{\eta}_k^2}{n_k \bar{\eta}_k} = \bar{\eta}_k,$$

and $\bar{S} > 0$. On the other hand, $\bar{S} < S(\bar{\eta}_k)$ since $S(\eta)$ is strictly concave. That is, we have $0 < \eta_k^\wedge \leq \bar{\eta}_k < \bar{y}_k$ and $0 < \bar{S} < S(\bar{\eta}_k)$. Then there exists an $\tilde{\eta}_k \in (\bar{\eta}_k, \bar{y}_k)$ that solves $S(\tilde{\eta}_k) = \bar{S}$. If $\eta_k^\wedge \geq \bar{y}_k - \bar{\eta}_k$, then $\tilde{\eta}_k \in (\bar{\eta}_k, \bar{y}_k)$ is the unique solution in $[\eta_k^\wedge, \eta_k^\vee]$. Otherwise, there might be another solution between η_k^\wedge and $\bar{y}_k - \bar{\eta}_k$.

Case (iii): $\eta_i < 0$ and $y_i > G(\eta_i)$ for all $i \in I_k$. We claim that the solution to (5) is unique in this case. Moreover, $-\infty < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < 0$,

$$\bar{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} > \frac{\sum_{i \in I_k} \eta_i^2}{\sum_{i \in I_k} \eta_i} = \frac{\sum_{i \in I_k} \eta_i^2}{n_k \bar{\eta}_k},$$

and

$$\bar{S} = \bar{y}_k \cdot \bar{\eta}_k - \frac{1}{n_k} \sum_{i \in I_k} \eta_i^2 < 0.$$

If $\bar{y}_k \geq 0$, then $S(\eta)$ is strictly monotone increasing on $[\eta_k^\wedge, \eta_k^\vee]$ and $\bar{y}_k = \bar{\eta}_{k,1} \in [\eta_k^\wedge, \eta_k^\vee]$ is the only solution. If $\bar{y}_k < 0$, then $S(\eta) \geq 0$ when $\eta \in [\frac{1}{2}\bar{y}_k, 0]$, there is no solution of $\bar{\eta}_k \in [\frac{1}{2}\bar{y}_k, 0]$ solving $S(\bar{\eta}_k) = \bar{S}$, or equivalently, (3). Therefore, $\bar{\eta}_k \in [\eta_k^\wedge, \eta_k^\vee]$ solving (3) is unique, which is $\bar{\eta}_{k,1} \in [\eta_k^\vee, \frac{1}{2}\bar{y}_k)$.

In summary, in this case, $\bar{\eta}_k = \bar{\eta}_{k,1}$ is the only solution.

Case (iv): $\eta_i < 0$ and $y_i < G(\eta_i)$ for all $i \in I_k$. In this case, (5) may not be unique in $[\eta_k^\wedge, \eta_k^\vee]$. Moreover, we have $-\infty < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < 0$,

$$\bar{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} < \frac{\sum_{i \in I_k} \eta_i^2}{\sum_{i \in I_k} \eta_i} \leq \frac{n_k \bar{\eta}_k^2}{n_k \bar{\eta}_k} = \bar{\eta}_k < 0,$$

and $\bar{S} > 0$. On the other hand, $\bar{S} < S(\bar{\eta}_k)$ since $S(\eta)$ is strictly concave. That is, we have $-\infty < \bar{y}_k < \bar{\eta}_k \leq \eta_k^\vee < 0$ and $0 < \bar{S} < S(\bar{\eta}_k)$. Then there exists an $\tilde{\eta}_k \in (\bar{y}_k, \bar{\eta}_k)$ that solves

$S(\tilde{\eta}_k) = \bar{S}$. If $\eta_k^\vee \geq \tilde{y}_k - \tilde{\eta}_k$, then $\tilde{\eta}_k \in (\tilde{y}_k, \tilde{\eta}_k)$ is the unique solution in $[\eta_k^\wedge, \eta_k^\vee]$. Otherwise, there might be another solution between $\tilde{y}_k - \tilde{\eta}_k$ and η_k^\vee .

The conclusions can be summarized based on the above cases. \square

Lemma S1. For the Bernoulli model with the logit link, $S_0(\eta)$ strictly increases before η_l and strictly decreases after η_l , where $\eta_l \approx -1.2784645$ is the unique root of the transcendent equation $1 + \eta + e^\eta = 0$; $S_1(\eta)$ strictly increases before $\eta_r = -\eta_l \approx 1.2784645$ and strictly decreases after η_r .

Proof of Lemma S1: For $Y_i \sim \text{Bernoulli}(\mu_i)$ with the logit link $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, which is also known as the logistic regression model, $v(\eta) \equiv \text{constant}$. In this case,

$$\begin{aligned} G(\eta) &= \frac{e^\eta}{1 + e^\eta}, \\ S(\eta) &= [\tilde{y}_k - G(\eta)]\eta, \\ S'(\eta) &= \tilde{y}_k - T(\eta), \\ T(\eta) &= \frac{e^\eta(1 + \eta + e^\eta)}{(1 + e^\eta)^2}, \\ T'(\eta) &= \frac{e^\eta}{(1 + e^\eta)^3}[\eta + 2 - e^\eta(\eta - 2)]. \end{aligned}$$

The following can be verified:

$$T(\eta) \begin{cases} < 0, & \text{if } \eta < \eta_0; \\ = 0, & \text{if } \eta = 0; \\ > 0, & \text{if } \eta > \eta_0, \end{cases}$$

with $\eta_0 \approx -1.278$, $T'(-\eta) = T'(\eta)$, and

$$T'(\eta) \begin{cases} < 0, & \text{if } \eta < -\eta_1 \text{ or } \eta > \eta_1; \\ = 0, & \text{if } \eta = \pm\eta_1; \\ > 0, & \text{if } \eta \in (-\eta_1, \eta_1), \end{cases}$$

where $\eta_1 \approx 2.399$. In other words, neither $S(\eta)$ nor $T(\eta)$ is monotone.

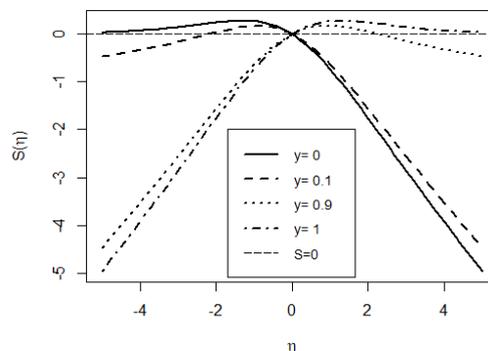


Figure S4. $S(\eta)$ with different \tilde{y}_k for the Bernoulli model with the logit link.

Figure S4 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k . Note that for the Bernoulli models, $y_i \in \{0, 1\}$ for all i . Suppose either $\eta_i > 0$ for all $i \in I_k$ or $\eta_i < 0$ for all $i \in I_k$. Then, $\tilde{y}_k \in [0, 1]$.

We further assume either $y_i < G(\eta_i)$ for all $i \in I_k$ or $y_i > G(\eta_i)$ for all $i \in I_k$. Since for the Bernoulli models, $y_i \in \{0, 1\}$ and $G(\eta_i) \in (0, 1)$ for all i , then $y_i < G(\eta_i)$ always implies $y_i = 0$ and $y_i > G(\eta_i)$ always implies $y_i = 1$.

