Statistica Sinica **33** (2023), 2545-2560
doi:https://doi.org/10.5705/ss.202021.0332

REGRESSION WITH SET-VALUED CATEGORICAL PREDICTORS

Ganghua Wang, Jie Ding and Yuhong Yang

University of Minnesota

Abstract: We address the regression problem with a new form of data that arises from data privacy applications. Instead of point values, the observed explanatory variables are subsets containing each individual's original value. In such cases, we cannot apply classical regression analyses, such as the least squares, because the set-valued predictors carry only partial information about the original values. We propose a computationally efficient subset least squares method for performing a regression on such data. We establish upper bounds of the prediction loss and risk in terms of the subset structure, model structure, and data dimension. The error rates are shown to be optimal in some common situations. Furthermore, we develop a model-selection method to identify the most appropriate model for prediction. Experiment results on both simulated and real-world data sets demonstrate the promising performance of the proposed method.

Key words and phrases: Model selection, regression, set-valued data.

1. Introduction

Data privacy is an emerging societal concern (Enserink and Chin (2015); Cohen and Nissim (2020)). For example, Rocher, Hendrickx and De Montjoye (2019) showed that even after removing common identifiers for each individual, 99.98% of Americans could be correctly re-identified using only 15 demographic attributes such as family size and vehicle type. As a result, privacy-preserving methods that protect individual identification and sensitive data values are receiving increasing attention. A popular choice is for the data owner to no longer release the exact value X of each individual. Instead, a quantity Z relevant to X is used to enhance individual privacy. Several such privacy-preserving methods have been proposed, including differential privacy (Dwork et al. (2006)), which uses a randomized response technique (Warner (1965)) or adds noise to X, k-anonymity (Aggarwal (2005)), which groups X with similar values to a representative value, and secure multi-party computing (Yao (1982); Chaum, Crépeau and Damgard (1988)), which encrypts X using cryptography techniques.

Corresponding author: Yuhong Yang, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: yangx374@umn.edu.

When developing data privacy techniques, a critical use scenario concerns the data collection procedure. A mechanism called subset privacy (Wang and Ding (2021)) was recently proposed to address the challenge of private data collection. Specifically, the data collector, such as a service provider, collects only a set A that contains the original value X held by the subject, such as an individual user. For example, in a study of income with respect to race using the Adult data set (Dua and Graff (2017)), a data collector might perform a survey that collects only a set of races, instead of the exact race from each participant. Here, A can be generated by a survey-based system, such as the independent design introduced in Subsection 2.2.

Subset privacy provides a privacy guarantee against de-identification. Nevertheless, regression and predictions using this new data format are highly nontrivial. We consider a general regression problem involving a real-valued response variable Y and set-valued predictor variables $A_1, \ldots, A_d$. Specifically, we study the regression model $Y = f(X_1, \ldots, X_d) + \epsilon$, where ϵ is a random noise. The goal is to estimate the underlying function f from n observations of $(Y, A_1, \ldots, A_d)$. This is a nontrivial problem, even when f is linear, because the predictors are no longer point-valued data. For example, we cannot apply the standard least squares method.

In this paper, we propose a computationally efficient subset least squares method. The main idea is to minimize the empirical modified mean squared error, given set-valued data. We derive a closed-form solution for this optimization problem above, and establish an upper bound for the prediction risk and show that it is rate-optimal in some circumstances. Examples include additive models, in which the effect of each variable is independent of the others, and saturated models, in which all variables interact with one another. We also discuss some practical strategies to improve the numerical stability and leverage fast matrix operations. Furthermore, we propose a method for selecting a model from different combinations of variables or interaction orders, and prove its asymptotic efficiency under some conditions. Finally, we perform experiments on simulated and real data to verify the proposed method.

2. Problem Formulation

2.1. Model

Notation. For a positive integer p, let $[p]$ and $2^{[p]}$ denote the set $\{1, 2, \ldots, p\}$ and its power set, respectively. For a set A, let $|A|$ and A^c denote its cardinality and complement, respectively. Let $\mathbb{1}$ and I_p denote the indicator function and

the $p \times p$ identity matrix, respectively. The Kronecker product is denoted by $\otimes$. The trace of a matrix M is $tr(M)$. The largest eigenvalue, smallest eigenvalue, and condition number of a positive-definite matrix M are denoted as $\sigma_{\max}(M)$, $\sigma_{\min}(M)$, and $\kappa(M) = \sigma_{\max}(M)\sigma_{\min}^{-1}(M)$, respectively. We sometimes represent a finite set $A \subseteq [p]$ using a vector $\mathbf{1}_A \in \{0,1\}^p$, the jth coordinate of which is one if $j \in A$, and zero otherwise. In addition, $\mathbf{1}_X$ is understood as $\mathbf{1}_{\{X\}}$ for a single element $X \in [p]$.

