A note on centering in subsample selection for linear regression

HaiYing Wang
University of Connecticut
haiying.wang@uconn.edu

July 25, 2022

Abstract

Centering is a commonly used technique in linear regression analysis. With centered data on both the responses and covariates, the ordinary least squares estimator of the slope parameter can be calculated from a model without the intercept. If a subsample is selected from a centered full data, the subsample is typically un-centered. In this case, is it still appropriate to fit a model without the intercept? We show that the least squares estimator on the slope parameter obtained from a model without the intercept is unbiased and it has a smaller variance covariance matrix in the Loewner order than that obtained from a model with the intercept. We further show that for noninformative weighted subsampling when a weighted least squares estimator is used, using the full data weighted means to relocate the subsample improves the estimation efficiency.

Keywords: Estimation efficiency, Ordinary Least Squares, Variance, Weighted Least Squares

1 Introduction

When fitting a linear regression model, the slope parameter is often the main focus and the intercept may not be of interest. In this scenario, a widely used trick to simplify the calculation is to center the data so that a linear model without the intercept can be used to calculate the slope estimator. If the intercept is need, e.g. for prediction, it can be calculated using the mean response and the means of the covariates together with the slope estimator.

To be specific, consider the following linear regression model for the full data $\mathcal{D}_n = (\mathbf{X}, \mathbf{y})$ of sample size n,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, ..., y_n)^{\mathrm{T}}$ is the response vector, $\mathbf{Z} = (\mathbf{1}_n \ \mathbf{X})$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ with α_0 and $\boldsymbol{\beta}$ being the intercept and slope vector respectively, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^{\mathrm{T}}$ is the model error satisfying $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

To estimate $\boldsymbol{\theta}$, the ordinary least squares (OLS) estimator with the full data $\mathcal{D}_n$ is

$$\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}. \tag{2}$$

The mean response is $\bar{y} = n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{y}$ and the vector of column means for $\mathbf{X}$ is $\bar{\mathbf{x}} = n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{1}_n$, so the centered data can be written as

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^{\mathrm{T}} = (\mathbf{I}_n - \boldsymbol{J}_n)\mathbf{X}, \tag{3}$$

$$\mathbf{y}_c = \mathbf{y} - \mathbf{1}_n\bar{y} = (\mathbf{I}_n - \boldsymbol{J}_n)\mathbf{y}, \tag{4}$$

where $\boldsymbol{J}_n = n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$. With centered data, it is well know that $\hat{\boldsymbol{\beta}}$ can be calculated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}\mathbf{X}_c\mathbf{y}_c, \tag{5}$$

and if α is of interest, $\hat{\alpha} = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

In recent years, due to the challenge of rapidly increasing volumes of data, one may have to select a small subsample from the full data so that available computational resources at hand can fully analyze the subsample and useful information can be drawn. Investigations in this direction have been fruitful for linear regression. A popular technique is to use statistical leverage scores or their variants to construct subsampling probabilities, see Drineas et al. (2012), Ma et al. (2015), Yang et al. (2015), Nie et al. (2018), and the references therein. Wang et al. (2019) proposed the information based optimal subdata selection (IBOSS) method and the proposed deterministic selection algorithm has a high estimation efficiency and a linear computational time complexity. Pronzato & Wang (2021) developed an online subsample selection algorithm that achieve the optimal variance under a general optimality criterion. Yu & Wang (2022) recommended using leverage scores to select subsamples deterministically.

An interesting question raises for centering in subsampling: if the full data is centered, do we have to center the subsample to calculate the slope estimate if the model does not contain an intercept? We will show in this short note that it is better to not center the subsample in this case. Since for a deterministically selected subsample, the OLS is applied (e.g., Wang et al. 2019, Pronzato & Wang 2021), while for a randomly selected subsample with nonuniform probabilities the weighted least squares (WLS) is often fitted (e.g., Yang et al. 2015, Ai et al. 2021, Zhang et al. 2021), we discuss these two types of estimators in Sections 2 and 3, respectively. Some numerical evaluations are provided in Section 4 and more technical details are given in the Appendix.

