Published in final edited form as:

Can J Stat. 2018 March ; 46(1): 123–146. doi:10.1002/cjs.11330.

Online updating method with new variables for big data streams

Chun Wang¹, Ming-Hui Chen², Jing Wu³, Jun Yan², Yuping Zhang², and Elizabeth Schifano^{2,*}

¹Liberty Mutual Insurance, Boston, MA, USA

²Department of Statistics, University of Connecticut, Storrs, CT, USA

³Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI, USA

Abstract

For big data arriving in streams, online updating is an important statistical method that breaks the storage barrier and the computational barrier under certain circumstances. In the regression context, online updating algorithms assume that the set of predictor variables does not change, and consequently cannot incorporate new variables that may become available midway through the data stream. A naive approach would be to discard all previous information and start updating with new variables from scratch. We propose a method that utilizes the information from earlier data in the online updating algorithm with bias corrections to improve efficiency. The method is developed for linear models first, and then extended to estimating equations for generalized linear models. Closed-form expressions for the efficiency gain over the naive approach are derived in a particular linear model setting. We compare the performance of our proposed bias-correcting approach and the naive approach in simulation studies with data generated from a normal linear model and a logistic regression model. The method is applied to a study on airline delay, where reasons for delays were only available more recently, starting in 2003.

Keywords

Added variable; estimating equation; data compression; regression

1. INTRODUCTION

Big data are data on a massive scale in terms of volume, intensity, and complexity that exceed the ability of standard analytic tools (e.g., Snijders et al., 2012). Standard statistical models have difficulty with fitting big data because the data can exceed even a super computer's memory. In a recent review, Wang et al. (2016) grouped the the statistical/ computational methodologies into three categories: sub-sampling based approaches (e.g., Liang et al., 2013; Ma et al., 2013; Maclaurin and Adams, 2014; Kleiner et al., 2014), divide and conquer approaches (e.g., Lin and Xi, 2011; Neiswanger et al., 2013; Scott et al., 2013; Chen and Xie, 2014; Song and Liang, 2014), and the online updating approach (Schifano et al., 2016). The online updating approach is distinct from the other two because it targets

* Author to whom correspondence may be addressed. elizabeth.schifano@uconn.edu.

problems where data arrive in streams or large chunks, and addresses statistical problems in an updating framework without storage requirements for previous data. Our focus is on this approach when new predictor variables become available at some point along the data stream.

In a typical regression setting, emergence of new variables is common, due to, for example, negligence in data collection in the past, change of protocol, or advances in technology. Our motivating application is a risk analysis of airline delays. US carriers were not required to report causes of flight delays to the Bureau of Transportation Statistics until June 2003 (e.g., Rupp, 2007). In linear or logistic regression models for flight delays, with these newly available variables only starting from June 2003, a naive application of the online updating method would start from scratch, discarding the possibly useful information in the data before June 2003 and, hence, losing efficiency in statistical inferences. Similar situations arise in many other applications, as well. In medical studies, new medical devices can monitor measurements that have never been seen before (e.g., Hood et al., 2004). In financial analyses at the company level, new companies become public every month, releasing unprecedented information at the initial public offering (e.g., Certo, 2003). In auto insurance pricing, insurance companies can take advantage of data from new devices capturing driver behavior (e.g., Desyllas and Sako, 2013).

We propose a novel extension of the online updating algorithm from Schifano et al. (2016) for the added variable situation. Under the online updating setting, statistical models such as the generalized linear model (GLM) or, more generally, estimating equations (EE) are fitted to blocks of data from the data stream. At each block, only a few lower-dimensional quantities such as coefficient estimates and variance estimates are updated; previous data and previous coefficient and variance estimates need not be stored. When new variables become available, under the assumption that the true model contains these new variables, the cumulative coefficient estimates (based on all previous data) for the existing predictor variables are biased. To improve efficiency using the existing information, we correct the bias in the cumulative coefficient estimates and the corresponding variance estimates for the existing variables; the online updating process then resumes at the next block using both the new and existing predictor variables. Our setting is different from “missing covariate designs” in meta-analysis (Liu et al., 2015), in which case a covariate is fixed at one value in one study (block) but at another value in another study (block), so that there is no “design” for this covariate, whose effect may be of interest but is not estimable using only the current study (block). In our setting, the covariate is completely missing (instead of fixed at a certain value) before it becomes available at a new block. The mapping method of Liu et al. (2015), which depends on the relationship between estimable parameters and the whole parameter vector, is not applicable because the relationship cannot be established with missing covariate values (not missing covariate design).

Since the data from previous blocks is not stored, the bias correction is constructed using the block where the new variables emerge, i.e., the ‘changing block’. The correction is the difference in the estimates of the coefficients of the existing variables with and without the new variables in the model. This problem has been studied as comparing regression coefficients between nested models (Clogg et al., 1995; Allison, 1995; Yan et al., 2013),

which is an important subject of mediation analysis with wide applications in social sciences. The recent asymptotic validation in Yan et al. (2013) provides a unified, intuitive way to make inferences about the difference in the linear model (LM) setting with clustered data. We extend the method of Yan et al. (2013) to the GLM setting, and in the LM case, our derivation matches the closed-form expressions in Clogg et al. (1995) and Allison (1995). Notably, however, Clogg et al. (1995), Allison (1995), and Yan et al. (2013) only focus on nested models within a *single* block of data. The online updating framework, in contrast, assumes limited data storage/access of *multiple* blocks of data, typically with different sample sizes, and thus requires sample size adjustments in the correction formulas. With the bias correction estimated from the changing block, we develop all the required quantities to adjust the cumulative coefficient estimates and variance estimates based on the previous data. We show that in a three-variable regression setting where y is the response variable, x is the existing variable, and z is the newly added covariate, the variance of the bias-corrected online updating estimator is always smaller than the variance of the naive estimator (based on data from the changing block only) as long as the squared correlation coefficient between x and z is smaller than $1/2$. The efficiency gain remains under a mild condition when the squared correlation coefficient between x and z is greater than $1/2$. Further, more accumulated data in the past yields more efficiency gains from the bias-corrected online updating estimator over the naive estimator that ignores previous information.

The rest of paper is organized as follows. In Section 2, under the linear regression setting, we propose our bias correction formula along with the associated online updating algorithm, and illustrate the benefit of our method in a three-variable-and-two-block case under some mild conditions. In Section 3, we generalize the method to the GLM setting with a unified EE approach. Simulation results are reported in Section 4. A case study of airline delays is presented in Section 5. A discussion concludes in Section 6.

2. LINEAR MODEL

2.1. Online Updating Method

Suppose data arrive in blocks $\kappa = 1, 2, \dots$, and that in block κ , there are n_κ independent observations $\{(y_{\kappa i}, \mathbf{x}_{\kappa i}) : i = 1, 2, \dots, n_\kappa\}$, where $y_{\kappa i}$ is the response variable, and $\mathbf{x}_{\kappa i}$ is a $p \times 1$ covariate vector. The working linear regression model is

$$y_{\kappa i} = \mathbf{x}_{\kappa i}^\top \mathbf{b} + \varepsilon_{\kappa i}, \quad (1)$$

where $\mathbf{b}$ is a p -dimensional vector of regression coefficients, $\varepsilon_{\kappa i}$ are uncorrelated with mean 0 and standard deviation σ_ε . Let $\mathbf{y}_\kappa = (y_{\kappa 1}, \dots, y_{\kappa n_\kappa})^\top$ and $\mathbf{X}_\kappa = (\mathbf{x}_{\kappa 1}, \dots, \mathbf{x}_{\kappa n_\kappa})^\top$. Assume that design matrix $\mathbf{X}_\kappa$ is of full rank $p < n_\kappa$. The least squares (LS) estimate of $\mathbf{b}$ and an appropriate estimator of σ_ε^2 from block κ are

$$\hat{\mathbf{b}}_\kappa = (\mathbf{X}_\kappa^\top \mathbf{X}_\kappa)^{-1} \mathbf{X}_\kappa^\top \mathbf{y}_\kappa \quad \text{and} \quad \hat{\sigma}_{\varepsilon, n_\kappa, \kappa}^2 = \frac{1}{n_\kappa - p} \mathbf{y}_\kappa^\top (\mathbf{I}_{n_\kappa} - \mathbf{X}_\kappa (\mathbf{X}_\kappa^\top \mathbf{X}_\kappa)^{-1} \mathbf{X}_\kappa^\top) \mathbf{y}_\kappa,$$

where $\mathbf{I}_a$ is the $a \times a$ identity matrix.

Schifano et al. (2016) formulated a computationally efficient and minimally storage-intensive procedure for estimating cumulative quantities for big stream data. Although the algorithm was designed for stream data, it also works for fixed sample size problems. The online-updated estimators are identical to the divide-and-recombine estimators (e.g., Lin and Xi, 2011) under the LM setting. At the κ^{th} block, the cumulative coefficient estimator of $\mathbf{b}$ (based on all data from blocks 1 through κ) is

$$\tilde{\mathbf{b}}_\kappa = (\mathbf{X}_\kappa^\top \mathbf{X}_\kappa + \mathbf{V}_{\kappa-1})^{-1} (\mathbf{X}_\kappa^\top \mathbf{X}_\kappa \hat{\mathbf{b}}_\kappa + \mathbf{V}_{\kappa-1} \tilde{\mathbf{b}}_{\kappa-1}), \quad (2)$$

where $\tilde{\mathbf{b}}_0 = \mathbf{0}$, $\mathbf{V}_0 = \mathbf{0}_p$ is a $p \times p$ matrix of zeros, and $\mathbf{V}_\kappa = \sum_{\ell=1}^{\kappa} \mathbf{X}_\ell^\top \mathbf{X}_\ell$ for $\kappa = 1, 2, \dots$. This can be viewed as an average of previous cumulative estimate $\tilde{\mathbf{b}}_{\kappa-1}$ and the estimate based on the current block $\hat{\mathbf{b}}_\kappa$, weighted by the inverses of their variance matrices

$$\tilde{\mathbf{b}}_\kappa = [\text{Var}^{-1}(\hat{\mathbf{b}}_\kappa) + \text{Var}^{-1}(\tilde{\mathbf{b}}_{\kappa-1})]^{-1} [\text{Var}^{-1}(\hat{\mathbf{b}}_\kappa) \hat{\mathbf{b}}_\kappa + \text{Var}^{-1}(\tilde{\mathbf{b}}_{\kappa-1}) \tilde{\mathbf{b}}_{\kappa-1}],$$

where $\text{Var}(\hat{\mathbf{b}}_\kappa) = \sigma_\varepsilon^2 (\mathbf{X}_\kappa^\top \mathbf{X}_\kappa)^{-1}$ and $\text{Var}(\tilde{\mathbf{b}}_{\kappa-1}) = \sigma_\varepsilon^2 \mathbf{V}_{\kappa-1}^{-1}$. The online-updated estimator of the residual sum of squares (SSE) based on cumulative data is

$$\text{SSE}_\kappa = \text{SSE}_{\kappa-1} + \text{SSE}_{n_\kappa, \kappa} + \tilde{\mathbf{b}}_{\kappa-1}' \mathbf{V}_{\kappa-1}^{-1} \tilde{\mathbf{b}}_{\kappa-1} + \hat{\mathbf{b}}_\kappa' \mathbf{X}_\kappa' \mathbf{X}_\kappa \hat{\mathbf{b}}_\kappa - \tilde{\mathbf{b}}_{\kappa-1}' \mathbf{V}_{\kappa-1}^{-1} \tilde{\mathbf{b}}_{\kappa-1},$$

where $\text{SSE}_{n_\kappa, \kappa}$ is the residual sum of squares from the κ th block, with corresponding residual mean square $\hat{\sigma}_{\varepsilon, n_\kappa, \kappa}^2 = \text{SSE}_{n_\kappa, \kappa} / (n_\kappa - p)$. The estimate of σ_ε^2 based on the cumulative data is then $\hat{\sigma}_{\varepsilon, \kappa}^2 = \text{SSE}_\kappa / (N_\kappa - p)$ where $N_\kappa = \sum_{\ell=1}^{\kappa} n_\ell$ for $\kappa = 1, 2, \dots$. Thus, the variance of the cumulative estimator $\tilde{\mathbf{b}}_\kappa$ can be estimated by $\hat{\sigma}_{\varepsilon, \kappa}^2 \mathbf{V}_\kappa^{-1}$. The online updating method is elegant because it only uses quantities $(\tilde{\mathbf{b}}_{\kappa-1}, \text{SSE}_{\kappa-1}, \mathbf{V}_{\kappa-1}, N_{\kappa-1})$, which are of low dimension from the previous accumulation point, to compute $\tilde{\mathbf{b}}_\kappa$ and $\hat{\sigma}_{\varepsilon, \kappa}^2$.