We consider the regression model $Y = f(\boldsymbol{X}) + \epsilon$, where $Y \in \mathbb{R}$ is the response, $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, with $X_j \in [p_j]$ and $j \in [d]$, is a d-dimensional categorical predictor, p_j is a positive integer, and ϵ is a noise term independent of $\boldsymbol{X}$ with mean zero and variance $\sigma^2 > 0$. We do not make other specific assumptions on the distribution of ϵ. We consider only categorical predictors. Continuous predictors (Ding and Ding (2020)) could be discretized in order to use our approach. For example, age can be divided into several groups. We parameterize $f(\boldsymbol{X})$ with $\Gamma(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\beta}$, where $\Gamma(\boldsymbol{X}) \in \mathbb{R}^q$ represents the postulated model structure, consisting of dummy encodings of the original variables and interactions between two or more variables, and $\boldsymbol{\beta} \in \mathbb{R}^q$ is the corresponding vector of unknown coefficients. We show how to encode the model structure using $\Gamma(\cdot)$ by means of three examples.

Example 1 (Additive model). In an additive model, also known as a main-effect model, the regression function is decomposed as $f(\boldsymbol{X}) = \sum_{j=1}^{d} f_j(X_j)$, and $f_j(X_j)$ is called the main effect for the variable X_j. To avoid collinearity, we reparameterize the model by adding a grand mean effect $\beta_0 \in \mathbb{R}$ and the constraint that for any j, $\sum_{k\in[p_j]} f_j(k) = 0$. In other words, $f(\boldsymbol{X}) = \beta_0 + \sum_{j=1}^{d} f_j(X_j)$, and

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_0, f_1(1), \ldots, f_1(p_1 - 1), \ldots, f_d(1), \ldots, f_d(p_d - 1))^{\mathrm{T}},\\ q &= 1 + \sum_{1\le j\le d} (p_j - 1), \quad \Gamma(\boldsymbol{X}) = (1, \boldsymbol{\gamma}_1(X_1), \ldots, \boldsymbol{\gamma}_d(X_d))^{\mathrm{T}},\end{aligned}$$

where $\boldsymbol{\gamma}_j(X_j) = (\mathbb{1}_{X_j=1} - \mathbb{1}_{X_j=p_j}, \ldots, \mathbb{1}_{X_j=p_j-1} - \mathbb{1}_{X_j=p_j})$.

Example 2 (Quadratic model). In addition to the main effects, a quadratic model considers pairwise interaction effects. In other words, $f(\boldsymbol{X}) = \sum_{j=1}^{d} f_j(X_j)$ $+ \sum_{1\le k<l\le d} h_{k,l}(X_k, X_l)$, and $h_{k,l}$ is the interaction effect between X_k and X_l. In addition to the parameterization of the additive model above, we add the constraints $\sum_{s\in[p_l]} h_{k,l}(X_k, s) = 0$ and $\sum_{s\in[p_k]} h_{k,l}(s, X_l) = 0$, for any $X_k \in [p_k]$, $X_l \in [p_l]$, and $k, l \in [d]$. The corresponding model structure is

$$\Gamma(\boldsymbol{X}) = (1, \boldsymbol{\gamma}_1(X_1), \ldots, \boldsymbol{\gamma}_d(X_d), \boldsymbol{\gamma}_1(X_1) \otimes \boldsymbol{\gamma}_2(X_2), \ldots, \boldsymbol{\gamma}_{d-1}(X_{d-1}) \otimes \boldsymbol{\gamma}_d(X_d))^{\mathrm{T}}.$$

The number of free parameters $q = 1 + \sum_{1\le k<l\le d}(p_k p_l - 1)$. The parameter $\boldsymbol{\beta}$ consists of the grand mean β_0, the main effects $f_j(1), \ldots, f_j(p_j - 1)$ for each variable X_j, and the interaction effects $h_{k,l}(1,1), \ldots, h_{k,l}(1, p_l - 1), \ldots, h_{k,l}(p_k - 1, 1), \ldots, h_{k,l}(p_k - 1, p_l - 1)$ for any two variables X_k and X_l.

Example 3 (Saturated model). In a saturated model, also known as a fully interactive model, every level of $\boldsymbol{X}$ corresponds to a free parameter. We have

$$\boldsymbol{\beta} = (f(X_1 = 1, \ldots, X_d = 1), \ldots, f(X_1 = p_1, \ldots, X_d = p_d))^{\mathrm{T}},$$
$$q = \prod_{1\le j\le d} p_j, \quad \Gamma(\boldsymbol{X}) = \bigotimes_{1\le j\le d} \mathbf{1}_{X_j}.$$

Let $p_{\boldsymbol{w}}$ denote the population distribution of $\boldsymbol{X}$, with $pr(\boldsymbol{X} = \boldsymbol{x}) = w_{\boldsymbol{x}}$ for any outcome $\boldsymbol{x}$ of $\boldsymbol{X}$. Furthermore $\boldsymbol{w}$ is the collection of $w_{\boldsymbol{x}}$, which is not required to be known in practice. We assume that the original data $\{\boldsymbol{X}_i, Y_i, i = 1, \ldots, n\}$ are independently and identically distributed, and we obtain set-valued data $\{\boldsymbol{A}_i, Y_i, i = 1, \ldots, n\}$. Here, each observation of $\boldsymbol{A} = (A_1, \ldots, A_d) \in \mathcal{A}$ is a subset associated with $\boldsymbol{X}$, where $\mathcal{A} = \{\boldsymbol{A} : A_j \in 2^{[p_j]}, j \in [d]\}$. The transition law $\boldsymbol{X} \to \boldsymbol{A}$ is explained in Subsection 2.2. The goal is to estimate the regression function f, or equivalently, the model parameters $\boldsymbol{\beta}$.