2 Deterministic selection with OLS

Let $(\mathbf{X}^*,\ \mathbf{y}^*)$ denote the subsample of size r corresponding to the un-centered full data $(\mathbf{X},\mathbf{y})$, and $(\mathbf{X}_c^*,\ \mathbf{y}_c^*)$ be the subsample corresponding to the centered full data $(\mathbf{X}_c,\mathbf{y}_c)$, i.e., $(\mathbf{X}_c^* = \mathbf{X}^* - \mathbf{1}_n\bar{\mathbf{x}}^{\mathrm{T}},\ \mathbf{y}_c^* = \mathbf{y}^* - \mathbf{1}_n\bar{y})$. In this section, we assume that the selection rule is nonrandom and it may depend on $\mathbf{X}$ but it does not depend on the response $\mathbf{y}$. Under this assumption, the subsample follows a linear regression model

$$\mathbf{y}^* = \mathbf{Z}^*\boldsymbol{\theta} + \boldsymbol{\varepsilon}^*, \tag{6}$$

where $\mathbf{Z}^* = (\mathbf{1}_r\ \mathbf{X}^*)$, $\mathbb{E}(\boldsymbol{\varepsilon}^*) = \mathbf{0}$, $\mathbb{V}(\boldsymbol{\varepsilon}^*) = \sigma^2\mathbf{I}_r$, $\mathbf{1}_r$ is a $r\times 1$ vector of ones, and $\mathbf{I}_r$ is the $r\times r$ identity matrix. The OLS based on the subsample is

$$\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}, \tilde{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{Z}^{*\mathrm{T}}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\mathrm{T}}\mathbf{y}^*. \tag{7}$$

Clearly, $\mathbf{X}_c^*$ and $\mathbf{y}_c^*$ may not be centered, i.e., their means are not zero. Can we simply use $(\mathbf{X}_c^*, \mathbf{y}_c^*)$ to fit a model without the intercept to estimate $\boldsymbol{\beta}$, i.e., use

$$\tilde{\boldsymbol{\beta}}_c = (\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}\mathbf{y}_c^* \tag{8}$$

to estimate $\boldsymbol{\beta}$? The answer is yes. Here, $\tilde{\boldsymbol{\beta}}_c$ is not only unbiased but also has a smaller variance compared with $\tilde{\boldsymbol{\beta}}$.

The unbiasedness of $\tilde{\boldsymbol{\beta}}_c$ has been noticed in Yu & Wang (2022). We show it here for completeness. Note that

$$\mathbf{y}_c^* = \mathbf{y}^* - \bar{y}\mathbf{1}_r = \alpha\mathbf{1}_r + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* - (\alpha + \bar{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta} + \bar{\varepsilon})\mathbf{1}_r = \mathbf{X}_c^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* - \bar{\varepsilon}\mathbf{1}_r, \tag{9}$$

where $\bar{\varepsilon}$ is the average of $\varepsilon_1, ..., \varepsilon_n$ and therefore $\bar{\varepsilon} \sim \mathbb{N}(0, n^{-1}\sigma^2)$. We then know that

$$\tilde{\boldsymbol{\beta}}_c = (\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}(\mathbf{X}_c^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* - \bar{\varepsilon}\mathbf{1}_r) = \boldsymbol{\beta} + (\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}(\boldsymbol{\varepsilon}^* - \bar{\varepsilon}\mathbf{1}_r). \tag{10}$$

Thus, we have the unbiasedness $\mathbb{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ from the above representation.

The following proposition shows that $\tilde{\boldsymbol{\beta}}_c$ has a smaller variance than $\tilde{\boldsymbol{\beta}}$ in the Loewner order.

Proposition 1. *Assume that* $\mathbf{Z}^*$ *is full rank. Let* $\bar{\mathbf{x}}^*$ *be the vector of subsample covariate means, i.e.,* $\bar{\mathbf{x}}^* = r^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{1}_r$*. The variances of* $\tilde{\boldsymbol{\beta}}_c$ *and* $\tilde{\boldsymbol{\beta}}$ *satisfy that*

$$\mathbb{V}(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}^*) - \mathbb{V}(\tilde{\boldsymbol{\beta}}_c \mid \mathbf{X}^*) = \left(\frac{r}{1-d} + \frac{r^2}{n}\right) \times \sigma^2(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}(\bar{\mathbf{x}}^* - \bar{\mathbf{x}})^{\otimes 2}(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}, \qquad (11)$$

where $d = r(\bar{\mathbf{x}}^* - \bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}(\bar{\mathbf{x}}^* - \bar{\mathbf{x}})^{\mathrm{T}} < 1$*, and the notation* $^{\otimes 2}$ *means* $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^{\mathrm{T}}$ *for a vector or matrix* $\mathbf{v}$*.*