2.2. Challenges from New Covariates

Suppose that the set of predictor variables $\mathbf{X}_\kappa$ is available for each block $\kappa = 1, 2, \dots$, but a new set of variables $\mathbf{Z}_\kappa = (\mathbf{z}_{\kappa 1}, \dots, \mathbf{z}_{\kappa n_\kappa})^\top$ become available only after k blocks, where $\mathbf{z}_{\kappa i}$ is a $q \times 1$ vector. That is, $\mathbf{Z}_\kappa$ is only observed for $\kappa = k+1, k+2, \dots$. The new design matrix $(\mathbf{X}_\kappa, \mathbf{Z}_\kappa)$ is assumed to have full column rank $p+q$. Suppose that, instead of (1), the true regression model is

$$y_{\kappa i} = \mathbf{x}_{\kappa i}^\top \boldsymbol{\beta} + \mathbf{z}_{\kappa i}^\top \boldsymbol{\theta} + \nu_{\kappa i}, \quad i = 1, 2, \dots, n_\kappa, \quad (3)$$

where $v_{\kappa i}$ are uncorrelated with mean 0 and standard deviation σ_v , β is a $p \times 1$ coefficient vector of $\mathbf{X}_\kappa$ and θ is a $q \times 1$ coefficient vector of $\mathbf{Z}_\kappa$ for all $\kappa = 1, 2, \dots$

Let H_R be the reduced model (1) and H_F the full model (3). As in Clogg et al. (1995), Allison (1995), and Yan et al. (2013), we assume the true model is H_F . Under true model H_F the updated estimate $\tilde{\mathbf{b}}_k$ from the first k blocks has not been estimating β in general. It is a biased estimator for β ; this kind of bias is called omitted variable bias in standard texts (e.g., Greene, 2003, p.148). Let $(\mathcal{X}_k, \mathcal{Z}_k)$ denote the cumulative predictor variables until block k , where $\mathcal{Z}_k$ is unobserved. Define $\delta = \lim_{N_k \rightarrow \infty} (\mathcal{X}_k^\top \mathcal{X}_k)^{-1} \mathcal{X}_k^\top \mathcal{Z}_k \theta$ in probability and $\mathbf{b} = \beta + \delta$. Then $\mathbf{b}$ is the limit of $\tilde{\mathbf{b}}_k$ as $N_k \rightarrow \infty$. With the new data on $\mathbf{Z}$ available starting from block $k+1$, a naive approach would be to start the updating process from scratch at block $k+1$, throwing away the information from earlier blocks. Note that at block $k+1$, we may obtain $\hat{\beta}_{k+1}$ and $\hat{\theta}_{k+1}$ from fitting H_F , as well as $\hat{\mathbf{b}}_{k+1}$ from fitting H_R . With H_F being the true model, $\hat{\beta}_{k+1}$ and $\hat{\theta}_{k+1}$ are unbiased estimators of β and θ , respectively, and $\hat{\mathbf{b}}_{k+1}$ is an unbiased estimator of $\mathbf{b}$.

Our goal is to fix the online updating at block $k+1$ such that the algorithm can continue as usual after this block. For θ , since there is no previous data for $\mathbf{Z}$, its estimate can only be obtained from current block $k+1$: $\tilde{\theta}_{k+1} = \hat{\theta}_{k+1}$. For β , however, it is more challenging. We propose to 1) use information from block $k+1$ to obtain $\tilde{\beta}_k$ and its variance; and 2) update from $\tilde{\beta}_k$ to $\tilde{\beta}_{k+1}$ and compute its variance and covariance with $\tilde{\theta}_{k+1}$. We will show that the cumulative estimate $\tilde{\beta}_{k+1}$ is still unbiased for β and the corresponding variance $\text{Var}(\tilde{\beta}_{k+1})$ is smaller in most practical applications than $\text{Var}(\hat{\beta}_{k+1})$, where all previous information has been naively discarded. The online updating procedure (2) then continues with $(\mathbf{X}, \mathbf{Z})$ in place of $\mathbf{X}$, and $(\beta^\top, \theta^\top)^\top$ in place of $\mathbf{b}$.

2.3. Updating at the Changing Block

The bias of the online updating estimator $\tilde{\mathbf{b}}_k$ up to block k in estimating β is $\delta = \mathbf{b} - \beta$, and can be estimated using data in block $k+1$ as $\hat{\delta}_{k+1} = \hat{\mathbf{b}}_{k+1} - \hat{\beta}_{k+1}$. This unbiased estimator of δ from block $k+1$ can be used to correct the bias of the cumulative estimate $\tilde{\mathbf{b}}_k$ from the previous block, such that

$$\tilde{\beta}_k = \tilde{\mathbf{b}}_k - \hat{\delta}_{k+1} \quad (4)$$

is an unbiased estimator of β . Using a weighted average as in (2), the updated estimate at block $k+1$ is

$$\tilde{\beta}_{k+1} = [\text{Var}^{-1}(\tilde{\beta}_k) + \text{Var}^{-1}(\hat{\beta}_{k+1})]^{-1} [\text{Var}^{-1}(\tilde{\beta}_k) \tilde{\beta}_k + \text{Var}^{-1}(\hat{\beta}_{k+1}) \hat{\beta}_{k+1}], \quad (5)$$

where $\text{Var}(\hat{\beta}_{k+1})$ can be estimated from block $k+1$, but $\text{Var}(\tilde{\beta}_k)$ needs extra care. As the blocks are mutually independent, we have

$$\text{Var}(\tilde{\boldsymbol{\beta}}_k) = \text{Var}(\tilde{\mathbf{b}}_k - \hat{\boldsymbol{\delta}}_{k+1}) = \text{Var}(\tilde{\mathbf{b}}_k) + \text{Var}(\hat{\boldsymbol{\delta}}_{k+1}). \quad (6)$$

Both terms on the right-hand side of (6) should be evaluated under H_F even though $\tilde{\mathbf{b}}_k$ and $\hat{\mathbf{b}}_{k+1}$ in $\hat{\boldsymbol{\delta}}_{k+1}$ are obtained under H_R .

The first component in Equation (6), $\text{Var}(\tilde{\mathbf{b}}_k)$, was derived in Clogg et al. (1995) and Allison (1995) for comparing regression coefficients between nested models. Clogg et al. (1995) condition on the realized values of $\mathbf{X}$ and $\mathbf{Z}$ in their derivations, while Allison (1995) regards them unconditionally as random as we do. Let $\mathbf{H}_{k+1} = (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})^{-1} \mathbf{X}_{k+1}^\top \mathbf{Z}_{k+1}$. Then by Allison (1995),

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\delta}}_{k+1}) &= \text{E}[\text{Var}(\hat{\boldsymbol{\delta}}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})] + \text{Var}[\text{E}(\hat{\boldsymbol{\delta}}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})] \\ &= \text{E}[\mathbf{H}_{k+1} \text{Var}(\hat{\boldsymbol{\theta}}_{k+1}) \mathbf{H}_{k+1}^\top] + \boldsymbol{\theta}^\top \Omega \boldsymbol{\theta} \text{E}[(\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})^{-1}], \end{aligned}$$

where Ω is a positive definite $q \times q$ variance matrix for the errors when regressing $\mathbf{Z}$ on $\mathbf{X}$. The variance of $\hat{\boldsymbol{\delta}}_{k+1}$ is estimated by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\delta}}_{k+1}) = \mathbf{H}_{k+1} \widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{k+1}) \mathbf{H}_{k+1}^\top + \hat{\boldsymbol{\theta}}_{k+1}^\top \mathbf{W}_{k+1} \hat{\boldsymbol{\theta}}_{k+1} (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})^{-1},$$

where $\mathbf{W}_{k+1}$ is a matrix of variances and covariances of the residuals obtained from regressing $\mathbf{Z}_{k+1}$ on $\mathbf{X}_{k+1}$.

Similarly, we can express $\text{Var}(\tilde{\mathbf{b}}_k)$ as

$$\text{Var}(\tilde{\mathbf{b}}_k) = \text{E}[\text{Var}(\tilde{\mathbf{b}}_k | \mathcal{X}_k, \mathcal{Z}_k)] + \text{Var}[\text{E}(\tilde{\mathbf{b}}_k | \mathcal{X}_k, \mathcal{Z}_k)],$$

pretending that the $\mathbf{Z}_\kappa$ are available for $\kappa = 1, \dots, k$. The conditional variance in the first component is shown to be $\text{Var}(\tilde{\mathbf{b}}_k | H_R) \sigma_\nu^2 / \sigma_\epsilon^2$, where $\text{Var}(\tilde{\mathbf{b}}_k | H_R)$ is the variance of $\tilde{\mathbf{b}}_k$ under the assumption of the reduced model H_R (Clogg et al., 1995, p.1274). For the second term,

$$\text{Var}[E(\tilde{\mathbf{b}}_k | \mathcal{X}_k, \mathcal{Z}_k)] = \text{Var}[\boldsymbol{\beta} + \mathcal{H}_k \boldsymbol{\theta}_k] = \text{Var}[\mathcal{H}_k \boldsymbol{\theta}_k] = \boldsymbol{\theta}^\top \Omega \boldsymbol{\theta} \text{E}[(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}]$$

, where $\mathcal{H}_k = (\mathcal{X}_k^\top \mathcal{X}_k)^{-1} \mathcal{X}_k^\top \mathcal{Z}_k$ is the cumulative version of $\mathbf{H}_k$ up to block k . Thus,

$$\text{Var}(\tilde{\mathbf{b}}_k) = \text{Var}(\tilde{\mathbf{b}}_k | H_R) \sigma_\nu^2 / \sigma_\epsilon^2 + \boldsymbol{\theta}^\top \Omega \boldsymbol{\theta} \text{E}[(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}].$$

Note that σ_ν^2 cannot be estimated until block $k+1$ because $\mathbf{Z}_\kappa$ is not observed for $\kappa = 1, \dots, k$. Consequently, we estimate the first term by $\widehat{\text{Var}}(\tilde{\mathbf{b}}_k | H_R) \hat{\sigma}_{\nu, n_{k+1}, k+1}^2 / \hat{\sigma}_{\epsilon, k+1}^2$, where