2.2. Subset-generating process

Here, we describe the transition law $\boldsymbol{X} \to \boldsymbol{A}$ and give some examples. In the data privacy literature, a desirable property is that a privatized observation $\boldsymbol{A}$ does not introduce selective bias related to $\boldsymbol{X}$. That is, we hope that the only information about $\boldsymbol{X}$ from $\boldsymbol{A}$ is that $\boldsymbol{X} \in \boldsymbol{A}$. That is also called the noninformative property. Here, we assume that the transition $\boldsymbol{X} \to \boldsymbol{A}$ is specified by the following mechanisms (Wang and Ding (2021)). First, we consider a one-dimensional variable $X \in [p]$.

Definition 1 (Conditional mechanism). A conditional mechanism determines the transition law $X \to A$ by

$$pr(A = a \mid X = j) = \mu_a \mathbb{1}_{j\in a}, \quad a \subseteq [p],\ j \in [p],$$

where μ_a satisfies $\sum_{a:j\in a} \mu_a = 1$, for all $j \in [p]$.

A particular choice $\{\mu_a, a \in \mathcal{A}\}$ of the conditional mechanism is referred to as a *conditional design*, which we denote as $\{\mu_a, a \in \mathcal{A}\}$.

For the multi-dimensional case, we introduce the following mechanism.

Definition 2 (Product mechanism). A product mechanism determines the transition law of $\boldsymbol{X} \to \boldsymbol{A}$ by

$$pr(\boldsymbol{A} = (a_1, \ldots, a_d) \mid \boldsymbol{X} = (j_1, \ldots, j_d)) = \prod_{l=1}^{d} pr(A_l = a_l \mid X_l = j_l) = \prod_{l=1}^{d} \mu_{a_l} \mathbb{1}_{j_l \in a_l},\ a_l \subseteq [p_l],\ j_l \in [p_l],\ l \in [d],$$

where $\{\mu_{a_l}, a_l \subseteq [p_l]\}$ is a conditional design for $l \in [d]$.

Unless mentioned otherwise, we assume $p_l \geq 4$ to avoid sampling trivial subsets that contain all categories. There are ways of addressing the cases of $p_l = 2$ and 3. First, we can combine two or more predictors into a single predictor so that all have sufficient categories. Second, we can generate dummy categories. For example, when the alphabet of X is $\{1, 2\}$, we independently generate some additional $X \in \{3, 4\}$ from a prespecified distribution. Hence, the alphabet is enlarged to $\{1, 2, 3, 4\}$, and the aforementioned mechanisms can be applied; see Wang and Ding (2021).

A particular case of a product mechanism is that, for a given X, any subset A that contains X, except for the trivial cases $A = \{X\}$ and $A = [p]$, has equal probability of being observed. This corresponds to the following design.

Design 1 (Uniform independence design). A uniform independence design $\{\mu_a, a \in \mathcal{A}\}$ satisfies

$$\mu_a = \begin{cases} 0, & \text{if } |a| = 0, 1, p-1, p \\ \dfrac{1}{2^{p-1} - p - 1}, & \text{otherwise.} \end{cases}$$

Another case is that only subsets that have cardinality k and contain X are chosen with equal probability.

Design 2 (Uniform k-card design.). A uniform k-card design $\{\mu_a, a \in \mathcal{A}\}$ satisfies

$$\mu_a = \begin{cases} \dfrac{(k-1)!(p-k)!}{(p-1)!}, & \text{if } |a| = k \\ 0, & \text{otherwise.} \end{cases}$$

3. Proposed Method

For technical convenience, the predictor variable $\boldsymbol{X}$ is represented by $X \in [p]$, where $p = \prod_{j=1}^{d} p_j$, using the mapping $\boldsymbol{X} \to X : (x_1, \ldots, x_d) \to x_d + \sum_{j=1}^{d-1} \{(x_j -$

$1)\prod_{k=j+1}^{d} p_k\}$, also known as the dictionary order of $\boldsymbol{X}$. For a subset $\boldsymbol{A}$, its corresponding mapping to $\mathbf{1}_A$ is $\mathbf{1}_A = \bigotimes_{j=1}^{d} \mathbf{1}_{A_j}$. We use one-dimensional X from now on, unless otherwise specified.