Remark 1. *The smaller variance of* $\tilde{\boldsymbol{\beta}}_c$ *indicates that even if the full data is not centered, it would be better to shift the subsample by* $(\bar{\mathbf{x}}, \bar{y})$ *and then fit a model without the intercept than fitting a model with an intercept directly.*

Remark 2. *The matrix on the right hand side of* (11) *is of rank one, so we do not expect the difference between* $\mathbb{V}(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}^*)$ *and* $\mathbb{V}(\tilde{\boldsymbol{\beta}}_c \mid \mathbf{X}^*)$ *to be large, especially when the dimension of* $\boldsymbol{\beta}$ *is high. Nevertheless, we recommend* $\tilde{\boldsymbol{\beta}}_c$ *because its calculation is as easy as that of* $\tilde{\boldsymbol{\beta}}$*.*

If the intercept α is of interest, it can be estimated by using the subsample means $\bar{\mathbf{x}}^*$ and $\bar{y}^*$. However, using the full data means $\bar{\mathbf{x}}$ and $\bar{y}$ is usually significantly more efficient. This is also observed in the numerical examples of Wang et al. (2019). Specifically, the estimator

$$\tilde{\alpha}_c = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_c \qquad (12)$$

is typically much more efficient than $\tilde{\alpha}$ defined in (7). By direct calculations, we obtain that

$$\mathbb{V}(\tilde{\alpha}_c \mid \mathbf{X}^*, \bar{\mathbf{x}}) = \sigma^2\Big\{\frac{1}{n} + \bar{\mathbf{x}}^{\mathrm{T}}\mathbb{V}(\tilde{\boldsymbol{\beta}}_c \mid \mathbf{X}^*)\bar{\mathbf{x}}\Big\} \quad \text{and} \quad \mathbb{V}(\tilde{\alpha} \mid \mathbf{X}^*) = \sigma^2\Big\{\frac{1}{r} + \bar{\mathbf{x}}^{*\mathrm{T}}\mathbb{V}(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}^*)\bar{\mathbf{x}}^*\Big\}.$$

Wang et al. (2019) have shown that $\mathbb{V}(\tilde{\boldsymbol{\beta}} \mid \mathbf{X}^*)$ converges to zero faster than r^{-1} if the support of covariate distribution is not bounded. For this scenario, the dominating term in $\mathbb{V}(\tilde{\alpha} \mid \mathbf{X}^*)$ is $\sigma^2 r^{-1}$, while the dominating term in $\mathbb{V}(\tilde{\alpha}_c \mid \mathbf{X}^*, \bar{\mathbf{x}})$ is often $\bar{\mathbf{x}}^{\mathrm{T}}\mathbb{V}(\tilde{\boldsymbol{\beta}}_c \mid \mathbf{X}^*)\bar{\mathbf{x}}$ which converges to zero faster than $\sigma^2 r^{-1}$.

3 Nonuniform random subsampling with WLS

A large class of subsample selection methods are through nonuniform random sampling such as the leverage sampling and optimal sampling. In this scenario the inverse probability WLS approach is typically applied on the subsample and the subsample estimator is often proposed as an "estimator" of the full data estimator. For this type of approaches, the exact variance of the resulting estimator may not be defined so our discussions are on the asymptotic variance which we use $\mathbb{V}_a$ to denote. Properties of random subsampling estimators are more complicated and we focus on the scenario that $r = o(n)$ so that the contribution of the randomness from the full data to the asymptotic variance is negligible.

Assume that a subsample of size r is randomly selected according to nonuniform probabilities $\pi_1, ..., \pi_n$. Here we abuse the notation and use $\mathbf{X}^*$, $\mathbf{y}^*$, $\boldsymbol{\varepsilon}^*$, and $\mathbf{Z}^*$ again to denote subsample quantities. We need to point out that (6) does not hold for a randomly selected subsample.