$\widehat{\text{Var}}(\tilde{\mathbf{b}}_k | H_R)$ is the cumulative variance estimate of $\tilde{\mathbf{b}}_k$ at block k under H_R , $\hat{\sigma}_{\nu, n_{k+1}, k+1}^2$ is the estimate of σ_ν^2 from block $k+1$, and $\hat{\sigma}_{\varepsilon, k+1}^2$ is the cumulative estimate of σ_ε^2 up to block $k+1$. Similarly, we estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_{k+1}$ and Ω by $\mathbf{W}_{k+1}$ from block $k+1$. Since $\mathbf{X}$ is available in all blocks, we would like to use as much information as possible to estimate $\text{E}[(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}]$; that is, we use $(\mathcal{X}_{k+1}^\top \mathcal{X}_{k+1})^{-1} = \mathbf{V}_{k+1}^{-1}$, scaled by the ratio in sample size N_{k+1}/N_k . Therefore, $\text{Var}(\tilde{\mathbf{b}}_k)$ can be estimated as:

$$\widehat{\text{Var}}(\tilde{\mathbf{b}}_k) = \widehat{\text{Var}}(\tilde{\mathbf{b}}_k | H_R) \frac{\hat{\sigma}_{\nu, n_{k+1}, k+1}^2}{\hat{\sigma}_{\varepsilon, k+1}^2} + \frac{N_{k+1}}{N_k} \hat{\boldsymbol{\theta}}_{k+1}^\top \mathbf{W}_{k+1} \hat{\boldsymbol{\theta}}_{k+1} \mathbf{V}_{k+1}^{-1}. \quad (7)$$

Remark 2.1—Under mild conditions, $N_k(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}$ converges almost surely to the inverse of the expectation for a single observation, $[\text{E}(x_1 x_1^\top)]^{-1}$, as $N_k \rightarrow \infty$. Using this relationship for both $N_k(\mathcal{X}_k^\top \mathcal{X}_k)^{-1}$ and $N_{k+1}(\mathcal{X}_{k+1}^\top \mathcal{X}_{k+1})^{-1}$ requires the sample size ratio adjustment N_{k+1}/N_k in (7). Alternative versions of (7) that do not use all available data on $\mathbf{X}$ are also possible (e.g., based on $\mathcal{X}_k^\top \mathcal{X}_k = \mathbf{V}_k$ or $\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1}$), but our simulation results confirm that using (7) based on $(\mathcal{X}_{k+1}^\top \mathcal{X}_{k+1})^{-1}$ is either better than or as good as the alternatives.

At this point, every term in Equation (5) that is needed for updating the estimates to block $k+1$ is available. Note that computation of $\text{Var}(\tilde{\boldsymbol{\beta}}_k)$ under the linear model is a special case of the computation under the GLM framework using influence functions discussed in Section 3.

2.4. Continue Updating with New Variables

In order to continue updating for future blocks, we need the variance-covariance matrix of the updated estimator $\tilde{\boldsymbol{\gamma}}_{k+1} = (\tilde{\boldsymbol{\beta}}_{k+1}, \tilde{\boldsymbol{\theta}}_{k+1})$ from the last subsection. That is,

$$\text{Var}(\tilde{\boldsymbol{\gamma}}_{k+1}) = \text{Var} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_{k+1} \\ \tilde{\boldsymbol{\theta}}_{k+1} \end{pmatrix} = \begin{pmatrix} \text{Var}(\tilde{\boldsymbol{\beta}}_{k+1}) & \text{Cov}(\tilde{\boldsymbol{\beta}}_{k+1}, \tilde{\boldsymbol{\theta}}_{k+1}) \\ \text{Cov}(\tilde{\boldsymbol{\beta}}_{k+1}, \tilde{\boldsymbol{\theta}}_{k+1})^\top & \text{Var}(\tilde{\boldsymbol{\theta}}_{k+1}) \end{pmatrix}.$$

Obviously, $\text{Var}(\tilde{\boldsymbol{\theta}}_{k+1})$ can be estimated from fitting the full model with data in block $k+1$. Detailed derivations for estimates of $\text{Var}(\tilde{\boldsymbol{\beta}}_{k+1})$ and $\text{Cov}(\tilde{\boldsymbol{\beta}}_{k+1}, \tilde{\boldsymbol{\theta}}_{k+1})$ are relegated to the Appendix. With these variance estimates available, the online updating algorithm in Equation (2) can be continued for future blocks with $(\mathbf{X}, \mathbf{Z})$ in place of $\mathbf{X}$ and $\boldsymbol{\gamma}$ in place of $\mathbf{b}$.

2.5. The Three-Variable Case

We consider a simple case with $p = q = 1$ and $k = 1$. That is, covariates x and z are both scalars, z becomes available after block 1, and we want to compare the naive estimator $\hat{\beta}_2$

and the bias-corrected online updating estimator $\tilde{\beta}_2$ at block 2. Assume that x and z are centered and that the correlation coefficient between x and z is ρ_{xz} . Let σ_η^2 be the error variance in the regression of z on x .

Define the relative efficiency of $\tilde{\beta}_2$ in comparison to $\hat{\beta}_2$ as the ratio of the Mean Squared Error (MSE) of the naive estimator $\hat{\beta}_2$ (throwing away block 1) to the MSE of the corrected estimator $\tilde{\beta}_2$, such that

$$\text{RE}(\tilde{\beta}_2, \hat{\beta}_2) = \frac{\text{MSE}(\hat{\beta}_2)}{\text{MSE}(\tilde{\beta}_2)}. \quad (8)$$

Since both $\hat{\beta}_2$ and $\tilde{\beta}_2$ are unbiased estimators of β , the ratio of their MSEs will be equal to the ratio of their variances, that is

$$\text{RE}(\tilde{\beta}_2, \hat{\beta}_2) = \frac{\text{MSE}(\hat{\beta}_2)}{\text{MSE}(\tilde{\beta}_2)} = \frac{\text{Var}(\hat{\beta}_2)}{\text{Var}(\tilde{\beta}_2)} = \frac{\left(\frac{n_1}{n_2} + \Delta\right)^2}{\frac{n_1}{n_2}\Delta + \Delta^2 + 2\frac{n_1}{n_2}\Delta\rho_{xz}^2}, \quad (9)$$

where $\Delta = (1 - \rho_{xz}^2) + \theta^2(\sigma_\eta^2/\sigma_\nu^2)(1 - \rho_{xz}^2)(n_1/n_2) + \theta^2(\sigma_\eta^2/\sigma_\nu^2)(1 - \rho_{xz}^2) + \rho_{xz}^2(n_1/n_2)$. See Appendix for detailed derivations.

It is interesting to see how the relative efficiency of $\tilde{\beta}_2$ with respect to $\hat{\beta}_2$ changes with other quantities. When both x and z are standardized, $\sigma_\eta^2 = 1 - \rho_{xz}^2$. If σ_ν^2 is additionally fixed, say at 1, then $\sigma_\eta^2/\sigma_\nu^2 = 1 - \rho_{xz}^2$. Note that $1 - \rho_{xz}^2$ is the inverse of the variance inflation factor (VIF).

Then, the relative efficiency becomes a function of three quantities: ρ_{xz} , θ , and n_1/n_2 . Figure 1 shows the relative efficiency when $n_1/n_2 = 1$: the left panel shows the relationship as a function of ρ_{xz} for fixed $\theta \in \{1, 2, 5\}$; the middle panel shows the relationship as a function of θ for fixed $\rho_{xz} \in \{\sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$. The relative efficiency of the corrected estimator decreases as either ρ_{xz} or θ increases. The relative efficiency is greater than one in most situations, but can be less than one when either ρ_{xz} or θ is sufficiently large. The magnitude of θ can be considered a measure of the importance of z relative to x , as β remains fixed. Alternatively, the importance of z can be measured as the partial R^2 of z given x in the model, given by $R_{y,z|x}^2 = (\text{SSE}(x) - \text{SSE}(x, z))/\text{SSE}(x)$, where $\text{SSE}(x)$ is the SSE from the reduced model and $\text{SSE}(x, z)$ is the SSE from the full model. Under the linear model setting, the partial R^2 has a closed form expression. The right panel in Figure 1 shows the relative efficiency as a function of the partial R^2 of z given x . As z contributes more to the model, the relative efficiency decreases. Figure 2 shows the relative efficiency as a function of n_1/n_2 for fixed $\rho_{xz} \in \{\sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$ with $\theta = 1$. As expected, the relative efficiency increases as n_1/n_2 increases.

The closed form relationships in this setting allow us to summarize the conditions for relative efficiency gain in the following proposition.

Proposition 2.2—Let $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2)$ be the relative efficiency of $\tilde{\beta}_2$ over $\hat{\beta}_2$ in the three-variable linear regression case, given in Equation (9).

1. If $\rho_{xz}^2 \leq 0.5$, then $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2) > 1$.
2. If $0.5 < \rho_{xz}^2 < 1$, then $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2) > 1$ if

$$\theta^2 < \frac{n_2(2\rho_{xz}^2 - 1) - n_1(2\rho_{xz}^2 + 1)\sigma_v^2}{(n_1 + n_2)(1 - 2\rho_{xz}^2)} \frac{\sigma_v^2}{\sigma_\eta^2} = \frac{n_1}{n_1 + n_2} \frac{2\rho_{xz}^2 + 1}{2\rho_{xz}^2 - 1} \frac{\sigma_v^2}{\sigma_\eta^2} - \frac{n_2}{n_1 + n_2} \frac{\sigma_v^2}{\sigma_\eta^2}.$$

The proof is provided in the Appendix. The inequality in condition 2 of Proposition 2.2 simplifies further if both covariates are standardized and $\sigma_v^2 = 1$; if $0.5 < \rho_{xz}^2 < 1$, we will see an efficiency gain in $\tilde{\beta}_2$ over $\hat{\beta}_2$ when

$$\theta^2 < \frac{n_1(2\rho_{xz}^2 + 1) - n_2(2\rho_{xz}^2 - 1)}{(n_1 + n_2)(2\rho_{xz}^2 - 1)(1 - \rho_{xz}^2)}.$$

Note that the right-hand side can be negative when ρ_{xz}^2 is close to 1 and when n_2 is much larger than n_1 . The closed form of $R_{y,z|x}^2$ under this simplified setting can be written as $R_{y,z|x}^2 = 1 - [\theta^2(1 - \rho_{xz}^2) + 1]^{-1}$. Clearly, the magnitude of θ has a positive relationship with $R_{y,z|x}^2$. This is expected, as a larger magnitude of θ indicates a stronger relationship between z and y . Thus when x and z are highly correlated, there will not be an efficiency gain in $\tilde{\beta}_2$ compared to $\hat{\beta}_2$ if the effect of z (measured by $R_{y,z|x}^2$ or magnitude of θ) is too large; see also Figure 1.