Case (i): $\eta_i > 0$ and $y_i < G(\eta_i)$ for all $i \in I_k$, that is, $y_i = 0$ for all $i \in I_k$. We claim that the solution to (5) is unique in this case. Moreover, $0 < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < \infty$, and

$$\tilde{y}_k = \frac{n_k^{-1} \sum_{i \in I_k} \eta_i y_i}{\bar{\eta}_k} = 0.$$

In this case, $S'(\eta) = -T(\eta) < 0$ for $\eta > -\eta_0 \approx -1.278$, that is, $S(\eta)$ is strictly monotone on $[\eta_k^\wedge, \eta_k^\vee]$. According to Theorem A1, the solution solving (3) is unique in $[\eta_k^\wedge, \eta_k^\vee]$.

Case (ii): $\eta_i > 0$ and $y_i > G(\eta_i)$ for all $i \in I_k$, that is, $y_i = 1$ for all $i \in I_k$. In this case, (5) may not be unique in $[\eta_k^\wedge, \eta_k^\vee]$. Moreover, we have $0 < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < \infty$,

$$\tilde{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} = 1 > \frac{\sum_{i \in I_k} \eta_i G(\eta_i)}{\sum_{i \in I_k} \eta_i} = \tilde{G}_k,$$

and $\bar{S} = \bar{\eta}_k(\tilde{y}_k - \tilde{G}_k) > 0$. If $[\eta_k^\wedge, \eta_k^\vee]$ is wide enough, $S(\eta) = \bar{S}$ may have a solution in $[\eta_k^\wedge, \eta_0]$ and another solution in $[\eta_0, \eta_k^\vee]$, where $\eta_0 \approx 1.278$.

Case (iii): $\eta_i < 0$ and $y_i > G(\eta_i)$ for all $i \in I_k$, that is, $y_i = 1$ for all $i \in I_k$. We claim that the solution to (5) is unique in this case. Moreover, $-\infty < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < 0$,

$$\tilde{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} = 1 > \frac{\sum_{i \in I_k} \eta_i G(\eta_i)}{\sum_{i \in I_k} \eta_i} = \tilde{G}_k,$$

and $\bar{S} = \bar{\eta}_k(\tilde{y}_k - \tilde{G}_k) < 0$. In this case, $S'(\eta) = 1 - T(\eta) > 0$ for $\eta < \eta_0 \approx 1.278$, that is, $S(\eta)$ is strictly monotone on $[\eta_k^\wedge, \eta_k^\vee]$. According to Theorem A1, the solution solving (3) is unique in $[\eta_k^\wedge, \eta_k^\vee]$.

Case (iv): $\eta_i < 0$ and $y_i < G(\eta_i)$ for all $i \in I_k$, that is, $y_i = 0$ for all $i \in I_k$. In this case, (5) may not be unique in $[\eta_k^\wedge, \eta_k^\vee]$. Moreover, we have $-\infty < \eta_k^\wedge \leq \bar{\eta}_k \leq \eta_k^\vee < 0$,

$$\tilde{y}_k = \frac{\sum_{i \in I_k} \eta_i y_i}{\sum_{i \in I_k} \eta_i} = 0 < \frac{\sum_{i \in I_k} \eta_i G(\eta_i)}{\sum_{i \in I_k} \eta_i} = \tilde{G}_k,$$

and $\bar{S} = \bar{\eta}_k(\tilde{y}_k - \tilde{G}_k) > 0$. If $[\eta_k^\wedge, \eta_k^\vee]$ is wide enough, $S(\eta) = \bar{S}$ may have a solution in $[\eta_k^\wedge, -\eta_0]$ and another solution in $[-\eta_0, \eta_k^\vee]$, where $\eta_0 \approx 1.278$.

For logit link,

$$S_0(\eta) = -\frac{\eta e^\eta}{1 + e^\eta}, \quad S_1(\eta) = \frac{\eta}{1 + e^\eta}.$$

Since $S_1(\eta) = S_0(-\eta)$, we only need to show the conclusion on S_0 . Moreover,

$$S'_0(\eta) = -(1 + \eta + e^\eta) \frac{e^\eta}{(1 + e^\eta)^2}.$$

Let $V(\eta) = -(1 + \eta + e^\eta)$. Then, $\text{sgn}(S'_0(\eta)) = \text{sgn}(V(\eta))$. Note that $V'(\eta) = -1 - e^\eta < 0$ for all η . Since $V(-2) = 1 - e^{-2} > 0$ and $V(0) = -2 < 0$, $V(\eta) = 0$ has one and only one solution $\eta_0 \in (-2, 0)$, which is approximately -1.278464543 . Before η_0 , $V(\eta) > 0$ and, thus, $S'_0(\eta) > 0$; after η_0 , $V(\eta) < 0$ and, thus, $S'_0(\eta) < 0$. \square

Lemma S2. For the Bernoulli model with a probit link, $S_0(\eta)$ strictly increases before $-\eta_0$ and strictly decreases after $-\eta_0$; $S_1(\eta)$ strictly increases before η_0 and strictly decreases after η_0 , where $\eta_0 \approx 0.839924$ is the unique positive root of the transcendent equation $(1 - \eta^2)\Phi(\eta) = \eta\phi(\eta)$.

Proof of Lemma S2: For $Y_i \sim \text{Bernoulli}(\mu_i)$ with a probit link $g(\mu_i) = \Phi^{-1}(\mu_i)$, we have the following:

$$\begin{aligned} G(\eta) &= \Phi(\eta), \\ \nu(\eta) &= \frac{\phi(\eta)}{\Phi(\eta)[1 - \Phi(\eta)]} = \frac{\phi(\eta)}{\Phi(\eta)\Phi(-\eta)}, \\ S(\eta) &= \nu(\eta)[\tilde{y}_k - G(\eta)]\eta = \begin{cases} -\frac{\eta\phi(\eta)}{1 - \Phi(\eta)}, & \text{if } \tilde{y}_k = 0; \\ \frac{\eta\phi(\eta)}{\Phi(\eta)}, & \text{if } \tilde{y}_k = 1. \end{cases} \end{aligned}$$

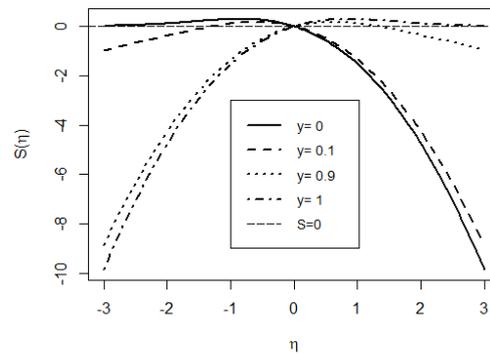


Figure S5. $S(\eta)$ with different \tilde{y}_k for the Bernoulli model with the probit link.

Figure S5 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k . Note that For the Bernoulli models, $y_i \in \{0, 1\}$ for all i . Suppose (i) either $\eta_i > 0$ for all $i \in I_k$ or $\eta_i < 0$ for all $i \in I_k$; and (ii) either $\eta_i > G(\eta_i)$ for all $i \in I_k$ or $\eta_i < G(\eta_i)$ for all $i \in I_k$. Then, \tilde{y}_k is either 0 or 1.

Depending on $\tilde{y}_k = 0$ or 1, we denote $S(\eta)$ as $S_0(\eta)$ or $S_1(\eta)$, respectively. Then, we have the following:

$$S_0(\eta) = -\frac{\eta\phi(\eta)}{1 - \Phi(\eta)} = S_1(-\eta).$$

In the literature, $r(\eta) = [1 - \Phi(\eta)]/\phi(\eta)$ is known as Mills' ratio of the standard normal distribution (see [6–9]), whose reciprocal, $\phi(\eta)/[1 - \Phi(\eta)]$, is also known as the hazard rate (see, for example, [10]).

