Statistica Sinica Preprint No: SS-2021-0332	
Title	Regression with Set-Valued Categorical Predictors
Manuscript ID	SS-2021-0332
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0332
Complete List of Authors	Ganghua Wang, Jie Ding and Yuhong Yang
Corresponding Author	Yuhong Yang
E-mail	yangx374@umn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Regression with Set-Valued Categorical Predictors

Ganghua Wang, Jie Ding and Yuhong Yang

School of Statistics, 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. The classical regression analyses such as least squares are not applicable since the set-valued predictors only carry partial information about the original values. We propose a computationally efficient subset least squares method to perform regression for such data. We establish upper bounds of the prediction loss and risk in terms of the subset structure, the model structure, and the data dimension. The error rates are shown to be optimal under 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 datasets demonstrate the promising performance of the proposed method.

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

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

1. Introduction

Data privacy has become an emerging societal concern (Enserink and Chin, 2015; Cohen and Nissim, 2020). For example, Rocher et al. (2019) showed that even if the common identifiers of each individual are removed, 99.98% of Americans could be correctly re-identified with only 15 demographic attributes such as family size and vehicle type. Therefore, privacy-preserving methods to protect individual identification and sensitive data values are receiving increasing attention nowadays. In particular, a popular choice is that the data owner no longer releases the exact value X of each individual. Instead, a quantity Z relevant to X will be used to enhance individual privacy. Along this direction, several privacy-preserving methods have been proposed, including differential privacy (Dwork et al., 2006) that uses the randomized response technique (Warner, 1965) or adds noises to X, k-anonymity (Aggarwal, 2005) that groups X with similar values to a representative value, and secure multi-party computing (Yao, 1982; Chaum et al., 1988) that encrypts X by cryptography techniques.

In the development of data privacy techniques, a critical use scenario concerns the data collection procedure. Recently, a new mechanism named subset privacy (Wang and Ding, 2021) was proposed to address the challenge of private data collection. Specifically, the data collector, such as a service provider, will only collect a set A that contains the original value X held by the subject, such as

an individual user. For example, in a study of the income with respect to race in the Adult Dataset (Dua and Graff, 2017), a data collector could perform a survey that only collects a set of races instead of the exact race from each participant. The generation of A can be easily realized through a survey-based system, such as the independent design introduced in Subsection 2.2.

Subset privacy provides a privacy guarantee against de-identification. Nevertheless, it is highly nontrivial to perform regression and prediction using the new format of data. This paper considers the 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, since 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 the set-valued data. We derive a closed-form solution for the optimization problem above. We establish an upper bound of the prediction risk and show that it is rate-optimal in some circumstances of particular interest. Examples are additive models where the effect of each variable is independent of

others and saturated models that all variables are interactive with one another. We also discuss some practical strategies to improve the numerical stability and leverage fast matrix operation. Furthermore, to select a model from different combinations of variables or interaction orders, we propose a selection method and prove its asymptotic efficiency under some conditions. Finally, experiments based on simulated and real data are performed to verify the developed 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 $|A|$, 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 the condition number of a positive definite matrix M are denoted as $\sigma_{\max}(M)$, $\sigma_{\min}(M)$, $\kappa(M) = \sigma_{\max}(M)\sigma^{-1}_{\min}(M)$, respectively. We sometimes represent a finite set $A \subseteq [p]$ with a vector $\mathbf{1}_A \in \{0, 1\}^p$, whose jth coordinate is one if $j \in A$ and zero otherwise. Also, $\mathbf{1}_X$ 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 $X_j \in [p_j], j \in [d]$ 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. In this paper, we only consider categorical predictors. In the presence of continuous-valued predictors (Ding and Ding, 2020, 2021), one may discretize those predictors to use our approach. For example, age could 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 the dummy encoding of original variables and interactions between two or more variables, and $\boldsymbol{\beta} \in \mathbb{R}^q$ is the corresponding unknown coefficients vector. We illustrate how to encode the model structure through $\Gamma(\cdot)$ by three examples.