Let $\mathbf{w} = (w_1, ..., w_n)^{\mathrm{T}}$ be the vector of weights where w_i's are proportional to π_i^{-1}'s. To ease the discussion, we assume that $\|\mathbf{w}\| = 1$. Let $\mathbf{W}$ be the corresponding $n \times n$ diagonal weighting matrix, i.e., $\mathbf{w} = \mathbf{W}\mathbf{1}_n$, and let $\mathbf{w}^*$ and $\mathbf{W}^*$ be the weighting vector and matrix for the selected subsample, respectively. The WLS estimator is

$$\tilde{\boldsymbol{\theta}}_w = (\tilde{\alpha}_w, \tilde{\boldsymbol{\beta}}_w^{\mathrm{T}})^{\mathrm{T}} = (\mathbf{Z}^{*\mathrm{T}}\mathbf{W}^*\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\mathrm{T}}\mathbf{W}^*\mathbf{y}^*. \tag{13}$$

It has been shown that $\tilde{\boldsymbol{\theta}}_w$ is asymptotically unbiased towards the full data OLS $\hat{\boldsymbol{\theta}}$, and its asymptotic variance has been derived in the literature (see Ma et al. 2015, Yu et al. 2022, Wang et al. 2022, etc). In our notation, the asymptotic variance of $\tilde{\boldsymbol{\theta}}_w$ given $\mathcal{D}_n$ is

$$\mathbb{V}_a(\tilde{\boldsymbol{\theta}}_w \mid \mathcal{D}_n) = \frac{C}{r}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{E}^2\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}, \tag{14}$$

where $C = \sum_{i=1}^n w_i^{-1}$ and $\mathbf{E} = \mathrm{diag}(e_1, ..., e_n)$ with e_i's being the residuals from the full data OLS estimator. Note that $\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{E}^2\mathbf{Z} = \sum_{i=1}^n w_i e_i^2 \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}$, so if the weights probabilities w_i's do not involve e_i's, then $\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{E}^2\mathbf{Z} = \sigma^2 \sum_{i=1}^n w_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}\{1 + o_P(1)\}$ under reasonable conditions. Thus the asymptotic variance can be written as

$$\mathbb{V}_a(\tilde{\boldsymbol{\theta}}_w \mid \mathcal{D}_n) = \frac{C\sigma^2}{r}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}, \tag{15}$$

from which we obtain that

$$\mathbb{V}_a(\tilde{\boldsymbol{\beta}}_w \mid \mathcal{D}_n) = \frac{C\sigma^2}{r}(\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\mathbf{X}_c(\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}. \tag{16}$$

If the centered data is sampled and used to construct an estimator of $\boldsymbol{\beta}$ directly

$$\tilde{\boldsymbol{\beta}}_{w,uwc} = (\mathbf{X}_c^{*\mathrm{T}}\mathbf{W}^*\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}\mathbf{W}^*\mathbf{y}_c^*, \tag{17}$$

then since the full data means $\bar{\mathbf{x}}$ and $\bar{y}$ are nonrandom functions of the full data $\mathcal{D}_n$, existing results (e.g., Ai et al. 2021, Yu et al. 2022, Wang et al. 2022) are applicable to $\tilde{\boldsymbol{\beta}}_{w,uwc}$. This tells us that $\tilde{\boldsymbol{\beta}}_{w,uwc}$ is asymptotically unbiased towards $\hat{\boldsymbol{\beta}}$ and its asymptotic variance is the same as that of $\tilde{\boldsymbol{\beta}}_w$ shown in (16). Thus for noninformative random subsampling with WLS, if the original full data is centered, we can ignore the intercept as well.

Interestingly, if we use weighted means of the full data to relocate the subsample, we have an improved estimator of $\boldsymbol{\beta}$. Denote $\bar{y}_w = \mathbf{w}^{\mathrm{T}}\mathbf{y}$ and $\bar{\mathbf{x}}_w = \mathbf{X}^{\mathrm{T}}\mathbf{w}$ as the weighted mean response and weighed mean covariate vector, respectively. Let $\mathbf{y}_{wc} = \mathbf{y} - \bar{y}_w\mathbf{1}_n$ and $\mathbf{X}_{wc} = \mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}_w^{\mathrm{T}}$ be the centered response vector and design matrix using the weighted means, respectively, and let $\mathbf{y}_{wc}^*$ and $\mathbf{X}_{wc}^*$ be the corresponding selected subsample quantities. A better subsample estimator for $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}}_{wc} = (\mathbf{X}_{wc}^{*\mathrm{T}}\mathbf{W}^*\mathbf{X}_{wc}^*)^{-1}\mathbf{X}_{wc}^{*\mathrm{T}}\mathbf{W}^*\mathbf{y}_{wc}^*. \tag{18}$$