Since $S_0(\eta) = S_1(-\eta)$, then $S'_1(\eta) = -S'_0(-\eta)$. We only need to show the conclusion on S_0 . Moreover,

$$S'_0(\eta) = \frac{\phi(\eta)}{1 - \Phi(\eta)} \left[\eta^2 - 1 - \frac{\eta\phi(\eta)}{1 - \Phi(\eta)} \right] = \frac{1}{r(\eta)} \left[\eta \left(\eta - \frac{1}{r(\eta)} \right) - 1 \right],$$

where $r(\eta) = [1 - \Phi(\eta)]/\phi(\eta)$ is the so-called Mills' ratio. Since $r(\eta) > 0$ for all η , then the sign of $S'_0(\eta)$, denoted as $\text{sgn}(S'_0(\eta))$, is the same as the sign of $U(\eta) = \eta(\eta - r(\eta)^{-1}) - 1$, where $\text{sgn}(x) = -1$ if $x < 0$; 0 if $x = 0$; and 1 if $x > 0$.

First of all, $U(\eta) = \eta^2 - 1 - \eta/r(\eta) > 0$ if $\eta \leq -1$. It can be verified that $U(\eta) = 0$ has a unique solution in $(-1, 0)$, which is approximately -0.839924 , denoted as $-\eta_0$. On the other hand, according to [6], $r(\eta) < 1/\eta$ if $\eta > 0$. Then, $U(\eta) < -1 < 0$ if $\eta > 0$. \square

Let $\eta_l = -\eta_0 \approx -0.839924$ and $\eta_r = \eta_0 \approx 0.839924$. A direct strategy for splitting sub-blocks inspired by Lemma S2 is as follows:

- 1 $^\circ$ Divide an original block into four sub-blocks: (i) $\eta_i > 0$ and $\eta_i < G(\eta_i)$; (ii) $\eta_i > 0$ and $y_i > G(\eta_i)$; (iii) $\eta_i < 0$ and $y_i > G(\eta_i)$; and (iv) $\eta_i < 0$ and $y_i > G(\eta_i)$.
- 2 $^\circ$ For cases (i) and (iii), $\tilde{\eta}_k$ solving (5) is unique.

- 3° For case (ii), if $\eta_r \notin (\eta_k^\wedge, \eta_k^\vee)$, $\tilde{\eta}_k$ is unique; otherwise, there are up to two solutions solving (5). In the case when $\eta_r \in (\eta_k^\wedge, \eta_k^\vee)$, we further divide the block according to $\eta_i < \eta_r$ or $\eta_i \geq \eta_r$. In each of the two sub-blocks, $\tilde{\eta}_k$ is unique.
- 4° For case (iv), if $\eta_l \notin (\eta_k^\wedge, \eta_k^\vee)$, $\tilde{\eta}_k$ is unique; otherwise, there are up to two solutions solving (5). In the case when $\eta_l \in (\eta_k^\wedge, \eta_k^\vee)$, we further divide the block according to $\eta_i < \eta_l$ or $\eta_i \geq \eta_l$. In each of the two sub-blocks, $\tilde{\eta}_k$ is unique.

Lemma S3. For the Bernoulli model with a complementary log–log link, $S_0(\eta)$ strictly increases before -1 and strictly decreases after -1 ; $S_1(\eta)$ strictly increases before η_r and strictly decreases after η_r , where $\eta_r \approx 0.729114$ is the unique positive root of the transcendent equation $1 + \eta = \exp(e^\eta)(1 + \eta - \eta e^\eta)$.

Proof of Lemma S3: For $Y_i \sim \text{Bernoulli}(\mu_i)$ with complementary-log-log link $g(\mu_i) = \log(-\log(1 - \mu_i))$, we have the following:

$$\begin{aligned} v(\eta) &= \frac{\exp(\eta)}{1 - \exp[-\exp(\eta)]}, \\ G(\eta) &= 1 - \exp\{-\exp(\eta)\}, \\ S(\eta) &= v(\eta)[\tilde{y}_k - G(\eta)]\eta = \begin{cases} -\eta \exp(\eta), & \text{if } \tilde{y}_k = 0; \\ \frac{\eta \exp(\eta)}{\exp\{\exp(\eta)\} - 1}, & \text{if } \tilde{y}_k = 1. \end{cases} \end{aligned}$$

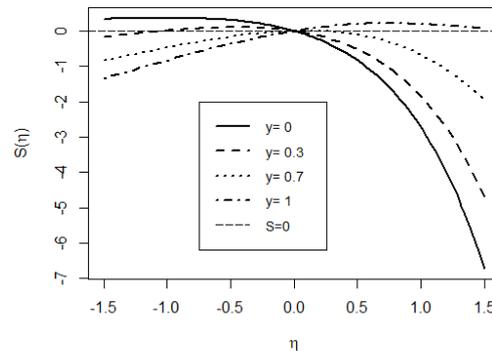


Figure S6. $S(\eta)$ with different \tilde{y}_k for the Bernoulli model with a cloglog link.

Figure S6 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k . Note that for the Bernoulli models with either $y_i > G(\eta_i)$ or $y_i < G(\eta_i)$, \tilde{y}_k is either 0 or 1. Depending on $\tilde{y}_k = 0$ or 1, we denote $S(\eta)$ as $S_0(\eta)$ or $S_1(\eta)$, respectively. Then, we have the following:

$$S_0(\eta) = -\eta \exp(\eta), \quad S_1(\eta) = \frac{\eta \exp(\eta)}{\exp\{\exp(\eta)\} - 1}.$$

Since $S'_0(\eta) = -e^\eta(1 + \eta)$, then $S'_0(\eta) > 0$ if $\eta < -1$; $= 0$ if $\eta = -1$; and < 0 if $\eta > -1$. That is, $S(\eta)$ strictly increases before $\eta_l = -1$ and strictly decreases after $\eta_l = -1$.

As for $S_1(\eta)$,

$$S'_1(\eta) = \frac{e^\eta}{[\exp(e^\eta) - 1]^2} [(1 + \eta)(\exp(e^\eta) - 1) - \eta e^\eta \exp(e^\eta)].$$

Let $V(\eta) = (1 + \eta)(\exp(e^\eta) - 1) - \eta e^\eta \exp(e^\eta) = \exp(e^\eta)(1 + \eta - \eta e^\eta) - (1 + \eta)$. Then, $\text{sgn}(S'_1(\eta)) = \text{sgn}(V(\eta))$. Note that

$$V'(\eta) = \exp(e^\eta) - 1 - \eta \exp(e^\eta + 2\eta).$$

Apparently, $V'(\eta) > 0$ for all $\eta < 0$. By applying L'Hospital's rule, we can verify that $\lim_{\eta \rightarrow -\infty} V(\eta) = 0$. Since $V(0) = e - 1 > 0$, we conclude that $V(\eta) > 0$ for all $\eta \leq 0$. Similarly, if $\eta \geq 1$, then $V'(\eta) \leq \exp(e^\eta) - 1 - \exp(e^\eta)e^{2\eta} < \exp(e^\eta) - \exp(e^\eta) = 0$. Along with $V(1) \approx -12.885 < 0$, we conclude that $V(\eta) < 0$ for all $\eta \geq 1$. It can be verified that there is one and only one solution solving $V'(\eta) = 0$ in $(0, 1)$. \square

Lemma S4. For the Bernoulli model with the log-log link, $S_0(\eta)$ strictly increases before η_l and strictly decreases after η_l , where $\eta_l \approx -0.729114$ is the unique positive root of the transcendent equation

$$1 - \eta = \exp(e^{-\eta})(1 - \eta + \eta e^{-\eta});$$

$S_1(\eta)$ strictly increases before $\eta_r = 1$ and strictly decreases after $\eta_r = 1$.