Example 1 (Additive model) *In an additive model, also known as the 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 variable* X_j. *To avoid collinearity, we reparameterize the model by adding a grand mean effect* $\beta_0 \in \mathbb{R}$ *and the constraints 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*

$$\boldsymbol{\beta} = (\beta_0, f_1(1), \ldots, f_1(p_1 - 1), \ldots, f_d(1), \ldots, f_d(p_d - 1))^{\mathrm{T}},$$

$$q = 1 + \sum_{1\le j\le d} (p_j - 1), \quad \Gamma(\boldsymbol{X}) = (1, \boldsymbol{\gamma}_1(X_1), \ldots, \boldsymbol{\gamma}_d(X_d))^{\mathrm{T}},$$

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) *Besides 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 constraints that* $\sum_{s\in[p_l]} h_{k,l}(X_k, s) = 0$ *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 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 the fully interactive model, every level of* $\boldsymbol{X}$ *corresponds to a free parameter. We have*

$$\boldsymbol{\beta} = (f(X_1 = 1, \ldots, X_d = 1), \ldots, f(X_1 = p_1, \ldots, X_d = p_d))^{\mathrm{T}},$$

$$q = \prod_{1\le j\le d} p_j, \Gamma(\boldsymbol{X}) = \bigotimes_{1\le j\le d} \mathbf{1}_{X_j}.$$

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 $\boldsymbol{X}$, and $\boldsymbol{w}$ 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. We only 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 of $\boldsymbol{X} \to \boldsymbol{A}$ will be elaborated next 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

We describe the transition law of $\boldsymbol{X} \to \boldsymbol{A}$ in this subsection and give some examples. In the data privacy literature, a desirable property is that a privatized observation $\boldsymbol{A}$ does not introduce selective bias regarding $\boldsymbol{X}$. Namely, we hope that the only information about $\boldsymbol{X}$ from $\boldsymbol{A}$ is that $\boldsymbol{X} \in \boldsymbol{A}$, also called the non-informative property. More specifically, we assume that the transition of $\boldsymbol{X} \to \boldsymbol{A}$ 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 of $X \to A$ *by*

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

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

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

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*

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

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 two techniques to address the case of $p_l = 2$ or 3. First, we may combine two or more predictors into a single predictor so that all of them have enough 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 pre-specified distribution. Hence, the alphabet is enlarged to $\{1, 2, 3, 4\}$, and the aforementioned mechanisms can be applied. We refer to Wang and Ding (2021) for the detailed description.

A particular case of 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 to be observed. This corresponds to the following design.

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

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

Another particular case is that only the subsets that have cardinality k and contain X will be chosen with equal probability.

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

$$\mu_a = \begin{cases} 1/\binom{p-1}{k-1}, & \text{if } |a| = k \\ 0, & \text{otherwise.} \end{cases}$$

3. Proposed Method

For technical convenience, the predictor variable $\boldsymbol{X}$ is represented by one-dimensional $X \in [p]$, 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 will use one-dimensional X from now on, unless otherwise specified.

We will 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.

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

Here, the second equality follows from the non-informative property that

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

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 numerical stability, which will be elaborated 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*

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

The requirements of Assumption 1 are reasonable. Firstly, we assume that $f(X)$ is bounded, so that the variance of the response given a set-valued observation is not too large. Secondly, the maximum cardinality of set-valued observation is upper bounded. If the cardinality of a subset is too large, there will be little information regarding the original value it contains. Thirdly, we assume that X has a balanced distribution so that there is no dominating or dominated category. Finally, the condition $\min_{a:\mu_a>0}\mathbf{1}_a^{\mathrm{T}}\boldsymbol{w}\ge\delta$ 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 only depends 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 the following Theorem 1 that bounds $E[\{f(X) - \widehat{f}(X)\}^2 \mid A_1, \ldots, A_n]$, which is the expected loss conditional on the observed set values. Here, X is a new predictor variable independent of the observations. This quantity is different from conventional loss or risk, since it averages both the noise terms $\{\epsilon_1, \ldots, \epsilon_n\}$ in training data and 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, for 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 Theorems 1 and 2 are in the supplementary document. Theorem 2 directly implies Corollary 1, which is at the optimal rate of the prediction risk using the original data $X_i, Y_i, i = 1, \ldots, n$. 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 will 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 2-card design. For example, if we use the uniform 2-card design to privatize a ten-digit phone number, then $\sigma_{\min}^{-1}(\widetilde{Q}) = (9/8)^{10} < 4$. The proof of Proposition 1 is in the supplementary document. 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 matrix $P^{\mathrm{T}}P$ satisfies $\kappa = \max_{1\le j\le d} p_j$. So, 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 document.*

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 necessary a full column rank matrix. To promote estimation stability, we suggest use the penalized estimator

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

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) and $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ (lasso-type regression).