Again, since $\bar{y}_w$ and $\bar{\mathbf{x}}_w$ are nonrandom functions of the full data $\mathcal{D}_n$, existing results show that $\tilde{\boldsymbol{\beta}}_{wc}$ is asymptotically unbiased with asymptotic variance

$$\mathbb{V}_a(\tilde{\boldsymbol{\beta}}_{wc} \mid \mathcal{D}_n) = \frac{C\sigma^2}{r}(\mathbf{X}_{wc}^{\mathrm{T}}\mathbf{X}_{wc})^{-1}\mathbf{X}_{wc}^{\mathrm{T}}\mathbf{W}\mathbf{X}_{wc}(\mathbf{X}_{wc}^{\mathrm{T}}\mathbf{X}_{wc})^{-1}. \tag{19}$$

The following result shows that $\tilde{\boldsymbol{\beta}}_{wc}$ has a smaller asymptotic variance than $\tilde{\boldsymbol{\beta}}_w$.

Proposition 2. *The asymptotic variances in* (16) *and* (19) *satisfy that* $\mathbb{V}_a(\tilde{\boldsymbol{\beta}}_{wc} \mid \mathcal{D}_n) \le \mathbb{V}_a(\tilde{\boldsymbol{\beta}}_w \mid \mathcal{D}_n)$*, and the equality holds if* $\bar{\mathbf{x}}_{wc} = \bar{\mathbf{x}}$.

Remark 3. *The above result relays on the asymptotic representation in* (15) *which requires that the weights do not involve the residuals. If the weights are inverses of the sampling probabilities, this means that the sampling probabilities are noninformative, i.e., they do not dependent on the responses. For informative subsampling such as the A- or L- optimal subsampling (Ai et al. 2021, Wang et al. 2022), there is no definite ordering between* $\mathbb{V}_a(\tilde{\boldsymbol{\beta}}_{wc} \mid \mathcal{D}_n)$ *and* $\mathbb{V}_a(\tilde{\boldsymbol{\beta}}_w \mid \mathcal{D}_n)$.

Similar to the case of deterministic selection, if the intercept is of interest, it can be estimated by

$$\tilde{\alpha}_{wc} = \bar{y}_w - \bar{\mathbf{x}}_w^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{wc} \quad \text{or} \quad \tilde{\alpha}_{w,uc} = \bar{y} - \bar{\mathbf{x}}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{wc}. \tag{20}$$

Note that the $\tilde{\alpha}_w$ defined in (13) satisfies $\tilde{\alpha}_w = \bar{y}_w^* - \bar{\mathbf{x}}_w^*\tilde{\boldsymbol{\beta}}_w$. Both $\bar{y}_w^*$ and $\bar{\mathbf{x}}_w^*$ are random given $\mathcal{D}_n$, and thus they both contribute to the variation of $\tilde{\alpha}_w$ in approximating $\hat{\alpha}$. For $\tilde{\alpha}_{wc}$ (or $\tilde{\alpha}_{w,uc}$), neither $\bar{y}_w$ nor $\bar{\mathbf{x}}_w$ (or $\bar{y}$ nor $\bar{\mathbf{x}}$) is random given $\mathcal{D}_n$, so the only source of variation in approximating $\hat{\alpha}$ is $\tilde{\boldsymbol{\beta}}_{wc}$. Thus $\tilde{\alpha}_{wc}$ is often significantly more efficient than $\tilde{\alpha}_w$.