Proof of Lemma S4: Consider $Y_i \sim \text{Bernoulli}(\mu_i)$ with the log-log link $g(\mu_i) = -\log(-\log \mu_i)$. Note that in this paper, we follow [5] and choose a strictly increasing link function. Then, we have the following:

$$\begin{aligned} G(\eta) &= \exp\{-e^{-\eta}\}, \\ v(\eta) &= \frac{e^{-\eta}}{1 - \exp(-e^{-\eta})}, \\ S(\eta) &= \begin{cases} \frac{\eta e^{-\eta}}{1 - \exp(e^{-\eta})}, & \text{if } \tilde{y}_k = 0; \\ \eta e^{-\eta}, & \text{if } \tilde{y}_k = 1. \end{cases} \end{aligned}$$

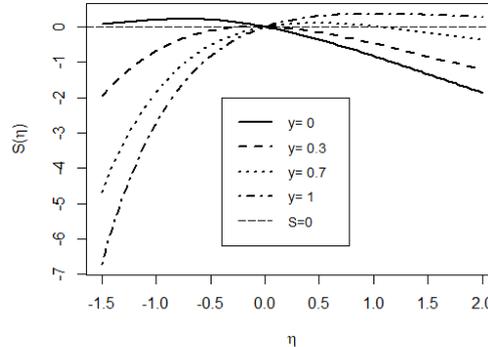


Figure S7. $S(\eta)$ with different \tilde{y}_k for the Bernoulli model with the loglog link.

Figure S7 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k . Note that For the Bernoulli models with either $y_i > G(\eta_i)$ or $y_i < G(\eta_i)$, \tilde{y}_k is either 0 or 1. Depending on $\tilde{y}_k = 0$ or 1, we denote $S(\eta)$ as $S_0(\eta)$ or $S_1(\eta)$, respectively. Then, we have the following:

$$S_0(\eta) = \frac{\eta e^{-\eta}}{1 - \exp(e^{-\eta})}, \quad S_1(\eta) = \eta e^{-\eta}.$$

Let $S_0^{(c)}(\eta) = -\eta \exp(\eta)$, $S_1^{(c)}(\eta) = \frac{\eta \exp(\eta)}{\exp\{\exp(\eta)\}-1}$, that is, the corresponding functions with the cloglog link. Since $S_0(\eta) = S_0^{(c)}(-\eta)$ and $S_1(\eta) = S_1^{(c)}(-\eta)$, the conclusion of Lemma S4 can be obtained as a corollary of Lemma S3. \square

Lemma S5. For the Bernoulli model with the cauchit link, $S_0(\eta)$ strictly increases before η_l and strictly decreases after η_l , where $\eta_l \approx -0.801916$ is the unique root of the transcendent equation $\eta = (\eta^2 - 1)[\pi/2 - \arctan(\eta)]$; $S_1(\eta)$ strictly increases before η_r and strictly decreases after η_r , where $\eta_r = -\eta_l \approx 0.801916$.

Proof of Lemma S5: For $Y_i \sim \text{Bernoulli}(\mu_i)$ with the cauchit link $g(\mu_i) = \tan\left(\pi(\mu_i - \frac{1}{2})\right)$, we have the following:

$$\begin{aligned} G(\eta) &= \arctan(\eta)/\pi + 1/2, \\ v(\eta) &= \frac{\pi}{(1 + \eta^2)(\pi^2/4 - \arctan^2(\eta))}, \\ S(\eta) &= \begin{cases} -\frac{\eta}{(1 + \eta^2)[\pi/2 - \arctan(\eta)]}, & \text{if } \tilde{y}_k = 0; \\ \frac{\eta}{(1 + \eta^2)[\pi/2 + \arctan(\eta)]}, & \text{if } \tilde{y}_k = 1. \end{cases} \end{aligned}$$

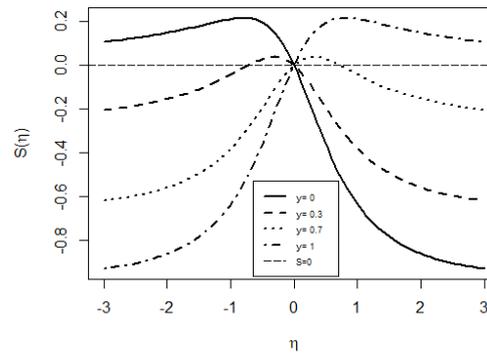


Figure S8. $S(\eta)$ with different \tilde{y}_k for the Bernoulli model with the cauchit link.

Figure S8 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k . Since we have either $y_i > G(\eta_i)$ or $y_i < G(\eta_i)$, \tilde{y}_k is either 0 or 1. Depending on $\tilde{y}_k = 0$ or 1, we denote $S(\eta)$ as $S_0(\eta)$ or $S_1(\eta)$, respectively. Then, we have the following:

$$S_0(\eta) = -\frac{\eta}{(1 + \eta^2)[\pi/2 - \arctan(\eta)]}, \quad S_1(\eta) = \frac{\eta}{(1 + \eta^2)[\pi/2 + \arctan(\eta)]}.$$

Since $S_1(\eta) = S_0(-\eta)$ in this case, we only need to justify the conclusion on $S_0(\eta)$. Moreover,

$$S'_0(\eta) = (1 + \eta^2)^{-2} \left[\frac{\pi}{2} - \arctan(\eta) \right]^{-2} \cdot \left\{ -\eta + (\eta^2 - 1) \left[\frac{\pi}{2} - \arctan(\eta) \right] \right\}.$$

Let $V(\eta) = -\eta + (\eta^2 - 1)[\pi/2 - \arctan(\eta)]$. Then, $\text{sgn}(S'_0(\eta)) = \text{sgn}(V(\eta))$. Note that

$$V'(\eta) = 2\eta \left[\frac{\pi}{2} - \arctan(\eta) - \frac{\eta}{1 + \eta^2} \right].$$

Apparently, $V'(\eta) < 0$ if $\eta < 0$. Since $V(-1) = 1 > 0$ and $V(0) = -\pi/2 < 0$, then $V(\eta) = 0$ has a unique solution $\eta_l \in (-\infty, 0]$, within $(-1, 0)$. Therefore, $V(\eta) > 0$ for $\eta < \eta_l$, and $V(\eta) < 0$ for $\eta \in (\eta_l, 0]$.

As for $\eta > 0$, we set $U(\eta) = \pi/2 - \arctan(\eta) - \eta/(1 + \eta^2)$. We claim that $U(\eta) > 0$ for all $\eta > 0$, which implies $V'(\eta) > 0$ for all $\eta > 0$. Moreover, $U(0) = \pi/2 > 0$, $U'(\eta) = -2(1 + \eta^2)^{-2} < 0$ for all $\eta > 0$, and $\lim_{\eta \rightarrow \infty} U(\eta) = 0$ since $\lim_{\eta \rightarrow \infty} \arctan(\eta) = \pi/2$, which imply $U(\eta) > 0$ for all $\eta > 0$.