Estimation of population distribution. If the population distribution $\boldsymbol{w}$ is unknown, we can estimate it by 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 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 the distribution of the noise ϵ to be known. If we assume the noise distribution is parameterized, an alternative way is to calculate the maximum likelihood estimator. The pioneering work of Dempster et al. (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, regarding $\{Y, A\}$ 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 em-

pirical 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 implemented the Expectation-Maximization algorithm with Gaussian noise and found its empirical performance was undesirable compared with the subset least squares method. Details about the algorithm derivation and time complexity are in the supplementary document.

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, or 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 estimated model α by subset least squares method. Since the observed data are set-valued, we consider the following modified squared error loss

$$L_n(\alpha) = n^{-1} \sum_{1 \le i \le n} \{f(X_i) - \widehat{f}_\alpha(A_i)\}^2, \tag{4.2}$$

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

$$\alpha_n^* = \arg\min_{\alpha \in \mathcal{M}_n} L_n(\alpha).$$

Since $L_n(\alpha)$ is not available as it involves the unknown f, we propose the following way to select the model.

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

$p_n(\alpha)$ 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 named the modified Mallows's C_p criterion (Mallows, 2000), 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 the 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 document.

5. Experiments

5.1 Simulated Data Experiments

We first verify the developed method in four simulated data experiments by showing the estimation error under different model structures and subset designs. There are six methods in comparison, including the least squares ('LS-Full') that uses the complete data $\boldsymbol{X}, Y$ for estimation, grand mean ('Mean') that only uses 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 by 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 element-wisely drawn from a uniform distribution on $[0, 1]$. The $\boldsymbol{w}$ is re-scaled to sum to one, and $\boldsymbol{\beta}$ is re-scaled to satisfy $|f(\boldsymbol{X})| \leq K$. We generate $n = 5000$ observations of $\boldsymbol{X}, \boldsymbol{A}, Y$ using the product uniform 2-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 pro-

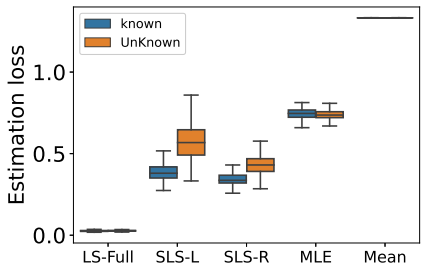

Figure 1: Box-plot showing the estimation loss of five methods defined in Subsection 5.1, from 100 replications under the saturated model. The left and right columns of each method correspond to known and unknown population distributions, respectively.

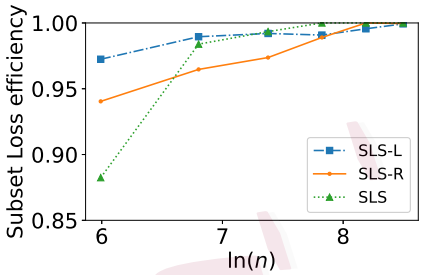

Figure 2: Loss efficiency using mC_p under different sample sizes, for the model selection experiment in Subsection 5.1.

cedure is replicated $k = 100$ times given $\boldsymbol{w}$ unknown or known. Since the QWP matrix is highly ill-conditioned, 'SLS' is not included in this experiment. The box-plot of the loss is reported in Figure 1. We find that 'MLE' is 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 the case that all predictor information is lacking ('Mean'), as is expected. Moreover, the prediction performance for subset least squares estimators is better when the population distribution is known.

Additive model. A suitable model can greatly reduce the number of parameters, hence improving the prediction accuracy. With the same setting as above,

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

we study the performance of subset least squares estimators when the underlying model is additive and sample size $n = 1000$. The results are summarized in Table 1. Even though the sample size is significantly fewer than the saturated model, we find that the estimation loss of the subset least squares estimators is greatly reduced, which is comparable to the loss using the complete data. However, 'MLE' has a relatively large loss.

Influence of the subset design. We compare the mean estimation loss of 100 replications on the previous additive model for subset lease square estimators and the maximum likelihood estimator, using uniform k-card design with $k = 2, 3, 4$, respectively. It can be seen from Table 2 that the average cardinality of the subset $\boldsymbol{A}$ influences the estimation accuracy. The results align with

the intuition that the higher the mean cardinality is, the less information we can learn from each subset observation, 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 uniform k-card design.