We propose an estimator for the parameter $\boldsymbol{\beta}$ in the model $Y = \Gamma(X)^{\mathrm{T}}\boldsymbol{\beta} + \epsilon$. The population distribution $\boldsymbol{w}$ is assumed to be known, otherwise, we replace it with a root-n consistent estimator $\widehat{\boldsymbol{w}}$, as described at the end of this section. Let $P = (\Gamma(X = 1), \ldots, \Gamma(X = p))^{\mathrm{T}} \in \mathbb{R}^{p\times q}$, $W \in \mathbb{R}^{p\times p}$ be the diagonal matrix expanded from $\boldsymbol{w}$, $\boldsymbol{q}_i = \mathbf{1}_{A_i}/\mathbf{1}_{A_i}^{\mathrm{T}}\boldsymbol{w}$, $Q = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)^{\mathrm{T}}$, and $\boldsymbol{y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. We propose the following estimator:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}} \sum_{1\le i\le n} \{Y_i - E(Y \mid A_i)\}^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}} \|\boldsymbol{y} - QWP\boldsymbol{\beta}\|_2^2. \tag{3.1}$$

Here, the second equality follows from the noninformative property that

$$\begin{aligned} E(Y \mid A_i) &= \sum_{j\in[p]} pr(X = j \mid A_i) E(Y \mid X = j) \\ &= \sum_{j\in[p]} \mathbb{1}_{j\in A_i} \frac{w_j}{\mathbf{1}_{A_i}^{\mathrm{T}}\boldsymbol{w}} \Gamma(j)\boldsymbol{\beta} = \boldsymbol{q}_i^{\mathrm{T}} WP\boldsymbol{\beta}. \end{aligned}$$

The solution of Equation (3.1) exists and is unique when $QWP \in \mathbb{R}^{n\times q}$ has full column rank; otherwise, we simply take $\widehat{\boldsymbol{\beta}}$ as zeros. In practice, if QWP is near-singular, we suggest adding a regularization term involving $\boldsymbol{\beta}$ to improve the numerical stability, which we elaborate on at the end of this section.

Next, we provide an upper bound for the estimation risk $E\{f(X) - \widehat{f}(X)\}^2$, where $\widehat{f}(X)$ is the estimated value of $f(X)$, and the expectation is taken over the training data $\{Y_i, A_i, i = 1, \ldots, n\}$ and a new predictor X. We make the following technical assumption.

Assumption 1 (Boundedness). *There exist positive values K, L, C, and δ such that*

$$\max_{1\le X\le p} |f(X)| \le K, \ \max_{a:\mu_a>0} |a| \le L, \ \frac{\max_{1\le j\le p} w_j}{\min_{1\le j\le p} w_j} \le C, \ \textit{and} \ \min_{a:\mu_a>0} \mathbf{1}_a^{\mathrm{T}}\boldsymbol{w} \ge \delta.$$

The requirements of Assumption 1 are reasonable. First, we assume that $f(X)$ is bounded, so that the variance of the response given a set-valued observation is not too large. Second, the maximum cardinality of a set-valued observation is upper bounded. If the cardinality of a subset is too large, there will be little information about the original value it contains. Third, we assume

that X has a balanced distribution so that there is no dominating or dominated category. Finally, the condition $\min_{a:\mu_a>0} \mathbf{1}_a^{\mathrm{T}}\boldsymbol{w} \geq \delta$ means that the subset design guarantees the privacy level is at least δ, which is a reasonable setting for privacy purposes (Wang and Ding (2021)).

Let $\widetilde{Q} = \sum_{a\in\mathcal{A}} \mu_a \mathbf{1}_a \mathbf{1}_a^{\mathrm{T}}$ be a matrix that depends only on the subset design, and $\kappa = \kappa(P^{\mathrm{T}}P)$ be the condition number of $P^{\mathrm{T}}P$. The (i,j)th element of $\widetilde{Q}$ is the probability that the subset A contains j when $X = i$. Intuitively, $\widetilde{Q}$ is a measurement of the ambiguity of the subset design. A design with less ambiguity will have $\widetilde{Q}$ closer to an identity matrix. We first introduce Theorem 1, which bounds $E[\{f(X) - \widehat{f}(X)\}^2 \mid A_1, \ldots, A_n]$, the expected loss conditional on the observed set values. Here, X is a new predictor variable, independent of the observations. This quantity differs from a conventional loss or risk, because it averages the noise terms $\{\epsilon_1, \ldots, \epsilon_n\}$ in both the training data and the predictor variables $\{X_1, \ldots, X_n\}$, given the sets $\{A_1, \ldots, A_n\}$.