We further claim that $V(\eta) < 0$ for all $\eta > 0$. Moreover, $V(0) = -\pi/2 < 0$. We have established that $V'(\eta) > 0$ for all $\eta > 0$. It is enough to justify $\lim_{\eta \rightarrow \infty} V(\eta) = 0$. Moreover, by the Taylor series, as η approaches ∞ ,

$$\frac{\pi}{2} - \arctan(\eta) = \frac{1}{\eta} - \frac{1}{3\eta^3} + O(\eta^{-4}).$$

Then, $V(\eta) = -\frac{4}{3\eta} + O(\eta^{-2}) \rightarrow 0$, as η goes to ∞ . \square

Lemma S6. Let $h(y)$ denote the solution solving $e^\eta(1 + \eta) = y$ for $y \geq 0$. Then, $h(y)$ exists uniquely in $[-1, \infty)$. Furthermore, $h(y)$ strictly increases on $y \geq 0$.

Proof of Lemma S6: Let $V(\eta) = e^\eta(1 + \eta)$. Then, $V'(\eta) = e^\eta(2 + \eta)$. For $y \geq 0$, since $V'(\eta) > 0$ for all $\eta > -2$, $V(\eta) < 0$ for all $\eta < -1$, and $V(-1) = 0$, then $h(y)$ exists and is unique in $[-1, \infty)$. As y increases, the solution $h(y)$ solving $V(\eta) = y$ strictly increases as well.

On the other hand, $V(\eta) < 0$ for $\eta < -1$, which implies that there is no solution in $(-\infty, -1)$ solving $V(\eta) = y$ if $y \geq 0$. So, $h(y)$ exists uniquely. \square

Proof of Theorem A4: For $Y_i \sim \text{Poisson}(\mu_i)$ with the log link $g(\mu_i) = \log(\mu_i)$, we have the following:

$$\begin{aligned} G(\eta) &= e^\eta, \\ v(\eta) &\equiv 1, \\ S(\eta) &= (\tilde{y}_k - e^\eta)\eta, \\ S'(\eta) &= \tilde{y}_k - e^\eta(1 + \eta). \end{aligned}$$

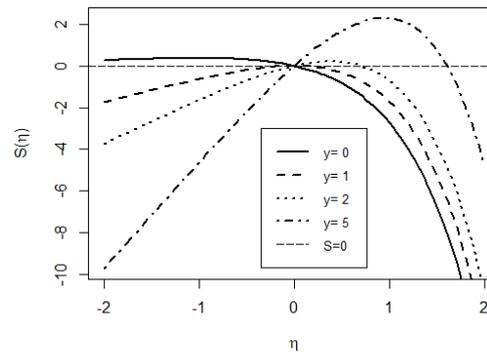


Figure S9. $S(\eta)$ with different \tilde{y}_k for the Poisson model with the log link.

Figure S9 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k .

Let $u(y)$ denote the solution solving $e^\eta(1 + \eta) = y$ for the given $y \geq 0$. That is, $u(\tilde{y}_k)$ solves $S'(\eta) = 0$. According to Lemma S6, $u(y)$ exists and is unique in $[-1, \infty)$. Specifically, $u(0) = -1$ and $u(1) = 0$. In the mathematical literature, $u(y)$ is associated with the Lambert W -function, also called the omega function, which is the inverse of the function $w \rightarrow we^w$ (see, for example, [11]). Moreover, one may use the function `lambertW` in the R package VGAM to calculate $u(y)$. More specifically, $u(y)$ can be obtained by `lambertW(exp(1)*y) - 1` using R.

If either $\eta_i > 0$ for all i or $\eta_i < 0$ for all i , then $\tilde{y}_k \geq 0$ since all $y_i \geq 0$ for the Poisson model. Since $S'(\eta) = \tilde{y}_k - e^\eta(1 + \eta)$, then $S'(\eta) > 0$ for all $\eta < -1$.

Note that $S'(-1) = \tilde{y}_k \geq 0$. According to Lemma S6, $S'(\eta) = 0$ has one and only one solution $u(\tilde{y}_k) \in [-1, \infty)$. Since $S''(\eta) = -e^\eta(2 + \eta) < 0$ for all $\eta > -2$, that is, $S'(\eta)$ strictly decreases on $\eta \in (-2, \infty)$, then $S'(\eta) > 0$ if $\eta \in (-\infty, u(\tilde{y}_k))$ and $S'(\eta) < 0$ if $\eta > u(\tilde{y}_k)$. In other words, $S(\eta)$ strictly increases before $u(\tilde{y}_k)$ and then strictly decreases after $u(\tilde{y}_k)$.

By using the L'Hospital's rule, we have that $\lim_{\eta \rightarrow -\infty} S(\eta) = 0$ if $\tilde{y}_k = 0$, and $-\infty$ if $\tilde{y}_k > 0$; $\lim_{\eta \rightarrow \infty} S(\eta) = -\infty$. Recall that $\bar{S} = \bar{\eta}_k(\tilde{y}_k - \bar{G}_k)$ and $S(\eta) = \bar{S}$ is equivalent to (3).

If $\tilde{y}_k > 0$, $S(\eta) = \bar{S}$ yields up to two solutions. One is before $u(\tilde{y}_k)$ and the other is after $u(\tilde{y}_k)$.

If $\tilde{y}_k = 0$, then $S(\eta) = -\eta e^\eta > 0$ if and only if $\eta < 0$. If all $\eta_i > 0$, then $\tilde{y}_k < \tilde{G}_k$ and, thus, $\bar{S} < 0$. The only possible solution exists in $[0, \infty)$, which can be solved in the form of either $1 + u(-\bar{S}/e)$ or $P(-\bar{S})$. If all $\eta_i < 0$, then $\tilde{y}_k < \tilde{G}_k$ and $\bar{S} > 0$. Since $u(0) = -1$, $S(\eta) = \bar{S}$ yields one solution in $(-\infty, -1]$ and one in $[-1, 0)$, both of which can be determined numerically. \square

Proof of Theorem A5: We consider $Y_i \sim \text{Gamma}(s, \mu_i/s)$ with a reciprocal or inverse link $g(\mu) = 1/\mu$. It has a probability density function, as follows:

$$f(y) = \frac{1}{\Gamma(s)} \left(\frac{s}{\mu_i} \right)^s y^{s-1} \exp\left\{ -\frac{s}{\mu_i} y \right\}, \quad y > 0,$$

where $s > 0$ is assumed to be known, called the shape parameter, and $\mu_i/s > 0$ is called the scale parameter. The gamma model here is described in Section 8.3.3 of [5]. In this case,

$$\begin{aligned} G(\eta) &= \eta^{-1}, \\ v(\eta) &\equiv -s < 0, \\ S(\eta) &= s(1 - \tilde{y}_k \eta). \end{aligned}$$

For the gamma model, $y_i > 0$ and $\eta_i > 0$ for all i . Then, $\tilde{y}_k > 0$ and, thus, $S(\eta)$ is strictly decreasing on $\eta > 0$. According to Theorem A1, the solution $\tilde{\eta}_k$ solving (3) is always unique.