3. GENERALIZED LINEAR MODEL

3.1. Online Updating for Estimating Equations

In the GLM setting with a known link function g , the reduced model H_R is

$$g(\mu_{\kappa i, R}) = \eta_{\kappa i, R} = \mathbf{x}_{\kappa i}^\top \mathbf{b} \quad (10)$$

at each block κ , where $\mu_{\kappa i, R} = \mathbb{E}[y_{\kappa i} | \mathbf{x}_{\kappa i}]$. Schifano et al. (2016) proposed an online updating cumulative EE (CEE) estimator for $\mathbf{b}$. At block κ , let $\mathbf{w}_{\kappa i} = (y_{\kappa i}, \mathbf{x}_{\kappa i})$, $i = 1, \dots, n_\kappa$. Let $\hat{\mathbf{b}}_\kappa$ denote the solution to EE

$$\mathbf{M}_{n_{\kappa}, \kappa}(\mathbf{b}) = \sum_{i=1}^{n_{\kappa}} \psi(\mathbf{w}_{\kappa i}, \mathbf{b}) = \mathbf{0},$$

and $\widehat{\mathbf{V}}_{n_{\kappa}, \kappa}$ its corresponding variance estimate at block κ . The CEE estimator $\tilde{\mathbf{b}}_{\kappa}$ takes the form

$$\tilde{\mathbf{b}}_{\kappa} = (\mathbf{A}_{\kappa-1} + \mathbf{A}_{n_{\kappa}, \kappa})^{-1} (\mathbf{A}_{\kappa-1} \tilde{\mathbf{b}}_{\kappa-1} + \mathbf{A}_{n_{\kappa}, \kappa} \widehat{\mathbf{b}}_{\kappa}), \quad (11)$$

where

$$\mathbf{A}_{n_{\kappa}, \kappa} = - \sum_{i=1}^{n_{\kappa}} \frac{\partial \psi(\mathbf{w}_{\kappa i}, \widehat{\mathbf{b}}_{\kappa})}{\partial \mathbf{b}},$$

$\tilde{\mathbf{b}}_0 = \mathbf{0}$, $\mathbf{A}_0 = \mathbf{0}_p$ and $\mathbf{A}_{\kappa} = \sum_{\ell=1}^{\kappa} \mathbf{A}_{n_{\ell}, \ell} = \mathbf{A}_{\kappa-1} + \mathbf{A}_{n_{\kappa}, \kappa}$. The cumulative variance estimator for $\tilde{\mathbf{b}}_{\kappa}$ is

$$\tilde{\mathbf{V}}_{\kappa} = (\mathbf{A}_{\kappa-1} + \mathbf{A}_{n_{\kappa}, \kappa})^{-1} (\mathbf{A}_{\kappa-1} \tilde{\mathbf{V}}_{\kappa-1} \mathbf{A}_{\kappa-1}^{\top} + \mathbf{A}_{n_{\kappa}, \kappa} \widehat{\mathbf{V}}_{n_{\kappa}, \kappa} \mathbf{A}_{n_{\kappa}, \kappa}^{\top}) [(\mathbf{A}_{\kappa-1} + \mathbf{A}_{n_{\kappa}, \kappa})^{-1}]^{\top}, \quad (12)$$

with $\tilde{\mathbf{V}}_0 = \mathbf{0}_p$ and $\mathbf{A}_0 = \mathbf{0}_p$. Specifically, if $\mathbf{M}_{n_{\kappa}, \kappa}(\mathbf{b})$ is the score function under likelihood inference, then $\mathbf{A}_{n_{\kappa}, \kappa} = \widehat{\mathbf{V}}_{n_{\kappa}, \kappa}^{-1}$. Thus, from Equation (12), we have $\mathbf{A}_{\kappa} = \tilde{\mathbf{V}}_{\kappa}^{-1}$, and Equation (11) becomes

$$\tilde{\mathbf{b}}_{\kappa} = (\tilde{\mathbf{V}}_{\kappa-1}^{-1} + \widehat{\mathbf{V}}_{n_{\kappa}, \kappa}^{-1})^{-1} (\tilde{\mathbf{V}}_{\kappa-1}^{-1} \tilde{\mathbf{b}}_{\kappa-1} + \widehat{\mathbf{V}}_{n_{\kappa}, \kappa}^{-1} \widehat{\mathbf{b}}_{\kappa}). \quad (13)$$

Equation (5) for the LM setting is a special case of Equation (13).

3.2. Challenges from New Covariates

When a new set of predictor variables becomes available after block k , we assume that the correct, full model H_F includes these variables:

$$g(\mu_{\kappa i, F}) = \eta_{\kappa i, F} = \mathbf{x}_{\kappa i}^{\top} \boldsymbol{\beta} + \mathbf{z}_{\kappa i}^{\top} \boldsymbol{\theta}, \quad i = 1, \dots, n_{\kappa}, \quad (14)$$

where $\mu_{\kappa i, F} = \mathbb{E}(y_{\kappa i} | \mathbf{x}_{\kappa i}, \mathbf{z}_{\kappa i})$. As in the LM case, we still assume that design matrix $(\mathbf{X}_{\kappa}, \mathbf{Z}_{\kappa})$ is of full column rank $p + q$ and denote by N_{κ} the cumulative sample size, i.e.,

$N_k = \sum_{\ell=1}^K n_{\ell}$. This raises a question: if model (14) is the true model, what has $\tilde{\mathbf{b}}_k$ been estimating up to block k ?

As $N_k \rightarrow \infty$, $\tilde{\mathbf{b}}_k$ converges in probability to a limit $\mathbf{b}$, the solution to the EE corresponding to all the data (Lin and Xi, 2011). When the EEs are the score equations, $\mathbf{b}$ minimizes the Kullback–Leiber divergence (Kullback and Leibler, 1951) between model (10) and model (14). In the LM setting with identity link function g , $\mathbf{b}$ has a closed-form expression in terms of $\boldsymbol{\beta}$ plus some bias. In the GLM setting with a general link function g , we need to assume that $\mathbf{b}$ is the unique root of $\mathbf{M}_{n_k,k}(\mathbf{b})$ as $n_k \rightarrow \infty$. It is $\mathbf{b}$ that $\tilde{\mathbf{b}}_k$ has been estimating. Our strategy is the same as in the LM setting. For $\boldsymbol{\theta}$, we start the updating with $\tilde{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_{k+1}$; for $\boldsymbol{\beta}$, we construct $\tilde{\boldsymbol{\beta}}_{k+1}$ by correcting the bias in the cumulative estimate $\tilde{\mathbf{b}}_k$ in estimating $\boldsymbol{\beta}$. The bias of $\hat{\mathbf{b}}_k$ in estimating $\boldsymbol{\beta}$ is $\boldsymbol{\delta} = \mathbf{b} - \boldsymbol{\beta}$, and can be estimated using the data in block $k + 1$ as $\hat{\boldsymbol{\delta}}_{k+1} = \hat{\mathbf{b}}_{k+1} - \hat{\boldsymbol{\beta}}_{k+1}$.

3.3. Updating at the Changing Block

The bias correction formula (4) is unchanged, but the variance terms in the updating formulas (5)–(6) need to be replaced with the asymptotic variances under the GLM. We extend the influence functions approach for the LM used by Yan et al. (2013) to the GLM setting for these terms. The method could be further extended to handle clustered data with generalized estimating equations (GEE), and an implementation is available in R package *geepack* (Halekoh et al., 2006). Here we focus, however, on independent data. The quantities to estimate are $\text{var}(\tilde{\mathbf{b}}_k)$ and $\text{var}(\hat{\boldsymbol{\delta}}_{k+1})$, both under the full model H_F .

Assuming both $\mathbf{X}$ and $\mathbf{Z}$ are available at block κ , the EE for $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ under H_F is

$$\sum_{i=1}^{n_\kappa} \mathbf{U}_{\kappa i, F}(\boldsymbol{\gamma}) = \sum_{i=1}^{n_\kappa} \mathbf{D}_{\kappa i, F}^\top V_{\kappa i, F}^{-1} \left(y_{\kappa i} - \mu_{\kappa i, F} \right) = \mathbf{0},$$

where $\mathbf{D}_{\kappa i, F}^\top = \partial \mu_{\kappa i, F} / \partial \boldsymbol{\gamma}$ is a $(p+q) \times 1$ vector, $V_{(k+1)i, F}$ is the variance as a function of $\mu_{\kappa i, F}$. Let $\boldsymbol{\gamma}_0$ be the true value of $\boldsymbol{\gamma}$. The estimator $\hat{\boldsymbol{\gamma}}_\kappa$ has an asymptotic representation

$$\sqrt{n_\kappa}(\hat{\boldsymbol{\gamma}}_\kappa - \boldsymbol{\gamma}_0) = \frac{1}{\sqrt{n_\kappa}} \sum_{i=1}^{n_\kappa} i_{\kappa i, F}(\boldsymbol{\gamma}_0) + o_p(1),$$

where $i_{\kappa i, F}(\boldsymbol{\gamma}_0) = \Gamma_F(\boldsymbol{\gamma}_0) \mathbf{U}_{\kappa i, F}(\boldsymbol{\gamma}_0)$ and $\Gamma_F(\boldsymbol{\gamma}_0) = [\lim_{n_\kappa \rightarrow \infty} n_\kappa^{-1} \sum_{i=1}^{n_\kappa} \mathbf{D}_{\kappa i, F}^\top V_{\kappa i, F}^{-1} \mathbf{D}_{\kappa i, F}]^{-1}$ in probability are evaluated at the true value $\boldsymbol{\gamma}_0$. For the inverse in the definition of Γ_F to exist, the variance-covariance matrix of $(\mathbf{x}_{\kappa, i}^\top, \mathbf{z}_{\kappa, i}^\top)^\top$ needs to be of full rank and the link function needs to be strictly monotone; these are common assumptions (Liang and Zeger, 1986). The $(p+q) \times 1$ vectors $i_{\kappa i, F}$ are the influence functions of $\hat{\boldsymbol{\gamma}}_\kappa$. Thus, $\sqrt{n_\kappa}(\hat{\boldsymbol{\gamma}}_\kappa - \boldsymbol{\gamma}_0)$ converges in

distribution to $N(\mathbf{0}, \Gamma_F(\boldsymbol{\gamma}_0)\Omega_F(\boldsymbol{\gamma}_0)\Gamma_F(\boldsymbol{\gamma}_0))$, where $\Omega_F(\boldsymbol{\gamma}_0) = \lim_{n_K \rightarrow \infty} n_K^{-1} \sum_{i=1}^{n_K} \mathbf{U}_{\kappa i, F} \mathbf{U}_{\kappa i, F}^\top$ is assumed to be positive definite.

From the sample, we can obtain estimates $\hat{\mathbf{D}}_{\kappa i, F}$, $\hat{V}_{\kappa i, F}$ and $\hat{\mathbf{U}}_{\kappa i, F}(\hat{\boldsymbol{\gamma}}_K)$. Thus, $\Gamma_F(\boldsymbol{\gamma}_0)$, $\Omega_F(\boldsymbol{\gamma}_0)$, and the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}_K$ can be estimated by

$$\begin{aligned} \hat{\Gamma}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K) &= [n_K^{-1} \sum_{i=1}^{n_K} \hat{\mathbf{D}}_{\kappa i, F}^\top \hat{V}_{\kappa i, F}^{-1} \hat{\mathbf{D}}_{\kappa i, F}]^{-1}, \hat{i}_{\kappa i, F}(\hat{\boldsymbol{\gamma}}_K) = \hat{\Gamma}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K) \hat{\mathbf{U}}_{\kappa i, F}(\hat{\boldsymbol{\gamma}}_K), \quad \hat{\Omega}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K), \text{ and} \\ &= n_K^{-1} \sum_{i=1}^{n_K} \hat{\mathbf{U}}_{\kappa i, F} \hat{\mathbf{U}}_{\kappa i, F}^\top \\ \widehat{\text{Var}}(\hat{\boldsymbol{\gamma}}_K) &= n_K^{-1} \hat{\Gamma}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K) \hat{\Omega}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K) \hat{\Gamma}_{\kappa, F}(\hat{\boldsymbol{\gamma}}_K), \text{ respectively.} \end{aligned}$$

Similarly, under the reduced model H_R (10), the EE is

$$\sum_{i=1}^{n_K} \mathbf{U}_{\kappa i, R} = \sum_{i=1}^{n_K} \mathbf{D}_{\kappa i, R}^\top V_{\kappa i, R}^{-1} \left(y_{\kappa i} - \mu_{\kappa i, R} \right) = \mathbf{0},$$

where $\mathbf{D}_{\kappa i, R}^\top = \partial \mu_{\kappa i, R} / \partial \mathbf{b}$ is a $p \times 1$ vector, $V_{\kappa i, R}$ is the variance as a function of $\mu_{\kappa i, R}$, and subscript R is used to denote explicitly the corresponding quantities under H_R . The resulting estimator belongs to a wide class of estimators, the M-estimators; see van de Geer (2000) for a unified treatment. Under the assumption that $\mathbf{b}$ is the solution to this EE as $n_K \rightarrow \infty$ with true value $\mathbf{b}_0$, the asymptotic representation of the resulting estimator $\hat{\mathbf{b}}_K$ is

$$\sqrt{n_K}(\hat{\mathbf{b}}_K - \mathbf{b}_0) = \frac{1}{\sqrt{n_K}} \sum_{i=1}^{n_K} i_{\kappa i, R}(\mathbf{b}_0) + o_p(1),$$

where $i_{\kappa i, R}(\mathbf{b}_0) = \Gamma_R(\mathbf{b}_0) \mathbf{U}_{\kappa i, R}(\mathbf{b}_0)$ and $\Gamma_R(\mathbf{b}_0) = \left[\lim_{n_K \rightarrow \infty} n_K^{-1} \sum_{i=1}^{n_K} \mathbf{D}_{\kappa i, R}^\top V_{\kappa i, R}^{-1} \mathbf{D}_{\kappa i, R} \right]^{-1}$ in probability are evaluated at $\mathbf{b}_0$. The variance-covariance of $\mathbf{x}_{\kappa, i}$ needs to be of full rank for the inverse in the definition of Γ_R to exist. Here, $\sqrt{n_K}(\hat{\mathbf{b}}_K - \mathbf{b}_0)$ converges in distribution to $N(\mathbf{0}, \Gamma_R(\mathbf{b}_0)\Omega_R(\mathbf{b}_0)\Gamma_R(\mathbf{b}_0))$, where $\Omega_R(\mathbf{b}_0) = \lim_{n_K \rightarrow \infty} n_K^{-1} \sum_{i=1}^{n_K} \mathbf{U}_{\kappa i, R} \mathbf{U}_{\kappa i, R}^\top$ is assumed to be positive definite.