4 Numerical comparisons

We provide some numerical simulation results that compare the performance of estimators discussed in previous sections. We generated data from model (1) with $n = 10^5$, $\alpha = 1$,

$\boldsymbol{\beta} = \mathbf{1}_{19}$, and $\boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ with $\sigma^2 = 9$. To generate rows of $\mathbf{X}$, we considered the following three distributions. Case 1: multivariate normal distribution $\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$, Case 2: multivariate log normal distribution $\exp\{\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})\}$, and Case 3: multivariate t distribution with degrees of freedom 5 $\mathbb{T}(\mathbf{0}, \boldsymbol{\Sigma}, 5)$. Here the (i, j)-th element of $\boldsymbol{\Sigma}$ is $0.5^{|i-j|}$ in all cases. We implemented three subsampling methods: uniform sampling, IBOSS (Wang et al. 2019), and leverage sampling (Ma et al. 2015). We run the simulation for 1,000 times to calculate the empirical mean squared errors (MSE) reported in Table 1.

We see that for the slope parameter, although not very significant, an estimator based on centered full data (un-centered subsample) has a smaller MSE than the counterpart based on un-centered full data (centered subsample). For the intercept, the estimator based on the full data means is better than the counterpart based on the subsample only, and the improvement is quite significant.

Table 1: Empirical MSEs of subsample estimators for the intercept and slope.*

		Uniform		IBOSS		Leverage	
		WI	WOI	WI	WOI	WI	WOI
Case 1	α	88.422	76.342	27.094	26.608	83.463	73.260
	$\boldsymbol{\beta}$	18.745	18.482	12.874	12.740	17.834	17.641
Case 2	α	81.065	79.155	12.693	12.579	152.350	22.085
	$\boldsymbol{\beta}$	0.683	0.668	0.064	0.062	0.400	0.396
Case 3	α	54.336	45.025	1.505	0.655	67.008	15.929
	$\boldsymbol{\beta}$	13.418	13.259	0.564	0.558	10.520	10.500

* "WI": α and $\boldsymbol{\beta}$ are estimated from a model with an intercept;
"WOI": $\boldsymbol{\beta}$ is estimated from a model without an intercept, and α is estimated using the full data means.

5 Summary

For a subsample selected from centered full data, although the subsample is un-centered, it is better to fit a model without an intercept to estimate the slope parameter if the subsampling rule does not depend on the responses. If the full data is un-centered, it would be better to shift the location of the data by the full data (weighted) means for the OLS (WLS) and then fit a model without an intercept.

Acknowledgement

This work is supported by NSF grant CCF 2105571.

A Technical details

Proof of Proposition 1. Let $\boldsymbol{S}$ be the $r \times n$ selection matrix consisting of zeros and ones that maps the full data to the subsample, i.e., $\mathbf{y}^* = \boldsymbol{S}\mathbf{y}$ and $\mathbf{X}^* = \boldsymbol{S}\mathbf{X}$. We know that

$$\begin{align}
\mathbf{X}_c^* &= \mathbf{X}^* - \mathbf{1}_r\bar{\mathbf{x}}^{\mathrm{T}} = \mathbf{X}^* - n^{-1}\mathbf{1}_r\mathbf{1}_n^{\mathrm{T}}\mathbf{X} = \boldsymbol{S}(\mathbf{I}_n - \boldsymbol{J}_n)\mathbf{X}, \tag{21}\\
\mathbf{y}_c^* &= \mathbf{y}^* - \bar{y}\mathbf{1}_r = \boldsymbol{S}\mathbf{y} - n^{-1}\mathbf{1}_r\mathbf{1}_n^{\mathrm{T}}\mathbf{y} = \boldsymbol{S}(\mathbf{I}_n - \boldsymbol{J}_n)\mathbf{y}. \tag{22}
\end{align}$$

Thus,

$$\mathbb{V}(\mathbf{y}_c^*) = \mathbb{V}\{\boldsymbol{S}(\mathbf{I}_n - \boldsymbol{J}_n)\boldsymbol{\varepsilon})\} = \sigma^2(\mathbf{I}_r - rn^{-1}\boldsymbol{J}_r), \tag{23}$$

and therefore

$$\begin{align}
\mathbb{V}(\tilde{\boldsymbol{\beta}}_c) &= (\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}\mathbb{V}(\mathbf{y}^*)\mathbf{X}_c^*(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1} \tag{24}\\
&= \sigma^2(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}(\mathbf{I}_r - rn^{-1}\boldsymbol{J}_r)\mathbf{X}_c^*(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1} \tag{25}\\
&= \sigma^2(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1} - \sigma^2 rn^{-1}(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}\mathbf{X}_c^{*\mathrm{T}}\boldsymbol{J}_r\mathbf{X}_c^*(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}. \tag{26}
\end{align}$$