Model selection. While other settings remain unchanged, now we 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 $Y \sim X_k \times X_l$, $1 \leq k < l \leq 3$, and the saturated model. Let the true model be $Y \sim X_1 \times X_2$. We use the proposed mC_p to perform model selection. The uniform 2-card design is applied for subset generation. The average loss efficiency $L_n(\alpha_n^*)/L_n(\widehat{\alpha}_n)$ of 100 replications against sample size is presented in Figure 2. Here, the loss is the modified squared error loss defined in Equation (4.2). The loss efficiency is close to one.

5.2 Student Performance dataset

This dataset contains 649 students in secondary education (Cortez and Silva, 2008). We use the students' first-period grades ('G1') in the Portuguese language as the response variable. The dataset 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, the 'School' represents the place of a student, while the 'Failure' is the number of failed courses in the past. An interesting problem is whether the Portuguese language grade is associated with the past study performance and potential differences among schools. The original dataset collected the exact value of 'School' and 'Failure'. However, the historical record of student grades is highly sensitive information, and the school information may be used to identify a particular student. To promote individual privacy, we can instead use the subset privacy mechanism to collect them and apply the proposed subset least squares method for 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 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'} with a ridge-type subset least squares estimator. Table 3 summarizes mC_p values of 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

Table 3: The mC_p values $S_n(\alpha)$ of five models on the student performance dataset.

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

Table 4: Mean test error on the Student Performance dataset. 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

methods. We split the whole dataset into training and test datasets with a ratio of two to one. The uniform independence design (Design 1) is chosen to generate subsets on the training dataset. The evaluation criterion is the mean squared error of the response on the test dataset. The average test errors are summarized in Table 4 with $k = 100$ replications. It can be seen that all methods actively using the complete or incomplete data have significantly smaller test errors than the grand mean method. Also, we observe that the subset-valued data using the proposed method has a similar test error compared with using the complete data. This is because the test error involves noise in the response. Such noise can be large compared with the estimation error brought by using incomplete data. The result indicates that we may not need to collect the exact sensitive individual information such as the history of failed classes to study statistical relationships.

6. Concluding remarks

Motivated by the set-valued predictors obtained from privacy-oriented data collection mechanisms, we propose the 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. Additionally, we develop an asymptotically loss efficient method mC_p for model selection. The subset least squares method has shown promising performances compared with the maximum likelihood estimator in our numerical studies. The numerical results indicate that when the regression model is complex relative to the sample size, the subset least square estimator may perform poorly due to the ill-conditioned design matrix. In contrast, the regularized subset least squares can stabilize the estimation and hence have a much smaller prediction risk. Moreover, the set-valued data using the proposed method tend to have similar estimation risks as if we observed the original data, which justifies the use of subset privacy.

There are some interesting problems left for future work. First, the asymptotic distributions for the regularized subset least squares estimators remain unclear. Second, one may explore the inference on the parameters $\boldsymbol{\beta}$, which seems highly non-trivial.

Acknowledgement

We sincerely thank the two anonymous reviewers, the associate editor and the co-editor for their insightful comments, which have helped us improve the paper. Ganghua Wang was supported by the Army Research Laboratory and the Army Research Office under grant number W911NF-20-1-0222.

Supplementary Document

The supplementary document includes detailed proofs, additional experiments, extended discussions, and experimental codes.

References

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proc. VLDB*, Volume 5, pp. 901–909.

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

Cohen, A. and K. Nissim (2020). Towards formalizing the gdpr's notion of singling out. *Proc. Natl. Acad. Sci. 117*(15), 8344–8352.

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

REFERENCES

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B 39*(1), 1–22.

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

Ding, J. and B. Ding (2021). Interval privacy: A framework for privacy-preserving data collection. *arXiv preprint arXiv:2106.09565*.

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

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

Enserink, M. and G. Chin (2015). The end of privacy. *Science 347*(6221), 490–491.

Mallows, C. L. (2000). Some comments on cp. *Technometrics 42*(1), 87–94.

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

Wang, G. and J. Ding (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. Am. Stat. Assoc. 60*(309), 63–69.

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