We can again obtain estimates $\hat{\mathbf{D}}_{\kappa i, R}$, $\hat{V}_{\kappa i, R}$ and $\hat{\mathbf{U}}_{\kappa i, R}(\hat{\mathbf{b}}_K)$ from the sample. $\Gamma_R(\mathbf{b}_0)$, $\Omega_R(\mathbf{b}_0)$, and the variance-covariance matrix can be estimated by

$$\begin{aligned} \hat{\Gamma}_{\kappa, R}(\hat{\mathbf{b}}_K) &= [n_K^{-1} \sum_{i=1}^{n_K} \hat{\mathbf{D}}_{\kappa i, R}^\top \hat{V}_{\kappa i, R}^{-1} \hat{\mathbf{D}}_{\kappa i, R}]^{-1}, \hat{i}_{\kappa i, R}(\hat{\mathbf{b}}_K) = \hat{\Gamma}_{\kappa, R}(\hat{\mathbf{b}}_K) \hat{\mathbf{U}}_{\kappa i, R}(\hat{\mathbf{b}}_K), \quad \hat{\Omega}_{\kappa, R}(\hat{\mathbf{b}}_K), \text{ and} \\ &= n_K^{-1} \sum_{i=1}^{n_K} \hat{\mathbf{U}}_{(\kappa)i, R} \hat{\mathbf{U}}_{(\kappa)i, R}^\top \\ \widehat{\text{Var}}(\hat{\mathbf{b}}_K) &= n_K^{-1} \hat{\Gamma}_{\kappa, R}(\hat{\mathbf{b}}_K) \hat{\Omega}_{\kappa, R}(\hat{\mathbf{b}}_K) \hat{\Gamma}_{\kappa, R}(\hat{\mathbf{b}}_K), \text{ respectively.} \end{aligned}$$

The influence functions in the asymptotic representation make the variance estimation of the needed quantities very easy. Notice that the influence functions of $\hat{\mathbf{b}}_K$, and, hence, its

variance and variance estimator, are the same as those obtained under H_R because of the definition of $\mathbf{b}$. Therefore, the estimator of $\text{Var}(\tilde{\mathbf{b}}_k)$ is the online updated $\widehat{\text{Var}}(\tilde{\mathbf{b}}_k) = \tilde{\mathbf{V}}_k$ at block k under H_R and does not require any further adjustments. The variance of $\hat{\delta}_{k+1}$ is estimated by

$$\widehat{\text{Var}}(\hat{\delta}_{k+1}) = \frac{1}{n_{k+1}^2} \sum_{i=1}^{n_{k+1}} \hat{j}_{(k+1)i, \delta}(\hat{\gamma}_{k+1}, \hat{\mathbf{b}}_{k+1}) \hat{j}_{(k+1)i, \delta}^T(\hat{\gamma}_{k+1}, \hat{\mathbf{b}}_{k+1}),$$

where $\hat{j}_{(k+1)i, \delta}(\hat{\gamma}_{k+1}, \hat{\mathbf{b}}_{k+1}) = \hat{i}_{(k+1)i, R}(\hat{\mathbf{b}}_{k+1}) - \hat{i}_{(k+1)i, F, \beta}(\hat{\gamma}_{k+1})$ and $\hat{i}_{(k+1)i, F, \beta}(\hat{\gamma}_{k+1})$ is the subvector of $\hat{i}_{(k+1)i, F}(\hat{\gamma}_{k+1})$ containing its first p elements.

3.4. Continue Updating with New Variables

To continue updating with new variables, we adapt the derivations for the LM case in the Appendix to find two key quantities, $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1})$ and $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1})$. Again, they can be readily estimated with the influence functions. First, we estimate $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1})$ by

$$\widehat{\text{Cov}}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1}) = \frac{1}{n_{k+1}^2} \sum_{i=1}^{n_{k+1}} \hat{i}_{(k+1)i, R}(\hat{\mathbf{b}}_{k+1}) \hat{i}_{(k+1)i, F, \beta}^T(\hat{\gamma}_{k+1}).$$

Further define $\hat{i}_{(k+1)i, F, \theta}(\hat{\gamma}_{k+1})$ as the subvector of $\hat{i}_{(k+1)i, F}(\hat{\gamma}_{k+1})$ containing its last q elements. Then, we estimate $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1})$ by

$$\widehat{\text{Cov}}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1}) = \frac{1}{n_{k+1}^2} \sum_{i=1}^{n_{k+1}} \hat{i}_{(k+1)i, R}(\hat{\mathbf{b}}_{k+1}) \hat{i}_{(k+1)i, F, \theta}^T(\hat{\gamma}_{k+1}).$$

All the other derivations are the same as in the LM setting.

4. SIMULATION STUDY

4.1. Three variable, two blocks

We first consider the case where there are three variables (y, x, z) and two blocks of data, as in Section 2.5. Covariates x and z were generated from a bivariate normal distribution with standard normal margins and correlation $\rho_{xz} \in \{\sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$. The first data generating model is a LM with $y = \beta_0 + \beta_1 x + \theta z + \varepsilon$, where $\beta_0 = \beta_1 = \theta = 1$ and ε follows a standard normal distribution. The second data generating model is a logistic model with y generated from Bernoulli distributions with probability $\text{logit}^{-1}(\beta_0 + \beta_1 x + \theta z)$. We use the same parameter values as in the LM.

The sample sizes of the two blocks were set at $n_1 = 10,000$ and $n_2 = 1,000$, respectively. For each scenario, 5,000 replicates were generated. First we verify that both the naive $\hat{\beta}_2$ and updated $\tilde{\beta}_2$ estimators are both unbiased under the LM and GLM settings. The estimated biases of these two estimators are reported in Table 1. The estimated biases are all very

small, but it is still observed that the estimated bias of $\tilde{\beta}_2$ is generally smaller than that of $\hat{\beta}_2$, because the updated estimator $\tilde{\beta}_2$ incorporates a much larger sample size in the first block. Next, we verify that the variance estimates of these two estimators are consistent with their empirical variance matrix. Table 1 provides the formula-based estimates of the standard errors and correlations for the regression coefficient estimators (averaged over simulations; denoted by ‘Average’) as well as the empirical standard errors and correlations of the regression coefficient estimators (denoted by ‘Empirical’). A close agreement between the average and empirical standard error is observed for the linear regression. For the logistic regression model, the agreement is not as tight as in the linear regression setting, but still generally good.

In Section 2.5, we have proved that RE is only affected by the absolute values of θ and ρ_{xz} under LM. This is also true for the logistic model and the symmetric pattern is demonstrated in Figure 3, and thus we only present results for positive θ and ρ_{xz} in Table 1. Also, as shown in Table 1 the estimated bias is very close to 0, so the estimated relative efficiencies will be very close to the corresponding ratio of variances. For the remainder of this paper, we will thus use the ratio of variances to represent the relative efficiency.

Next we study the efficiency gain under the GLM settings of the bias-corrected cumulative estimator relative to the naive estimator obtained from discarding previous data under different conditions. Unlike in the LM setting, there is no closed-form expression about the relative efficiency $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2)$. First, we set $n_1 = n_2 = 1,000$ and plot $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2)$, which was calculated using the empirical variances of the two estimators from 500 replicates, against ρ_{xz} and θ in Figure 3 (the analog of Figure 1). The left panel shows the change in the relative efficiency as ρ_{xz} increases from -0.9 to 0.9 with fixed $\theta \in \{-5, -2, -1, 1, 2, 5\}$. The curves for the negative and positive values of θ are symmetric about $\rho_{xz} = 0$. All three curves are above 1 in the range shown. The pattern seems to be in agreement with the change of $\text{Corr}(\tilde{\beta}_2, \hat{\beta}_2)$ in Table 1. As $\text{Corr}(\tilde{\beta}_2, \hat{\beta}_2)$ becomes closer to zero, relative efficiency increases. The right panel of Figure 3 shows the change in relative efficiency as θ varies from -5 to 5 for fixed $\rho_{xz} \in \{-\sqrt{0.7}, -\sqrt{0.5}, -\sqrt{0.1}, \sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$. The six curves are symmetric about $\theta = 0$ and all have the same increasing-then-decreasing trend. The peak of the curve is shifting to the left as ρ_{xz} decreases, and shifting to the right as ρ_{xz} increases. Again, all the curves are above 1 in the range shown.

Figure 4 is the analog of Figure 2 under the logistic model using empirical relative efficiencies. With $\theta = 1$ and $n_2 = 1,000$, the empirical relative efficiency $\text{RE}(\tilde{\beta}_2, \hat{\beta}_2)$ obtained from 100 replicates is shown as n_1/n_2 increases from 1 to 20 for fixed $\rho_{xz} \in \{\sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$. All three curves are above 1, increasing as n_1/n_2 increases. The sample size appears to have a larger effect on $\rho_{xz} = \sqrt{0.1}$. The magnitude of the relative efficiency here seems much greater than that in Figure 2 in the LM setting, suggesting that using information from the previous data in the logistic model can be even more beneficial than in the LM setting.

4.2. Beyond three variables and two blocks

We now increase the sample sizes, number of variables and number of blocks to more closely mimic the real data analysis in Section 5. Covariates $x_1, x_2, x_3, x_4, x_5, z_1, z_2, z_3$ and

z_4 were generated from a multivariate normal distribution with standard normal margins and correlation $\rho \in \{\sqrt{0.1}, \sqrt{0.5}, \sqrt{0.7}\}$ between any two different covariates. The first data generating model is a LM with $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4 + \varepsilon$, where $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, $\beta_3 = 0.5$, $\beta_4 = -0.5$, $\beta_5 = -1$, $\theta_1 = 0.5$, $\theta_2 = 1$, $\theta_3 = -0.5$, $\theta_4 = -1$, and ε follows a standard normal distribution. The second data generating model is a logistic model with y generated from Bernoulli distributions with probability $\text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \theta_1 z_1 + \theta_2 z_2 + \theta_3 z_3 + \theta_4 z_4)$. We use the same parameter values as in the LM setting. At block 1, only x_1 , x_2 , x_3 , x_4 , and x_5 are available while z_1 and z_2 become available at block 2, and all variables are available at block 3. Each block has sample size of 50,000. In this case, the efficiency gains for β_1 to β_5 can be computed at block 2 when the model with x 's only is corrected to the model with all x 's, z_1 , and z_2 . Similarly, the efficiency gains for all five β 's, θ_1 , and θ_2 can be computed at block 3 when the model is further corrected to the full model.