Theorem 1. *Under Assumption* 1*, for any* $\tau \in (0, 1/2]$*, with probability at least* $1 - \exp[-2n\{(LC)^{-1}\sigma_{\min}(\widetilde{Q})\tau\delta\}^2]$*, we have*

$$E[\{f(X) - \widehat{f}(X)\}^2 \mid A_1, \ldots, A_n] \leq n^{-1}q\kappa LC^2(\sigma^2 + K^2)\sigma_{\min}^{-1}(\widetilde{Q})(1 + 2\tau),$$

for any conditional design $\{\mu_a, a \in \mathcal{A}\}$*.*

Theorem 2. *Under Assumption* 1*, the prediction risk satisfies*

$$E[\{f(X) - \widehat{f}(X)\}^2] \leq 3n^{-1}q\kappa LC^2(\sigma^2 + K^2), \sigma_{\min}^{-1}(\widetilde{Q})$$

for all sufficiently large n *and any conditional design* $\{\mu_a, a \in \mathcal{A}\}$*.*

Corollary 1. *Under Assumption* 1*, if* κ *is upper bounded by a constant and* $\sigma_{\min}(\widetilde{Q})$ *is lower bounded away from zero, we have*

$$E[\{f(X) - \widehat{f}(X)\}^2] = O\left(\frac{q}{n}\right).$$

The proofs of Theorem 1 and 2 are given in the Supplementary Material. Theorem 2 directly implies Corollary 1, which is at the optimal rate of the prediction risk using the original data X_i, Y_i, for $i = 1, \ldots, n$. Recall that $\sigma_{\min}(\widetilde{Q})$ is only associated with the subset design, and the condition number κ represents the inherent property of the model structure $\Gamma(\cdot)$. We show that the conditions of Corollary 1 hold in many common situations with a proper subset design and model structure. We first give the following result.

Proposition 1. *For the uniform independence design (Design* 1*) and uniform k-card design (Design* 2*), we have*

$$\sigma_{\min}^{-1}(\widetilde{Q}) = \prod_{1\le j\le d} \frac{a_j}{a_j - 1},$$

where $a_j = 2$ *for the uniform independence design, and* $a_j = (p_j - 1)/(k - 1)$ *for the uniform k-card design.*

Proposition 1 implies that $\sigma_{\min}^{-1}(\widetilde{Q})$ is at most 2^d, and almost a constant for the uniform two-card design. For example, if we use the uniform 2-card design to privatize a 10 digit phone number, then $\sigma_{\min}^{-1}(\widetilde{Q}) = (9/8)^{10} < 4$. The proof of Proposition 1 is provided in the Supplementary Material. Next, we show that Corollary 1 holds for two widely used model structures.

Example 1: Additive model, continued. It can be shown that for the additive model parameterized as in Example 1, the condition number of the matrix $P^{\mathrm{T}}P$ satisfies $\kappa = \max_{1\le j\le d} p_j$. Thus, when the maximum value of p_j is bounded by a constant, the risk bound is rate optimal. The proof is included in the Supplementary Material.

Example 3: Saturated model, continued. For the saturated model, P is an identity matrix, so $\kappa = 1$ and the risk bound is rate optimal.

Regularized subset least squares estimator. In practice, the matrix QWP is not necessarily a full-column rank matrix. To improve the estimation stability, we suggest using the penalized estimator

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}} \|\boldsymbol{y} - QWP\boldsymbol{\beta}\|_2^2 + \lambda J(\boldsymbol{\beta}),$$

where λ is a tuning parameter and $J(\cdot)$ is a regularization function, such as $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ (ridge-type regression) or $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ (lasso-type regression).

Estimation of population distribution. If the population distribution $\boldsymbol{w}$ is unknown, we can estimate it using the method of moments with the following equation:

$$E(\mathbf{1}_A) = \sum_{a\in\mathcal{A}} pr(A = a)\mathbf{1}_a = \sum_{a\in\mathcal{A}} \mathbf{1}_a \mu_a \mathbf{1}_a^{\mathrm{T}} \boldsymbol{w} = \widetilde{Q}\boldsymbol{w}.$$

In other words, given the set observations $\{A_1, \ldots, A_n\}$, the estimator $\widehat{\boldsymbol{w}}$ is solved from

$$\widetilde{Q}\widehat{\boldsymbol{w}} = n^{-1}\sum_{i=1}^{n}\mathbf{1}_{A_i}.$$

It can be shown that this moment-based estimator is consistent and root-n asymptotically normal under some regularity conditions (Wang and Ding (2021)).

Maximum likelihood method. The proposed subset least squares estimator does not require that we know the distribution of the noise ϵ. If we assume the noise distribution is parameterized, an alternative way is to calculate the maximum likelihood estimator. In their pioneering work, Dempster, Laird and Rubin (1977) studied a general class of incomplete data and proposed the expectation-maximization algorithm to find the maximum likelihood estimator. We extend the concept of incomplete data to our problem, reviewing $\{Y, A\}$ as the incomplete data and $\{Y, X, A\}$ as the complete data. Nevertheless, we do not recommend this method for our problem because of its computational cost and empirical performance, even if the noise distribution assumption is justifiable. The total computational cost of the earlier proposed estimator is $O(nq^2)$. In contrast, the cost of the expectation-maximization algorithm is at least $O(knq^2)$ per iteration, where k is the average cardinality of the observed sets. We implement the expectation-maximization algorithm with Gaussian noise, and find that its empirical performance is undesirable compared with that of the subset least squares method. Details about the algorithm derivation and time complexity are provided in the Supplementary Material.