Let $\mathbf{X}_{cr}^*$ be the centered subsample design matrix with the subsample means, i.e., $\mathbf{X}_{cr}^* = \mathbf{X}^* - \mathbf{1}_r\bar{\mathbf{x}}^{*\mathrm{T}}$. From the facts that $\mathbf{1}_r^{\mathrm{T}}\mathbf{X}_{cr}^* = \mathbf{0}^{\mathrm{T}}$ and

$$\mathbf{X}_c^* = \mathbf{X}^* - \mathbf{1}_r\bar{\mathbf{x}}^{\mathrm{T}} = \mathbf{X}_{cr}^* - \mathbf{1}_r(\bar{\mathbf{x}} - \bar{\mathbf{x}}^*)^{\mathrm{T}}, \tag{27}$$

we know

$$\mathbf{X}_c^{*\mathrm{T}}\boldsymbol{J}_r\mathbf{X}_c^* = r^{-1}\mathbf{X}_c^{*\mathrm{T}}\mathbf{1}_r\mathbf{1}_r^{\mathrm{T}}\mathbf{X}_c^* = r(\bar{\mathbf{x}} - \bar{\mathbf{x}}^*)^{\otimes 2}. \tag{28}$$

Thus (26) and (28) give

$$\mathbb{V}(\tilde{\boldsymbol{\beta}}_c) = \sigma^2(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1} - \sigma^2 r^2 n^{-1}(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}^*)^{\otimes 2}(\mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^*)^{-1}. \tag{29}$$

From (7), the variance of $\tilde{\boldsymbol{\beta}}$ is

$$\mathbb{V}(\tilde{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}_{cr}^{*\mathrm{T}}\mathbf{X}_{cr}^*)^{-1}. \tag{30}$$

Note that (27) implies

$$\mathbf{X}_{cr}^{*\mathrm{T}}\mathbf{X}_{cr}^* = \mathbf{X}_c^{*\mathrm{T}}\mathbf{X}_c^* - r\{\bar{\mathbf{x}}^* - \bar{\mathbf{x}}\}^{\otimes 2}. \tag{31}$$

Thus we obtain

$$(\mathbf{X}_{cr}^{*\mathrm{T}}\mathbf{X}_{cr}^{*})^{-1} = (\mathbf{X}_{c}^{*\mathrm{T}}\mathbf{X}_{c}^{*})^{-1} + \frac{r(\mathbf{X}_{c}^{*\mathrm{T}}\mathbf{X}_{c}^{*})^{-1}(\bar{\mathbf{x}}^{*} - \bar{\mathbf{x}})^{\otimes 2}(\mathbf{X}_{c}^{*\mathrm{T}}\mathbf{X}_{c}^{*})^{-1}}{1-d}, \tag{32}$$

where $d = r(\bar{\mathbf{x}}^{*} - \bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{X}_{c}^{*\mathrm{T}}\mathbf{X}_{c}^{*})^{-1}(\bar{\mathbf{x}}^{*} - \bar{\mathbf{x}})^{\mathrm{T}}$. Here $1-d$ must be positive because (31) implies that $(\mathbf{X}_{cr}^{*\mathrm{T}}\mathbf{X}_{cr}^{*})^{-1} \geq (\mathbf{X}_{c}^{*\mathrm{T}}\mathbf{X}_{c}^{*})^{-1}$ and they are both positive-definite.

Combining (29), (30), and (32) finishes the proof.

□

Proof of Proposition 2. For positive definite matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_1$, and $\mathbf{B}_2$, if $\mathbf{A}_1 \leq \mathbf{A}_2$ and $\mathbf{B}_1 \geq \mathbf{B}_2$, then

$$\begin{aligned}\mathbf{B}_2^{1/2}\mathbf{A}_1^{-1}\mathbf{B}_2^{1/2} \geq \mathbf{B}_2^{1/2}\mathbf{A}_2^{-1}\mathbf{B}_2^{1/2} \Rightarrow& (\mathbf{B}_2^{1/2}\mathbf{A}_1^{-1}\mathbf{B}_2^{1/2})^2 \geq (\mathbf{B}_2^{1/2}\mathbf{A}_2^{-1}\mathbf{B}_2^{1/2})^2 \\ \Rightarrow& \mathbf{A}_1^{-1}\mathbf{B}_2\mathbf{A}_1^{-1} \geq \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{A}_2^{-1},\end{aligned}$$