Results for the LM simulation are summarized in Table 2. It can be observed that the magnitude and sign of the true coefficients for covariates in the “old” model do not make a difference in their relative efficiencies. We have proved this for the three-variable case in Section 2.5. It can also be observed that the efficiency gains for the β estimates are all larger than those for θ_1 and θ_2 at block 3. This is because more historical information is available for the x covariates (from both blocks 1 and 2) than for the z covariates (from block 2 only). It is interesting to see that the efficiency gains for the β estimates at block 3 are not necessarily larger than those at block 2. When more values of z are added in the model, the relative importance of x may decrease. This is consistent with the finding from Section 2.5 where larger values of θ led to smaller RES. In this simulation, the RES are affected by correlations, importance of new variables, and sample sizes.

Similar results for GLM are summarized in Table 3. This is a more complicated situation than LM where the magnitudes and signs of the true coefficients have an impact on the RE's. As with the LM setting, the efficiency gains for the β estimates are all larger than those for θ_1 and θ_2 at block 3.

Finally, the computing time is always crucial in the big data setting. Since our methodologies are based on the online updating algorithm, it is assumed that the data arrive (or are subset) into blocks of manageable size. The additional computation time required for the proposed bias-correction method over the naive approach just amounts to some simple algebra on small dimensional matrices. When the number of new variables is not too big, the computing time for our method should be more or less double that of the naive method only at the changing block, since we need to fit both the full and reduced models at the changing block. This computational efficiency loss is acceptable since online updating method has already limited the computing time for each block to a small amount.

5. AIRLINE DELAYS

We examined the airline on-time statistics using the data from the 2009 ASA DataExpo (available at <http://stat-computing.org/dataexpo/2009/the-data.html>). The raw data consists of flight departure/arrival information for all commercial flights within the USA, from

October 1987 to April 2008, with $N = 123,534,969$ observations and 29 variables (~ 11 GB). In June 2003, the US Department of Transportation Bureau of Transportation Statistics began collecting data on the causes of flight delays. The categories of delays created by the Air Carrier On-Time Reporting Advisory Committee are:

- Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control.
- Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delay or prevent the operation of a flight such as tornado, blizzard or hurricane.
- National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.
- LateAircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.

These five variables are recorded in minutes and are not mutually exclusive to each other. They were not available until June 2003 and are apparently important in explaining flight delays. Two models were considered: a linear regression for the arrival delay in minutes and a logistic model for whether or not the arrival delay is longer than 15 minutes.

For the linear regression, we first consider the three-variable case where x is the distance in thousands of miles, and add each of the delay causes separately as z into the model with data of May–June, 2003. Table 4 shows the correlation coefficient of x and z , the t -ratio (estimate over standard error) for the newly added variable z , and the ratio of the variance estimate of $\tilde{\beta}_2$ to that of $\hat{\beta}_2$. The variables are sorted in the order of their t -ratios. Note that, because this is a single data analysis, we can only report the variance ratio estimate instead of the relative efficiency, where the variances are estimated from the formulas directly. This variance ratio is much smaller than 1, ranging from 0.49 to 0.67, suggesting that, by including the data in May 2003, the cumulative estimate for the coefficient of distance has a smaller standard error than the estimate from the June 2003 data alone. As the values of the t -ratios increase, the variance ratios also increase. This behavior is expected, as larger t -ratios indicate greater importance of the variable z relative to x , and, hence, including existing information on x becomes less effective.

Next we consider a multiple linear regression with x being the distance, and $\mathbf{z}$ being the five delay causes. We also let the dimension q of $\mathbf{z}$ vary from 1 to 5 by adding the different types of airline delay to the model one by one, in the same order as in Table 4. The variance ratios of the coefficient estimate for distance at all five steps are summarized in Table 5. As the number of new variables q increases from 1 to 5, the importance of the whole set of $\mathbf{z}$ relative to x increases. Consequently, the variance ratio increases and the relative advantage

of the adjusted cumulative estimate decreases. The variance ratio is also impacted by the significance of the added variable. For example, from Table 4, we see that Security delay is the least significant covariate in all the five causes of delay. In Table 4, the variance ratio when adding just Weather delay is 0.53, while in Table 5 the variance ratio when adding both Weather and Security is 0.54; adding Security delay has almost no impact on the ratio. This suggests that, in practice, when too many new variables or some extremely significant variables become available, it may not always be beneficial to use the previous information. Notably, however, these results are based on sample size ratio $n_1/n_2 \approx 1$.

The impact of the sample sizes of the previous data and current data can be investigated by including more data before June 2003. We kept the data of June 2003 as the new block of data, and expanded the previous data to include those from January–May 2003, which increases the ratio of sample sizes n_1/n_2 roughly from 1 to 5. The left panel of Figure 5 shows the variance ratio against n_1/n_2 for the three-variable case as each delay cause serves as z . As more months of data are included as existing data, the variance ratio drops below 0.5 for even the most important delay cause. The right panel of Figure 5 shows the variance ratio against n_1/n_2 in the multiple linear regression with different numbers of new variables in z . The heights of the lines increase as the number of new variables increases, but even for the model with 5 new variables, an increased sample size corresponding to previous information still has a beneficial effect on the variance ratio.

Similar analyses were performed for the logistic regression model for the probability of over 15 minutes arrival delay; that is, the response variable is binary, 1 if late by more than 15 minutes and 0 otherwise. Tables 6 and 7 are, respectively, the analogs of Tables 4 and 5. Figure 6 is the analog of Figure 5. Note that in Table 6, the order of importance of the delay causes is different from that in Table 4, suggesting that the dichotomized arrival delay does have quite different behavior than the pre-dichotomized continuous version. The conclusions about the impact of the added variables, however, remain the same as those in the LM setting.

6. DISCUSSION

We focused on the online updating method in analyzing big data when new covariates become available, in which case, the updating method of Schifano et al. (2016) would need to start from scratch. Our strategy was to correct the bias in the cumulative coefficient estimate of the existing variables by generalizing a method that makes inferences about the difference in regression coefficients between nested models (Yan et al., 2013). After appropriate handling of the weights needed in the online updating algorithm after the correction, the online updating algorithm may resume. We have shown that including the cumulative estimates from existing data helps improve efficiency under a variety of practical situations. The efficiency gain is large when the existing and new variables are not highly correlated, the new variables are not dominating the predictions, or the cumulative sample size is large relative to the sample size of the changing block. Note that the convergence rate of the cumulative estimator established in Schifano et al. (2016) no longer holds because the bias correction at block $k + 1$ is of order $1/\sqrt{n_k + 1}$. If another batch of new variables become

available at some block later, the order will be adjusted again by the sample size of that block.

No storage of the past data is a special feature of stream data, which makes missing data approaches such as multiple imputation infeasible. The opposite of the newly available variables scenario is that some variables may become unavailable. If this is only temporary, these blocks could be treated as “old” blocks, and their information could be used when a future block with full data becomes available using our approach, before being combined with the existing cumulative estimates; multiple imputation could be useful in this case as well if an imputation model has been updated cumulatively. If these variables become unavailable permanently, one would probably need to change the model.

The rank deficiency problem addressed in Schifano et al. (2016) may also arise in the specific setting of newly available variables. Our method assumes that the design matrices under the reduced and full models are both of full rank at the changing block; otherwise, the bias cannot be estimated using the data in this block. The rank deficiency problem in other blocks could be solved using generalized inverses as in Schifano et al. (2016). This problem is common when the new variable is binary. Intervention analysis often introduces a binary variable which takes value 0 before the changing block and 1 from the changing block forward. Another example is the 9–11 terrorist attack effect on airline delays. In this specific setting, one can augment the vector of coefficient estimates and the matrix of variance-covariance estimates by zeros, then replace all inverses with generalized inverses in the correction and updating formulas at the changing block.

While the proposed method was motivated by big data applications, it may also prove beneficial in other contexts, e.g., in meta-analysis with systematically missing covariates in some studies. This is an interesting area devoted for future research.

Acknowledgments

Dr. M.-H. Chen’s research was partially supported by the United States National Institutes of Health. Dr. Y. Zhang’s research was partially supported by the Connecticut Institute for Clinical and Translational Science, the University of Connecticut Research Excellence Program and Institute for Collaboration on Health, Intervention, and Policy.

BIBLIOGRAPHY

- Allison PD. The Impact of Random Predictors on Comparisons of Coefficients between Models: Comment on Clogg, Petkova, and Haritou. *American Journal of Sociology*. 1995; 100:1294–1305.
- Certo ST. Influencing Initial Public Offering Investors with Prestige: Signaling with Board Structures. *Academy of Management Review*. 2003; 28:432–446.
- Chen X, Xie M-g. A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data. *Statistica Sinica*. 2014; 24:1655–1684.
- Clogg CC, Petkova E, Haritou A. Statistical Methods for Comparing Regression Coefficients between Models. *American Journal of Sociology*. 1995; 100:1261–1293.
- Desyllas P, Sako M. Profiting from Business Model Innovation: Evidence from Pay-As-You-Drive Auto Insurance. *Research Policy*. 2013; 42:101–116.
- Greene, WH. *Econometric Analysis*. Pearson Education India; 2003.
- Halekoh U, Højsgaard S, Yan J. The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*. 2006; 15:1–11.

- Hood L, Heath JR, Phelps ME, Lin B. Systems Biology and New Technologies Enable Predictive and Preventative Medicine. *Science*. 2004; 306:640–643. [PubMed: 15499008]
- Kleiner A, Talwalkar A, Sarkar P, Jordan MI. A Scalable Bootstrap for Massive Data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2014; 76:795–816.
- Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951; 22:79–86.
- Liang F, Cheng Y, Song Q, Park J, Yang P. A Resampling-based Stochastic Approximation Method for Analysis of Large Geostatistical Data. *Journal of the American Statistical Association*. 2013; 108:325–339.
- Liang K-Y, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*. 1986; 73:13–22.
- Lin N, Xi R. Aggregated Estimating Equation Estimation. *Statistics and Its Interface*. 2011; 4:73–83.
- Liu D, Liu RY, Xie M. Multivariate Meta-Analysis of Heterogeneous Studies Using only Summary Statistics: Efficiency and Robustness. *Journal of the American Statistical Association*. 2015; 110:326–340. [PubMed: 26190875]
- Ma P, Mahoney MW, Yu B. A Statistical Perspective on Algorithmic Leveraging. arXiv preprint arXiv: 1306.5362. 2013
- Maclaurin D, Adams RP. Firefly Monte Carlo: Exact MCMC with Subsets of Data. arXiv preprint arXiv: 1403.5693. 2014
- Neiswanger W, Wang C, Xing E. Asymptotically Exact, Embarrassingly Parallel MCMC. arXiv preprint arXiv: 1311.4780. 2013
- Rupp, NG. Tech. rep. Department of Economy, East Carolina University; 2007. Further Investigations into the Causes of Flight Delays.
- Schifano ED, Wu J, Wang C, Yan J, Chen M-H. Online Updating of Statistical Inference in the Big Data Setting. *Technometrics*. 2016; 58:393–403. [PubMed: 28018007]
- Scott SL, Blocker AW, Bonassi FV, Chipman H, George E, McCulloch R. Bayes and Big Data: The Consensus Monte Carlo Algorithm. *EFaBBayes 250 conference*. 2013; 16
- Snijders C, Matzat U, Reips U-D. Big Data: Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science*. 2012; 7:1–5.
- Song Q, Liang F. A Split-and-merge Bayesian Variable Selection Approach for Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2014; 77:947–972.
- van de Geer, SA. Empirical Processes in M-estimation. Vol. 6. Cambridge university press; 2000.
- Wang C, Chen M-H, Schifano ED, Wu J, Yan J. Statistical Methods and Computing for Big Data. *Statistics and Its Interface*. 2016; 9:399–414. [PubMed: 27695593]
- Yan J, Aseltine RH, Harel O. Comparing Regression Coefficients Between Nested Linear Models for Clustered Data With Generalized Estimating Equations. *Journal of Educational and Behavioral Statistics*. 2013; 38:172–189.