4. Model Selection

In practice, it is rare that we know the structure $\Gamma(\cdot)$ of the underlying model $f(X) = \Gamma(X)^{\mathrm{T}}\boldsymbol{\beta}$. This section focuses on the selection of an appropriate model, such as the additive or quadratic model, and the selection of variables in the models. Suppose that we have a set of candidate models $\mathcal{M}_n = \{\alpha : \Gamma_\alpha(\cdot)\}$, indexed by α. Let $X \mapsto \widehat{f}_\alpha(X)$ denote the model α estimated using the subset least squares method. Because the observed data are set-valued, we consider the following modified squared error loss:

$$L_n(\alpha) = n^{-1}\sum_{1\le i\le n}\{f(X_i) - \widehat{f}_\alpha(A_i)\}^2, \tag{4.1}$$

where $\widehat{f}_\alpha(A_i)$ is the estimated mean of Y conditional on A_i. The model with the smallest loss is

$$\alpha_n^* = \operatorname*{argmin}_{\alpha\in\mathcal{M}_n} L_n(\alpha).$$

Note that $L_n(\alpha)$ is not available, because it involves the unknown f. Therefore, we propose selecting the model as follows:

$$\widehat{\alpha}_n = \underset{\alpha \in \mathcal{M}_n}{\operatorname{argmin}} S_n(\alpha), \quad \text{where } S_n(\alpha) = n^{-1} \sum_{1 \le i \le n} \{y_i - \widehat{f}_\alpha(A_i)\}^2 + 2n^{-1}\widehat{\sigma}^2 p_n(\alpha),$$

$p_n(\alpha)$ is the number of free parameters of model α, and $\widehat{\sigma}$ is an estimator of the noise level σ. The above selection method is called the modified Mallows's C_p criterion (Mallows (2000)), and is denoted by mC_p.

Theorem 3. *Assume that $E[\{E(Y \mid X) - E(Y \mid A)\}^2]$ is bounded away from zero, $|\mathcal{M}_n|/n \to 0$ and $p/n \to 0$ as $n \to \infty$, and $\widehat{\sigma}$ is a consistent estimator of σ. The model selected by mC_p is asymptotically loss efficient, meaning that $L_n(\alpha_n^*)/L_n(\widehat{\alpha}_n) \to 1$ in probability as $n \to \infty$.*

When p is fixed, the condition of Theorem 3 is automatically satisfied, and thus mC_p is asymptotically loss efficient. A consistent estimator $\widehat{\sigma}$ can be obtained by solving an equation based on the law of total variance of $var(Y)$. More details are included in the Supplementary Material.

5. Experiments

5.1. Simulated data experiments

We first verify the proposed method using four simulated data experiments by showing the estimation error under different model structures and subset designs. We compare six methods: the least squares ("LS-Full"), which uses the complete data $\boldsymbol{X}, Y$ for the estimation; grand mean ("Mean"), which uses only Y; subset least squares ("SLS"); ridge-type subset least squares ("SLS-R"); lasso-type subset least squares ("SLS-L"); and maximum likelihood estimator based on the expectation-maximization algorithm ("MLE"). The lasso-type and ridge-type subset least squares estimators are tuned using five-fold cross-validation with the parameter $\lambda \in \{0, 0.1, 1, 10\}$.

Saturated model. First, we consider a saturated model (Example 3) with dimension $d = 3$, $p_j = 5, j = 1, 2, 3$, Gaussian noise with standard deviation $\sigma = 1$, and the maximum of $|f(\boldsymbol{X})|$ being smaller than $K = 3$. The population distribution $\boldsymbol{w}$ and parameters $\boldsymbol{\beta}$ are drawn element-wisely from a uniform distribution on $[0, 1]$; $\boldsymbol{w}$ is re-scaled to sum to one, and $\boldsymbol{\beta}$ is re-scaled to satisfy $|f(\boldsymbol{X})| \le K$. We generate $n = 5{,}000$ observations of $\boldsymbol{X}, \boldsymbol{A}, Y$ using the product uniform two-card design (Design 2). For each method, we evaluate the estimation loss $E[\{f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X})\}^2 \mid \boldsymbol{A}_i, Y_i, i = 1, \ldots, n]$. The procedure is replicated

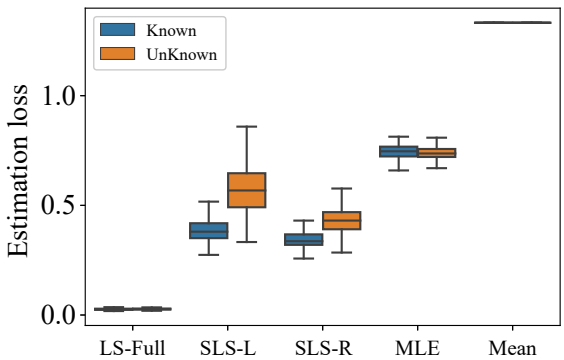

Figure 1. Box plot showing the estimation loss of five methods defined in Subsection 5.1, from 100 replications under the saturated model. The left and right columns of each method correspond to known and unknown population distributions, respectively.