The objective is to find a solution for $S(\eta) = n_k^{-1} \sum_{i \in I_k} S(\eta_i)$, which does not depend on s . We simply let $s = 1$ and re-define $S(\eta) = 1 - \tilde{y}_k \eta$. In this case, we have the following:

$$\bar{S} = \frac{1}{n_k} \sum_{i \in I_k} S(\eta_i) = \frac{1}{n_k} \sum_{i \in I_k} v(\eta_i) \eta_i (\tilde{y}_k - \tilde{G}_k) = -\bar{\eta}_k (\tilde{y}_k - \tilde{G}_k),$$

where $\bar{\eta}_k = n_k^{-1} \sum_{i \in I_k} \eta_i$, and $\tilde{G}_k = \left[\sum_{i \in I_k} v(\eta_i) \eta_i G(\eta_i) \right] / \left[\sum_{i \in I_k} v(\eta_i) \eta_i \right] = \bar{\eta}_k^{-1}$ in this case. Note that (3) is equivalent to $S(\eta) = \bar{S}$, whose solution is as follows:

$$\tilde{\eta}_k = \bar{\eta}_k = \frac{1}{n_k} \sum_{i \in I_k} \eta_i.$$

Furthermore, if $y_i < G(\eta_i)$ for all $i \in I_k$, then $\tilde{y}_k < \tilde{G}_k$, $\bar{S} > 0$, and, thus, $\tilde{\eta}_k \in (0, \tilde{y}_k^{-1})$; if $y_i > G(\eta_i)$ for all $i \in I_k$, then $\tilde{y}_k > \tilde{G}_k$, $\bar{S} < 0$ and, thus, $\tilde{\eta}_k \in (\tilde{y}_k^{-1}, \infty)$. \square

Proof of Theorem A6: for the inverse Gaussian model, $Y_i \sim \text{IG}(\mu_i, \phi)$ with the inverse-square link $g(\mu_i) = \mu_i^{-2}$. It has a probability density function, as follows:

$$f(y) = (2\pi\phi)^{-\frac{1}{2}} y^{-\frac{3}{2}} \exp\left\{ -\frac{(y - \mu_i)^2}{2\phi\mu_i^2 y} \right\}, \quad y > 0,$$

where $\phi > 0$ is assumed to be known, called the dispersion parameter, and $\mu_i = E(Y_i) > 0$ is the mean parameter (see, for example, Chapter 11 of [12]). In this case, we have the following:

$$\begin{aligned} G(\eta) &= \eta^{-\frac{1}{2}}, \\ v(\eta) &\equiv -\frac{1}{2\phi} < 0, \\ S(\eta) &= \frac{1}{2\phi} \left(\eta^{-\frac{1}{2}} - \tilde{y}_k \right) \eta. \end{aligned}$$

Since the objective is to find a solution for $S(\eta) = n_k^{-1} \sum_{i \in I_k} S(\eta_i)$, which does not depend on $\phi > 0$, we set $\phi = 1$ moving forward. That is, $S(\eta) = \frac{1}{2} \left(\eta^{-\frac{1}{2}} - \tilde{y}_k \right) \eta$.

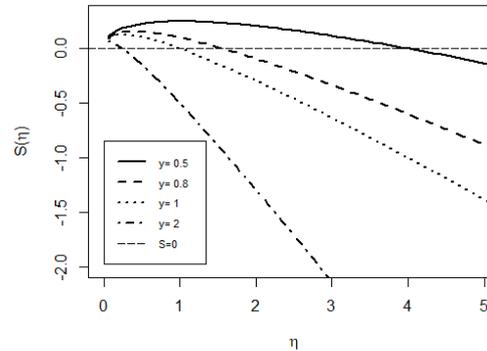


Figure S10. $S(\eta)$ with different \tilde{y}_k for the inverse Gaussian model with the inverse-square link.

Figure S10 shows graphs of $S(\eta)$ with different possible values of \tilde{y}_k for the inverse Gaussian model.

Since $y_i > 0$ and $\eta_i > 0$ for all i , then $\tilde{y}_k = \left[\sum_{i \in I_k} \eta_i y_i \right] / \left[\sum_{i \in I_k} \eta_i \right]^{-1} > 0$. Recall that we set $\phi = 1$. The following can be verified:

$$S'(\eta) = \frac{1}{2} \left(\frac{1}{2} \eta^{-\frac{1}{2}} - \tilde{y}_k \right) \begin{cases} > 0, & \text{if } 0 < \eta < (4\tilde{y}_k^2)^{-1}; \\ = 0, & \text{if } \eta = (4\tilde{y}_k^2)^{-1}; \\ < 0, & \text{if } \eta > (4\tilde{y}_k^2)^{-1}. \end{cases}$$

That is, $S(\eta)$ strictly increases before $(4\tilde{y}_k^2)^{-1}$ and strictly decreases after $(4\tilde{y}_k^2)^{-1}$.

On the other hand, $S(\eta) > 0$ if and only if $\eta \in (0, \tilde{y}_k^{-2})$. The following can be verified:

$$\bar{S} = \frac{1}{n_k} \sum_{i \in I_k} S(\eta) = -\frac{1}{2} \bar{\eta}_k (\tilde{y}_k - \tilde{G}_k)$$

with $\bar{\eta}_k = n_k^{-1} \sum_{i \in I_k} \eta_i$ and $\tilde{G}_k = \left[\sum_{i \in I_k} \eta_i G(\eta_i) \right] / \left[\sum_{i \in I_k} \eta_i \right]$.

If $y_i < G(\eta_i)$ for all $i \in I_k$, then $\tilde{y}_k < \tilde{G}_k$ and, thus, $\bar{S} > 0$. In this case, $S(\eta) = \bar{S}$, which is equivalent to (3), has up to two solutions. If $y_i > G(\eta_i)$ for all $i \in I_k$, then $\tilde{y}_k > \tilde{G}_k$ and, thus, $\bar{S} < 0$. In this case, $S(\eta) = \bar{S}$ has only one solution, as $S(\eta)$ is strictly decreasing on $[\tilde{y}_k^{-2}, \infty)$. The corresponding formulae of solutions can be obtained by solving $(\eta^{-1/2} - \tilde{y}_k) \eta = 2\bar{S}$. \square

S7. More on Airline Data

Table S8 provides a list of brief definitions of variables considered in this study.

Table S8. Description of variables in the oracle model.

Variable Name	Definition
ArrDelayLabel	binary response variable: "1" if flight delay is more than or equal to 15 minutes, "0" otherwise.
QUARTER	"1": Jan. 1 - Mar. 31; "2": Apr. 1 - Jun. 30; "3": Jul. 1 - Sep. 30; "4" Oct. 1 - Dec. 31.
DayOfWeek	"1": Monday; "2" Tuesday; "3": Wednesday; "4": Thursday; "5": Friday; "6": Saturday; "7", Sunday.

Table S8. Description of variables in the oracle model.

Variable Name	Definition
DepTimeBlk	"1": 12:00 AM - 05:59 AM; "2": 06:00 AM - 11:59 AM; "3": 12:00 PM - 05:59 PM; "4": 06:00 PM - 11:59 PM.
CRSTimeElapsed	CRS elapsed time of flights, in minutes.
DISTANCE	the distance that flights travel, in miles.
DepDelay	departure delay of flights, in minutes.

S8. More Figures

In this section, we provide more figures and output to support the previous discussions.