APPENDIX

Derivations for the Linear Model

Continuing with Section 2.4, to derive $\text{Var}(\tilde{\beta}_{k+1})$, we can decompose (5) and let

$$W_1 = [\text{Var}^{-1}(\tilde{\beta}_k) + \text{Var}^{-1}(\hat{\beta}_{k+1})]^{-1} \text{Var}^{-1}(\tilde{\beta}_k),$$

$$W_2 = [\text{Var}^{-1}(\tilde{\beta}_k) + \text{Var}^{-1}(\hat{\beta}_{k+1})]^{-1} \text{Var}^{-1}(\hat{\beta}_{k+1}).$$

thus,

$$\begin{aligned} \text{Var}(\tilde{\beta}_{k+1}) &= \text{Var}(W_1 \tilde{\beta}_k + W_2 \hat{\beta}_{k+1}) = W_1 \text{Var}(\tilde{\beta}_k) W_1^\top + W_1 \text{Cov}(\tilde{\beta}_k, \hat{\beta}_{k+1}) W_2^\top \\ &+ W_2 \text{Cov}(\hat{\beta}_{k+1}, \tilde{\beta}_k) W_1^\top + W_2 \text{Var}(\hat{\beta}_{k+1}) W_2^\top \end{aligned} \quad (\text{A.1})$$

where the estimate of $\text{Var}(\tilde{\beta}_k)$ is found in Section 2.3 and $\text{Var}(\hat{\beta}_{k+1})$ is estimated at block $k + 1$ as $\widehat{\text{Var}}(\hat{\beta}_{k+1})$. By decomposing $\tilde{\beta}_k$, the covariance between cumulative and current estimators of β can be decomposed as

$$\text{Cov}(\tilde{\beta}_k, \hat{\beta}_{k+1}) = \text{Var}(\hat{\beta}_{k+1}) - \text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1}).$$

By the law of total covariance, we have $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1}) = \mathbb{E}[\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1}) | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1}] + \text{Cov}[\mathbb{E}(\hat{\mathbf{b}}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1}), \mathbb{E}(\hat{\beta}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})]$. According to Clogg et al. (1995), the first component is equal to $\text{Var}(\hat{\mathbf{b}}_{k+1} | H_R) \sigma_\nu^2 / \sigma_\varepsilon^2$, which can be estimated by

$\widehat{\text{Var}}(\hat{\mathbf{b}}_{k+1} | H_R) \hat{\sigma}_{\nu, n_{k+1}, k+1}^2 / \hat{\sigma}_{\varepsilon, k+1}^2$ from the sample while the second component is zero.

Alternatively, the first component can be estimated directly by using linear algebra which will be demonstrated later.

The other quantity to derive is $\text{Cov}(\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1})$. We can decompose $\tilde{\beta}_{k+1}$ and finally get

$$\text{Cov}(\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1}) = (W_1 + W_2) \text{Cov}(\hat{\beta}_{k+1}, \hat{\theta}_{k+1}) - W_1 \text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1}),$$

where $\text{Cov}(\hat{\beta}_{k+1}, \hat{\theta}_{k+1})$ is estimated as part of the variance-covariance matrix of coefficients at block $k + 1$. For $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1})$, we have $\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1}) = \mathbb{E}[\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1}) | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1}] + \text{Cov}[\mathbb{E}(\hat{\mathbf{b}}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1}), \mathbb{E}(\hat{\theta}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})]$. The second component is zero, but the first component takes more effort to derive. From the Linear Model,

$$\hat{\mathbf{b}}_{k+1} = (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})^{-1} \mathbf{X}_{k+1}^\top \mathbf{y}_{k+1},$$

and

$$\begin{pmatrix} \hat{\beta}_{k+1} \\ \hat{\theta}_{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} & \mathbf{X}_{k+1}^\top \mathbf{Z}_{k+1} \\ \mathbf{Z}_{k+1}^\top \mathbf{X}_{k+1} & \mathbf{Z}_{k+1}^\top \mathbf{Z}_{k+1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{k+1}^\top \\ \mathbf{Z}_{k+1}^\top \end{pmatrix} \mathbf{y}_{k+1}.$$

Let

$$\mathbf{B} = \begin{pmatrix} \mathbf{X}_{k+1}^\top \mathbf{X}_{k+1} & \mathbf{X}_{k+1}^\top \mathbf{Z}_{k+1} \\ \mathbf{Z}_{k+1}^\top \mathbf{X}_{k+1} & \mathbf{Z}_{k+1}^\top \mathbf{Z}_{k+1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}.$$

where $\mathbf{B}_{11}$ is $p \times p$, $\mathbf{B}_{12}$ is $p \times q$, $\mathbf{B}_{21}$ is $q \times p$, and $\mathbf{B}_{22}$ is $q \times q$. All of them have closed-forms. So,

$$\begin{pmatrix} \hat{\beta}_{k+1} \\ \hat{\theta}_{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{11}\mathbf{X}_{k+1}^\top + \mathbf{B}_{12}\mathbf{Z}_{k+1}^\top \\ \mathbf{B}_{21}\mathbf{X}_{k+1}^\top + \mathbf{B}_{22}\mathbf{Z}_{k+1}^\top \end{pmatrix} \mathbf{y}_{k+1}.$$

Let $\mathbf{C} = \mathbf{B}_{11}\mathbf{X}_{k+1}^\top + \mathbf{B}_{12}\mathbf{Z}_{k+1}^\top$, $\mathbf{D} = \mathbf{B}_{21}\mathbf{X}_{k+1}^\top + \mathbf{B}_{22}\mathbf{Z}_{k+1}^\top$, and $\mathbf{E} = (\mathbf{X}_{k+1}^\top \mathbf{X}_{k+1})^{-1} \mathbf{X}_{k+1}^\top$. The separate least square estimators of β and θ at block $k+1$ are $\hat{\beta}_{k+1} = \mathbf{C}\mathbf{y}_{k+1}$ and $\hat{\theta}_{k+1} = \mathbf{D}\mathbf{y}_{k+1}$, respectively. Thus, we have

$$\mathbb{E}[\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\beta}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})] = \mathbb{E}[\text{Cov}(\mathbf{E}\mathbf{y}_{k+1}, \mathbf{C}\mathbf{y}_{k+1})] = \mathbb{E}[\sigma_\nu^2 \mathbf{E}\mathbf{C}^\top],$$

$$\mathbb{E}[\text{Cov}(\hat{\mathbf{b}}_{k+1}, \hat{\theta}_{k+1} | \mathbf{X}_{k+1}, \mathbf{Z}_{k+1})] = \mathbb{E}[\text{Cov}(\mathbf{E}\mathbf{y}_{k+1}, \mathbf{D}\mathbf{y}_{k+1})] = \mathbb{E}[\sigma_\nu^2 \mathbf{E}\mathbf{D}^\top],$$

where $\mathbb{E}[\sigma_\nu^2 \mathbf{E}\mathbf{C}^\top]$ can be estimated by $\hat{\sigma}_{\nu, n_{k+1}, k+1}^2 \mathbf{E}\mathbf{C}^\top$ and $\mathbb{E}[\sigma_\nu^2 \mathbf{E}\mathbf{D}^\top]$ can be estimated by $\hat{\sigma}_{\nu, n_{k+1}, k+1}^2 \mathbf{E}\mathbf{D}^\top$ from the sample. Finally the variance-covariance matrix of $(\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1})$ can be estimated with the formulae above and $\hat{W}_1, \hat{W}_2$, the estimates of W_1 and W_2 . Updating with future blocks just follows the updating algorithms in Schifano et al. (2016).

Proof of Proposition 2.2

In this section, we present the proof for Proposition 2.2, the improvement of the new method under the linear regression setting in the three-variable-and-two-block case. For simplicity, both x and z are centered.

Define $S_{x_1}^2 = \sum_{i=1}^{n_1} x_{1i}^2/n_1$ and $S_{x_2}^2 = \sum_{i=1}^{n_2} x_{2i}^2/n_2$. We assume that variable x has standard deviation σ_x and the correlation between x and z is ρ_{xz} . Thus,

$$\text{Var}(\hat{b}_1 | H_R) = \mathbb{E}[\text{Var}(\hat{b}_1 | x_1)] + \text{Var}[\mathbb{E}(\hat{b}_1 | x_1)] = \mathbb{E}\left[\frac{\sigma_\varepsilon^2}{n_1 S_{x_1}^2}\right] + \text{Var}(b) = \frac{\sigma_\varepsilon^2}{n_1} \mathbb{E}\left[\frac{1}{S_{x_1}^2}\right].$$

Similarly we find $\text{Var}(\hat{b}_2 | H_R) = \mathbb{E}[1/S_{x_2}^2] \sigma_\varepsilon^2/n_2$, and $\text{Var}(\hat{\beta}_2) = \mathbb{E}[1/S_{x_2}^2] \sigma_\nu^2/(n_2(1 - \rho_{xz}^2))$.

To utilize the correction method by Allison (1995), we need to regress z on x . The variance of $\hat{b}_{zx}$, the estimated coefficient of x in a regression of z on x , is $\text{Var}(\hat{b}_{zx}) = \mathbb{E}[1/S_{x_2}^2] \sigma_\eta^2/n_2$.

Since x has the standard deviation σ_x , $E[1/S_{x_1}^2]$ and $E[1/S_{x_2}^2]$ are equal and can be determined from the distribution of x . We denote both by $E[1/S_{x_1}^2] = E[1/S_{x_2}^2] = 1/\Lambda_x$.