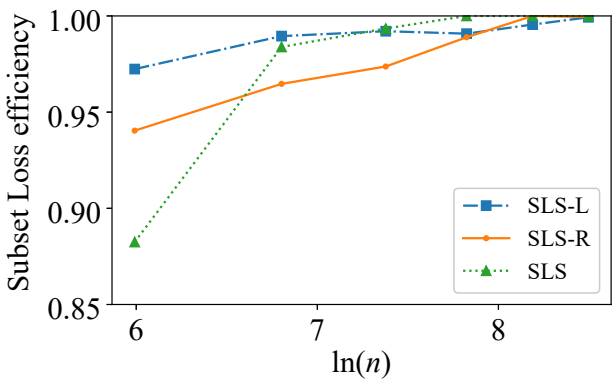

Figure 2. Loss efficiency using mC_p under different sample sizes for the model selection experiment in Subsection 5.1.

$k = 100$ times, with $\boldsymbol{w}$ unknown or known. Because the QWP matrix is highly ill-conditioned, "SLS" is not included in this experiment. A box plot of the loss is reported in Figure 1. We find that "MLE" performs worse than the proposed subset least squares estimators. The estimation loss using the set-valued data ("SLS-L," "SLS-R," or "MLE") is larger than the loss using the original data ("LS-Full'), but smaller than when all predictor information is lacking ("Mean"), as expected. Moreover, the prediction performance of the subset least squares estimators is better when the population distribution is known.

Table 1. Mean estimation loss of six methods under the additive model in Subsection 5.1. Standard errors are all within 0.01 from 100 replications.

$\boldsymbol{w}$	LS-Full	SLS	SLS-L	SLS-R	MLE	Mean
Unknown	0.01	0.11	0.12	0.11	0.55	2.00
Known	0.01	0.07	0.08	0.07	0.57	2.00

Table 2. Mean estimation loss under the additive model in Subsection 5.1, for different average subset cardinalities k. Standard errors are all within 0.03 from 100 replications.

k	SLS	SLS-L	SLS-R	MLE
2	0.07	0.07	0.06	0.57
3	0.21	0.25	0.21	1.12
4	0.74	0.71	0.69	1.73

Additive model. A suitable model can greatly reduce the number of parameters, hence improving the prediction accuracy. Using the same setting as above, we study the performance of subset least squares estimators when the underlying model is additive and the sample size $n = 1{,}000$. The results are summarized in Table 1. Even though the sample size is significantly smaller than that of the saturated model, we find that the estimation loss of the subset least squares estimators is greatly reduced, and is comparable to the loss based on the complete data. However, "MLE" still exhibits a relatively large loss.

Influence of the subset design. We compare the mean estimation loss of 100 replications on the previous additive model for the subset least square estimators and the maximum likelihood estimator, using a uniform k-card design with $k = 2, 3, 4$. The results in Table 2 show that the average cardinality of the subset $\boldsymbol{A}$ influences the estimation accuracy. This aligns with our intuitive understanding that the higher the mean cardinality is, the less information we can learn from each subset observation, and hence the worse the estimation is. It also matches the error bound given by Theorem 2, which is proportional to $\sigma_{\min}^{-1}(\widetilde{Q})$, and Proposition 1 tells us $\sigma_{\min}^{-1}(\widetilde{Q})$ is increasing with k under a uniform k-card design.

Model selection. With all other settings remaining unchanged, we now have a collection of models $\mathcal{M}$, instead of a given true model. Suppose $\mathcal{M}$ includes the grand mean model $Y \sim 1$, main-effect models for each variable $Y \sim X_j$, $j = 1, 2, 3$, quadratic models for any two variables $Y \sim X_k \times X_l$, $1 \le k < l \le 3$, and the saturated model. Let the true model be $Y \sim X_1 \times X_2$. We use the proposed mC_p to perform the model selection. We apply the uniform two-card design to generate the subsets. The average loss efficiency $L_n(\alpha_n^*)/L_n(\widehat{\alpha}_n)$ of 100

Table 3. The mC_p values $S_n(\alpha)$ of five models on the Student Oerformance data set.

"G1∼1"	"G1∼School"	"G1∼Failure"	"G1∼School×Failure"	"G1∼School+Failure"
7.55	6.62	7.19	6.59	6.44

replications against the sample size is presented in Figure 2. Here, the loss is the modified squared error loss defined in Equation (4.1). The loss efficiency is close to one.