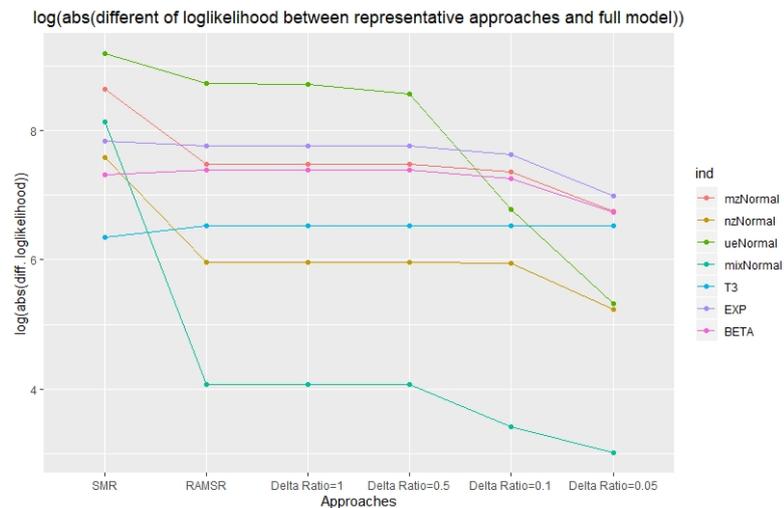


Figure S11. Differences between log-likelihood estimated from representative approaches and the log-likelihood directly from the full amount of data

Figure S11 shows the simulation results relevant to Section 4.3. From this figure, it is evident that as the threshold decreases, the performance of RAMSR with the delta ratio split generally improves in terms of approximating the maximum log-likelihood based on the full data. However, a significantly lower threshold also implies a longer running time. As a compromise, we recommend a delta ratio threshold of $\delta_0 = 0.1$ in this case.

Figure S12 (a) and (b) show the convergence performance of algorithms in various cases that are relevant to Section 4.3.

Figure S13 shows the comparison among the learning rate scheduling strategies with different numbers of variables generated from the mzNormal distribution, which is relevant to Section 2.5. According to Figure S13, the original RAMSR (i.e., RAMSR with a learning rate of 1) performs well with a small number of variables, such as 4 or 7, but not as well with 20 or 100. RAMSR with a constant learning rate of 0.1 performs well with up to 20 variables but struggles with 100. Compared to RAMSR with a constant learning rate, RAMSR with an exponential learning rate performs better over a larger number of iterations, such as 20. However, without truncation at the 10th iteration, the improvement with an exponential learning rate is minimal after this point. Conversely, with truncation at the 10th iteration (that is, at the learning rate $e^{-0.3 \times 10} \approx 0.05$), the RAMSR with an exponential learning rate still improves significantly, especially with 100 variables.

According to Figure S13, we recommend (1) the original RAMSR for GLMs with a small number of variables, such as 4 or 7; (2) the RAMSR with an exponential learning rate $e^{-0.3i}$ and truncation at $i = 10$ for GLMs with a moderate (such as 20) or large (such as 100) number of variables.

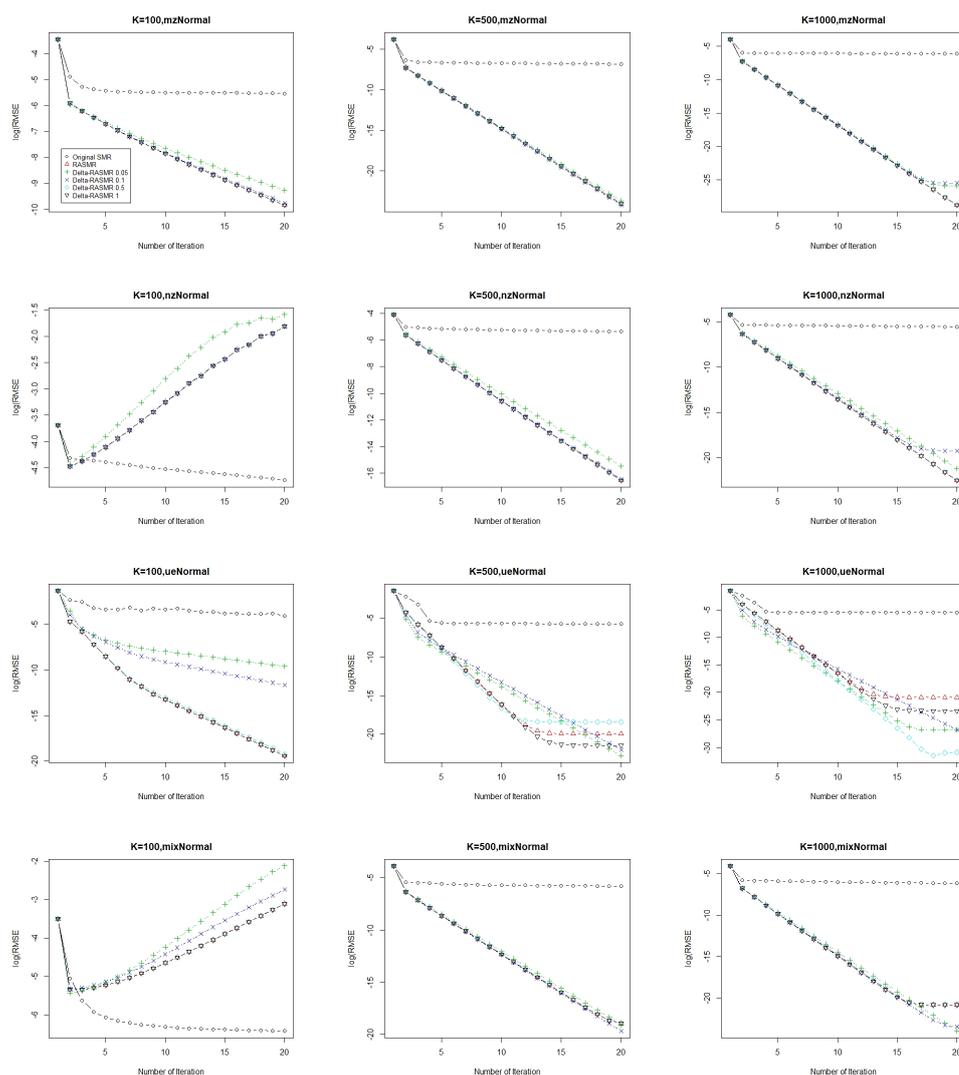


Figure S12. (a) Log(RMSE) of RASMR with various numbers of clusters and the delta ratio threshold

Figure S14 provides the fitted GLM with the logit link for the airline data.

Figure S15 shows the detailed summary from the R output of the “glm2” function for the selected model.

References

1. Li, K.; Yang, J. Score-matching representative approach for big data analysis with generalized linear models. *Electronic Journal of Statistics* **2022**, *16*, 592–635.
2. Lohr, S. *Sampling: Design and Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019.
3. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
4. Chen, X.; Xie, M.g. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **2014**, *24*, 1655–1684.
5. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2 ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 1989.
6. Gordon, R. Values of Mills’ ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Annals of Mathematical Statistics* **1941**, *12*, 364–366.
7. Birnbaum, Z. An inequality for Mill’s ratio. *Annals of Mathematical Statistics* **1942**, *13*, 245–246.
8. Mitrinovic, D.; Vasic, P. *Analytic Inequalities*; Springer: Berlin/Heidelberg, Germany, 1970.