so $\mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{A}_1^{-1} \geq \mathbf{A}_1^{-1}\mathbf{B}_2\mathbf{A}_1^{-1} \geq \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{A}_2^{-1}$. Thus we only need to prove that

$$\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c \leq \mathbf{X}_{wc}^{\mathrm{T}}\mathbf{X}_{wc}, \tag{33}$$

$$\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\mathbf{X}_c \geq \mathbf{X}_{wc}^{\mathrm{T}}\mathbf{W}\mathbf{X}_{wc}, \tag{34}$$

and the equality in both hold if $\bar{\mathbf{x}}_{wc} = \bar{\mathbf{x}}$. The proof finishes from the fact that

$$\mathbf{X}_c^{\mathrm{T}}\mathbf{W}\mathbf{X}_c - \mathbf{X}_{wc}^{\mathrm{T}}\mathbf{W}\mathbf{X}_{wc} = \mathbf{X}^{\mathrm{T}}(\mathbf{w} - n^{-1}\mathbf{1}_n)^{\otimes 2}\mathbf{X} = (\bar{\mathbf{x}}_{wc} - \bar{\mathbf{x}})^{\otimes 2} \geq \mathbf{0} \tag{35}$$

$$\mathbf{X}_{wc}^{\mathrm{T}}\mathbf{X}_{wc} - \mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c = n\mathbf{X}^{\mathrm{T}}(\mathbf{w} - n^{-1}\mathbf{1}_n)^{\otimes 2}\mathbf{X} = n(\bar{\mathbf{x}}_{wc} - \bar{\mathbf{x}})^{\otimes 2} \geq \mathbf{0}, \tag{36}$$

which can be verified by inserting $\mathbf{X}_c = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}})\mathbf{X}$ and $\mathbf{X}_{wc} = (\mathbf{I}_n - \mathbf{1}_n\mathbf{w}^{\mathrm{T}})\mathbf{X}$. □

References

Ai, M., Yu, J., Zhang, H. & Wang, H. (2021), 'Optimal subsampling algorithms for big data regressions', *Statistica Sinica* **31**(2), 749–772.

Drineas, P., Magdon-Ismail, M., Mahoney, M. & Woodruff, D. (2012), 'Faster approximation of matrix coherence and statistical leverage.', *Journal of Machine Learning Research* **13**, 3475–3506.

Ma, P., Mahoney, M. & Yu, B. (2015), 'A statistical perspective on algorithmic leveraging', *Journal of Machine Learning Research* **16**, 861–911.

Nie, R., Wiens, D. P. & Zhai, Z. (2018), 'Minimax robust active learning for approximately specified regression models', *Canadian Journal of Statistics* **46**(1), 104–122.

Pronzato, L. & Wang, H. (2021), 'Sequential online subsampling for thinning experimental designs.', *Journal of Statistical Planning and Inference* **212**, 169 – 193.

Wang, H., Yang, M. & Stufken, J. (2019), 'Information-based optimal subdata selection for big data linear regression', *Journal of the American Statistical Association* **114**(525), 393–405.

Wang, J., Zou, J. & Wang, H. (2022), 'Sampling with replacement vs poisson sampling: a comparative study in optimal subsampling', *IEEE Transactions on Information Theory* .
URL: *https://doi.org/10.1109/TIT.2022.3176955*

Yang, T., Zhang, L., Jin, R. & Zhu, S. (2015), An explicit sampling dependent spectral error bound for column subset selection, *in* 'Proceedings of The 32nd International Conference on Machine Learning', pp. 135–143.

Yu, J. & Wang, H. (2022), 'Subdata selection algorithm for linear model discrimination', *Statistical Papers* .

Yu, J., Wang, H., Ai, M. & Zhang, H. (2022), 'Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data', *Journal of the American Statistical Association* **117**(537), 265–276.

Zhang, T., Ning, Y. & Ruppert, D. (2021), 'Optimal sampling for generalized linear models under measurement constraints', *Journal of Computational and Graphical Statistics* **30**(1), 106–114.