5.2. Student performance data

This data set contains information on 649 secondary education students (Cortez and Silva (2008)). We use the students' first-period grades ("G1") in the Portuguese language as the response variable. The data set includes demographic, social, and school-related attributes, among which we choose "School" and "Failure" as the variables of interest. Both variables have four levels. Here, "School" represents the place of a student, and "Failure" is the number of failed courses in the past. An interesting problem is whether the Portuguese language grade is associated with past study performance and potential differences among schools. The original data set collected the exact values of "School" and "Failure." However, historical records of student grades are highly sensitive information, and may be used to identify a particular student. To promote individual privacy, we instead use the subset privacy mechanism to collect them, and apply the proposed subset least squares method for the regression. In this illustrative experiment, we adopt the uniform independence design (Design 1) to generate subsets, and show that the prediction error using the set-valued observations is comparable with that of the regression using the original data.

First, we illustrate the proposed mC_p method for selecting a regression model from the model class {"G1∼1," "G1∼School," "G1∼Failure," "G1∼School× Failure," "G1∼School+Failure"} using a ridge-type subset least squares estimator. Table 3 summarizes the mC_p values of the different models. The additive model "G1∼School+Failure" has the smallest value, and is thus selected. The mC_p values also suggest that both predictors are associated with the response.

Under the selected additive model, we compare the performance of all six methods. We split the whole data set into training and test data sets with a ratio of two to one. The uniform independence design (Design 1) is chosen to generate subsets on the training data set. The evaluation criterion is the mean squared error of the response on the test data set. The average test errors are summarized in Table 4, with $k = 100$ replications. All methods that actively use

Table 4. Mean test error on the Student Performance data set. Permutation standard errors are all within 0.06 from 100 replications.

Method	LS-Full	SLS	SLS-L	SLS-R	MLE	Mean
Loss	5.85	6.21	6.10	5.99	6.01	7.54

the complete or incomplete data have significantly smaller test errors than that of the grand mean method. In addition, we observe that the subset-valued data using the proposed method have a similar test error to that of the complete data. This is because the test error involves noise in the response. Such noise can be large compared with the estimation error from using incomplete data. Thus, we may not need to collect exact sensitive individual information, such as the history of failed classes, to study statistical relationships.

6. Conclusion

Motivated by set-valued predictors obtained from privacy-oriented data collection mechanisms, we propose a subset least squares method for regression. We derive an upper bound of the prediction risk for the proposed estimator, and show that it is rate-optimal under mild conditions. In addition, we develop an asymptotically loss-efficient method mC_p for model selection. The subset least squares method shows promising performance compared with that of the MLE in our numerical studies. Our numerical results indicate that when the regression model is complex relative to the sample size, the subset least square estimator may perform poorly, owing to the ill-conditioned design matrix. In contrast, the regularized subset least squares stabilizes the estimation, and hence has a much smaller prediction risk. Moreover, the set-valued data using the proposed method tend to have similar estimation risks to those observed from the original data, which justifies the use of subset privacy.

Some interesting problems are left for future work. First, the asymptotic distributions for the regularized subset least squares estimators remain unclear. Second, we would like to explore an inference on the parameters $\boldsymbol{\beta}$, which seems highly nontrivial.

Supplementary Material

The online Supplementary Material includes detailed proofs, additional experiments, extended discussions, and the code for the experiments.

Acknowledgements

We sincerely thank the two anonymous reviewers, associate editor, and co-editor for their helpful comments and suggestions. Ganghua Wang was supported by the Army Research Laboratory and the Army Research Office under grant number W911NF-20-1-0222.

References

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proc. VLDB*, 901–909.

Chaum, D., Crépeau, C. and Damgard, I. (1988). Multiparty unconditionally secure protocols. In *Proc. STOC*, 11–19.

Cohen, A. and Nissim, K. (2020). Towards formalizing the GDPR's notion of singling out. *Proc. Natl. Acad. Sci.* **117**, 8344–8352.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proc. FBTC*, 5–12.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39**, 1–38.

Ding, J. and Ding, B. (2020). "To tell you the truth" by interval-private data. In *Proceedings of 2020 IEEE International Conference on Big Data (IEEE Big Data)*, 25–32. IEEE, New York.

Dua, D. and Graff, C. (2017). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory Crypto. Conf.*, 265–284. Springer, Berlin.

Enserink, M. and Chin, G. (2015). The end of privacy. *Science* **347**, 490–491.

Mallows, C. L. (2000). Some comments on C_p. *Technometrics* **42**, 87–94.

Rocher, L., Hendrickx, J. M. and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**, 1–9.

Wang, G. and Ding, J. (2021). Subset privacy: Draw from an obfuscated urn. *arXiv preprint arXiv:2107.02013*.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63–69.

Yao, A. C. (1982). Protocols for secure computations. In *Proc. SFCS*, 160–164. IEEE, New York.

Ganghua Wang

School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA.

E-mail: wang9019@umn.edu

Jie Ding
School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA.
E-mail: dingj@umn.edu
Yuhong Yang
School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA.
E-mail: yangx374@umn.edu

(Received September 2021; accepted February 2022)