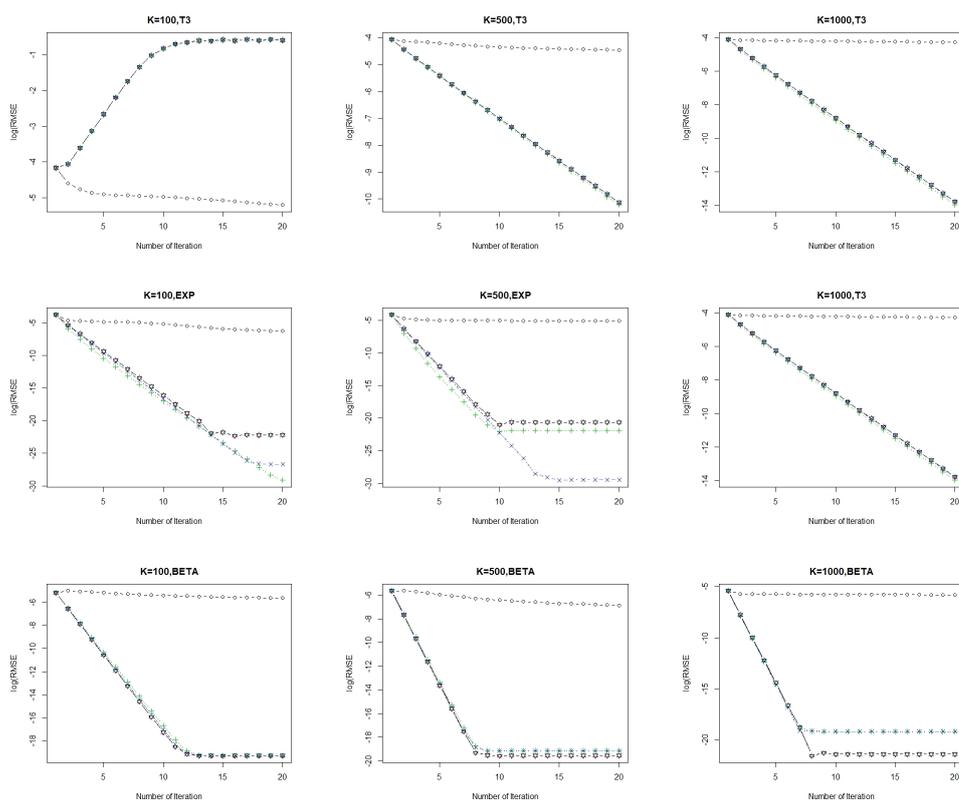


Figure S12. (b) Log(RMSE) of RASMR with various numbers of clusters and the delta ratio threshold

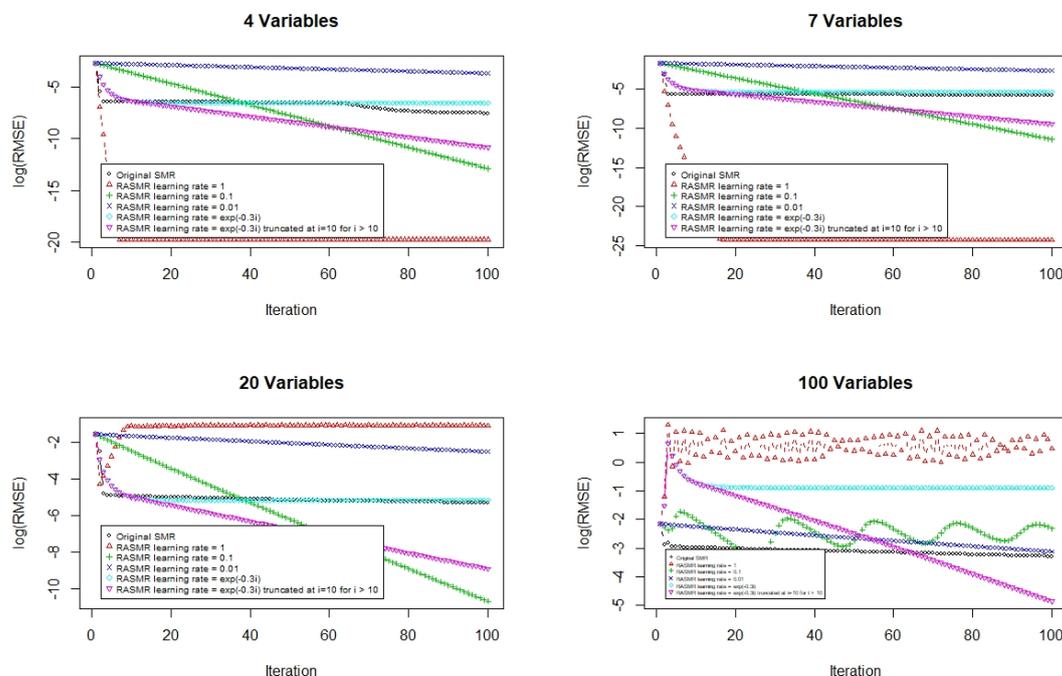


Figure S13. Learning curve comparison of learning rate scheduling strategy with different number of variables generated from mzNormal

9. Baricz, A. Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications* 2008, 340, 1362–1370.

```

*-----*
* GLM fitted using RASMR for Airline Data with Logit Link *
*-----*|

Call:
glm2(formula = formula.smr, family = binomial(link = "logit"),
      data = data.smr.total, weights = data.smr.total[, "n"])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-38.908  -5.754  -0.231   3.613  49.656

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Intercept    -3.124e+00  3.107e-03 -1005.37 <2e-16 ***
QUARTER2     -6.300e-02  7.782e-04  -80.96 <2e-16 ***
QUARTER3     -9.987e-02  7.865e-04 -126.98 <2e-16 ***
QUARTER4     -4.526e-02  7.766e-04  -58.28 <2e-16 ***
DAY_OF_WEEK2  4.421e-02  1.030e-03   42.93 <2e-16 ***
DAY_OF_WEEK3  8.922e-02  1.017e-03   87.72 <2e-16 ***
DAY_OF_WEEK4  1.262e-01  1.001e-03  126.14 <2e-16 ***
DAY_OF_WEEK5  9.030e-02  1.001e-03   90.21 <2e-16 ***
DAY_OF_WEEK6 -2.034e-01  1.098e-03 -185.23 <2e-16 ***
DAY_OF_WEEK7 -8.174e-02  1.044e-03  -78.33 <2e-16 ***
DEP_TIME_BLK2 1.057e-01  2.555e-03   41.36 <2e-16 ***
DEP_TIME_BLK3 1.089e-01  2.555e-03   42.62 <2e-16 ***
DEP_TIME_BLK4 6.190e-02  2.583e-03   23.97 <2e-16 ***
DEP_DELAY     1.469e-01  2.734e-05 5372.79 <2e-16 ***
CRS_ELAPSED_TIME 8.123e-03  4.816e-05  168.65 <2e-16 ***
DISTANCE     -8.491e-04  5.962e-06 -142.40 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 248235882 on 1102741 degrees of freedom
Residual deviance: 90283990 on 1102725 degrees of freedom
AIC: 90284022

Number of Fisher Scoring iterations: 8

```

Figure S14. R output for the GLM with all variables and the logit link.

10. Marshall, A.; Olkin, I. *Life Distributions*; Springer: Berlin/Heidelberg, Germany, 2007.
11. Corless, R.; Gonnet, G.; Hare, D.; Jeffrey, D.; Knuth, D. On the Lambert W function. *Advances in Computational Mathematics* **1996**, *5*, 329–359.
12. Dunn, P.; Smyth, G. *Generalized Linear Models with Examples in R*; Springer: Berlin/Heidelberg, Germany, 2018.

```
*-----*
*   The final GLM for Airline Data Selected by MR + RASMR   *
*-----*

Call:
glm2(formula = formula.smr, family = binomial(link = "logit"),
      data = data.smr.total, weights = data.smr.total[, "n"])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-38.908  -7.178   0.151   6.158  50.214

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
Intercept -2.660e+00  3.357e-04  -7922  <2e-16 ***
DEP_DELAY  1.466e-01  2.662e-05   5508  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 248046740 on 861855 degrees of freedom
Residual deviance: 90044924 on 861853 degrees of freedom
AIC: 90044928

Number of Fisher Scoring iterations: 8
```

Figure S15. R output for the GLM selected by MR + RASMR.