Following the derivations in Section 2 and earlier in the Appendix, we get the following simplified formulae in the three-variable-and-two-block case,

$$\text{Var}(\hat{b}_1 | H_F) = \frac{\sigma_\nu^2}{n_1 \Lambda_x} + \theta_2^2 \frac{\sigma_\eta^2}{n_1 \Lambda_x},$$

$$\text{Var}(\hat{\delta}_2) = \frac{\sigma_\nu^2}{n_2 \Lambda_x (1 - \rho_{xz}^2)} - \frac{\sigma_\nu^2}{n_2 \Lambda_x} + \theta_2^2 \frac{\sigma_\eta^2}{n_2 \Lambda_x},$$

$$\begin{aligned} \text{Var}(\tilde{\beta}_1) &= \frac{\sigma_\nu^2}{n_1 \Lambda_x} + \theta_2^2 \frac{\sigma_\eta^2}{n_2 \Lambda_x} + \theta_2^2 \frac{\sigma_\eta^2}{n_1 \Lambda_x} + \frac{\rho_{xz}^2 \sigma_\nu^2}{n_2 \Lambda_x (1 - \rho_{xz}^2)} \\ &= \frac{\sigma_\nu^2 [n_2 (1 - \rho_{xz}^2) + n_1 \theta_2^2 \frac{\sigma_\eta^2}{\sigma_\nu^2} (1 - \rho_{xz}^2) + n_2 \theta_2^2 \frac{\sigma_\eta^2}{\sigma_\nu^2} (1 - \rho_{xz}^2) + n_1 \rho_{xz}^2]}{n_1 n_2 \Lambda_x (1 - \rho_{xz}^2)}. \end{aligned}$$

According to Equation (A.1), we have

$$\text{Var}(\tilde{\beta}_2) = W_1^2 \text{Var}(\tilde{\beta}_1) + W_2^2 \text{Var}(\hat{\beta}_2) + 2W_1 W_2 \text{Cov}(\tilde{\beta}_1, \hat{\beta}_2),$$

where

$$W_1 = [\text{Var}^{-1}(\tilde{\beta}_1) + \text{Var}^{-1}(\hat{\beta}_2)]^{-1} \text{Var}^{-1}(\tilde{\beta}_1),$$

$$W_2 = [\text{Var}^{-1}(\tilde{\beta}_1) + \text{Var}^{-1}(\hat{\beta}_2)]^{-1} \text{Var}^{-1}(\hat{\beta}_2),$$

$$\text{Cov}(\tilde{\beta}_1, \hat{\beta}_2) = \text{Var}(\hat{\beta}_2) - \text{Var}(\hat{b}_2 | H_R) \sigma_\nu^2 / \sigma_\varepsilon^2 = \frac{\rho_{xz}^2 \sigma_\nu^2}{n_2 \Lambda_x (1 - \rho_{xz}^2)}.$$

Letting

$$\Delta = (1 - \rho_{xz}^2) + \frac{n_1}{n_2} \theta^2 \frac{\sigma_\eta^2}{\sigma_\nu^2} (1 - \rho_{xz}^2) + \theta^2 \frac{\sigma_\eta^2}{\sigma_\nu^2} (1 - \rho_{xz}^2) + \frac{n_1}{n_2} \rho_{xz}^2,$$

we get

$$\text{Var}(\tilde{\beta}_2) = \left(\frac{\frac{n_1}{n_2}}{\frac{n_1}{n_2} + \Delta} \right)^2 \frac{\sigma_\nu^2 \Delta}{n_1 \Lambda_x (1 - \rho_{xz}^2)} + \left(\frac{\Delta}{\frac{n_1}{n_2} + \Delta} \right)^2 \frac{\sigma_\nu^2}{n_2 \Lambda_x (1 - \rho_{xz}^2)} + \frac{\frac{n_1}{n_2} \Delta}{\left(\frac{n_1}{n_2} + \Delta \right)^2} \frac{\rho_{xz}^2 \sigma_\nu^2}{n_2 \Lambda_x (1 - \rho_{xz}^2)}.$$

Our goal is find the conditions for

$$\frac{\text{Var}(\hat{\beta}_2)}{\text{Var}(\tilde{\beta}_2)} = \frac{\left(\frac{n_1}{n_2} + \Delta \right)^2}{\frac{n_1}{n_2} \Delta + \Delta^2 + 2 \frac{n_1}{n_2} \Delta \rho_{xz}^2} \geq 1. \quad (\text{A.2})$$

Expanding and rearranging gives the following cases:

When $0 \leq \rho_{xz}^2 < 0.5$, we need

$$\theta^2 \geq \frac{n_2(2\rho_{xz}^2 - 1) - n_1(2\rho_{xz}^2 + 1)}{(n_1 + n_2)(1 - 2\rho_{xz}^2)} \frac{\sigma_\nu^2}{\sigma_\eta^2} = -\frac{n_2}{n_1 + n_2} \frac{\sigma_\nu^2}{\sigma_\eta^2} - \frac{n_1}{n_1 + n_2} \frac{2\rho_{xz}^2 + 1}{1 - 2\rho_{xz}^2} \frac{\sigma_\nu^2}{\sigma_\eta^2}.$$

Because

$$-\frac{n_2}{n_1 + n_2} \frac{\sigma_\nu^2}{\sigma_\eta^2} - \frac{n_1}{n_1 + n_2} \frac{2\rho_{xz}^2 + 1}{1 - 2\rho_{xz}^2} \frac{\sigma_\nu^2}{\sigma_\eta^2} < 0$$

and $\theta^2 \geq 0$, the inequality always holds.

When $\rho_{xz}^2 = 0.5$, we need

$$\left(\frac{n_1}{n_2} + \Delta \right)^2 \geq \frac{n_1}{n_2} \Delta + \Delta^2 + 2 \frac{n_1}{n_2} \Delta \rho_{xz}^2,$$

where $\rho_{xz}^2 = 0.5$. After simplification, that is $n_1/n_2 \geq 0$ which is always true.

When $0.5 < \rho_{xz}^2 < 1$,

$$\theta^2 \leq \frac{n_2(2\rho_{xz}^2 - 1) - n_1(2\rho_{xz}^2 + 1)\sigma_\nu^2}{(n_1 + n_2)(1 - 2\rho_{xz}^2)} \frac{\sigma_\nu^2}{\sigma_\eta^2} = \frac{n_1}{n_1 + n_2} \frac{2\rho_{xz}^2 + 1}{2\rho_{xz}^2 - 1} \frac{\sigma_\nu^2}{\sigma_\eta^2} - \frac{n_2}{n_1 + n_2} \frac{\sigma_\nu^2}{\sigma_\eta^2}.$$

In this case there is an upper bound to the values of θ^2 for which $\text{var}(\tilde{\beta}_2)$ is less than $\text{var}(\hat{\beta}_2)$. The limit is restrictive especially when ρ_{xz}^2 is close to 1 or n_2 is larger than n_1 . The equality in Equation (A.2) holds when

$$\theta^2 = \frac{n_1}{n_1 + n_2} \frac{2\rho_{xz}^2 + 1}{2\rho_{xz}^2 - 1} \frac{\sigma_\nu^2}{\sigma_\eta^2} - \frac{n_2}{n_1 + n_2} \frac{\sigma_\nu^2}{\sigma_\eta^2}, \quad 0.5 < \rho_{xz}^2 < 1.$$

When $\rho_{xz}^2 = 1$, the variances and covariances do not exist.

Figure 1.

Under the linear model, fixing n_1/n_2 at 1, the relative efficiency of $\tilde{\beta}_2$ with respect to $\hat{\beta}_2$ for varying correlation ρ_{xz} between x and z , θ , or the Partial R^2 of z given x .

Figure 2. Under the linear model, fixing θ at 1, the relative efficiency of $\tilde{\beta}_2$ with respect to $\hat{\beta}_2$ as a function of n_1/n_2 for different correlations ρ_{xz} between x and z .

Figure 3.
The empirical relative efficiency of $\tilde{\beta}_2$ with respect to $\hat{\beta}_2$ as a function of ρ_{xz} and θ under the logistic model, with $n_1 = n_2 = 1,000$.

Figure 4. Under the logistic model, fixing θ at 1, the relative efficiency of $\tilde{\beta}_2$ with respect to $\hat{\beta}_2$ as a function of n_1/n_2 for different correlations ρ_{xz} between x and z .

Figure 5.

Under the linear model, block 2 data is fixed as the data of June 2003, while block 1 data varies from May 2003 to January–May 2003. That is, the ratio of sample sizes n_1/n_2 changes from 1 to 5. In the left plot, each line comes from a three-variable case where the new variable z changes among the five types of delay. In the right plot, each line corresponds to a multiple linear regression with different dimension q of $\mathbf{z}$. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2)/\text{Var}(\hat{\beta}_2)$.

Figure 6.

Under the logistic model, block 2 data is fixed as the data of June 2003, while block 1 data varies from May 2003 to January–May 2003. That is, the ratio of sample sizes n_1/n_2 changes from 1 to 5. In the left plot, each line comes from a three-variable case where the new variable z changes among the five types of delay. In the right plot, each line is a regression with different dimension q of $\mathbf{z}$. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2)/\text{Var}(\hat{\beta}_2)$.

Table 1

Under linear or logistic models, biases of $\hat{\beta}_2$ and $\tilde{\beta}_2$, standard errors (SE; $\times 10$) of $\hat{\beta}_2$ and $\tilde{\beta}_2$, and the correlation between $\tilde{\beta}_2$ and $\tilde{\theta}_2$.

ρ_{xz}	Bias($\hat{\beta}_2$)	Bias($\tilde{\beta}_2$)	SE($\hat{\beta}_2$)		SE($\tilde{\beta}_2$)		Corr($\tilde{\beta}_2, \hat{\theta}_2$)	
			Average	Empirical	Average	Empirical	Average	Empirical
Linear Model								
$\sqrt{0.1}$	0.000315	0.000230	0.334	0.330	0.252	0.251	-0.418	-0.420
$\sqrt{0.5}$	0.000491	0.000087	0.448	0.449	0.375	0.380	-0.845	-0.847
$\sqrt{0.7}$	0.000549	0.000264	0.578	0.575	0.517	0.516	-0.935	-0.935
EE Approach under Logistic Model								
$\sqrt{0.1}$	0.005012	0.000433	0.988	0.986	0.448	0.453	0.033	0.035
$\sqrt{0.5}$	0.005381	-0.000549	1.282	1.305	0.738	0.742	-0.761	-0.759
$\sqrt{0.7}$	0.013245	0.003062	1.616	1.609	1.176	1.176	-0.896	-0.892

Table 2

Under linear regression with five variables (x_1 to x_5) in the first block, adding z_1 and z_2 in the second block, and further adding z_3 and z_4 in the third block, the relative efficiencies for estimates corresponding to existing variables at block 2 and block 3.

Variables	Relative Efficiency					
	$\rho = \sqrt{0.1}$		$\rho = \sqrt{0.5}$		$\rho = \sqrt{0.7}$	
	Block 2	Block 3	Block 2	Block 3	Block 2	Block 3
β_1	1.563	1.361	1.655	1.627	1.719	1.861
β_2	1.563	1.361	1.654	1.627	1.718	1.862
β_3	1.562	1.361	1.655	1.628	1.718	1.862
β_4	1.563	1.361	1.655	1.627	1.718	1.862
β_5	1.563	1.361	1.655	1.627	1.718	1.861
θ_1		1.318		1.498		1.626
θ_2		1.318		1.498		1.625

Table 3

Under logistic regression with five variables (x_1 to x_5) in the first block, adding z_1 and z_2 in the second block, and further adding z_3 and z_4 in the third block, the relative efficiencies for estimates corresponding to existing variables at block 2 and block 3.

Variables	Relative Efficiency					
	$\rho = \sqrt{0.1}$		$\rho = \sqrt{0.5}$		$\rho = \sqrt{0.7}$	
	Block 2	Block 3	Block 2	Block 3	Block 2	Block 3
β_1	1.852	1.829	1.920	2.180	1.921	2.381
β_2	1.856	1.967	1.880	2.271	1.881	2.439
β_3	1.832	2.071	1.851	2.330	1.860	2.472
β_4	1.744	2.206	1.789	2.442	1.820	2.551
β_5	1.702	2.175	1.765	2.474	1.804	2.582
θ_1		1.671		1.785		1.841
θ_2		1.614		1.749		1.820

Table 4

In the linear model for the airline data analysis, y is the delay time, x is the distance, and the newly added variable z is one of the five types of delay. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2)/\text{Var}(\hat{\beta}_2)$.

	ρ_{xz}	t -ratio	Variance Ratio
Security	0.004	48.52	0.49
Weather	-0.006	231.79	0.53
NAS	0.016	414.26	0.61
LateAircraft	-0.015	484.90	0.65
Carrier	0.017	538.63	0.67

Table 5

In the multiple linear regressions where y is the delay time, x is the distance, types of delay are added into the model one by one as new variables. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2) / \text{Var}(\hat{\beta}_2)$.

New Variables (Type of Delay)	Variance Ratio
Security	0.49
Security, Weather	0.54
Security, Weather, NAS	0.66
Security, Weather, NAS, LateAircraft	0.79
Security, Weather, NAS, LateAircraft, Carrier	0.96

Table 6

In a logistic regression where y is the arrival delay (binary), x is the distance, and the newly added variable z is one of the five types of delay. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2) / \text{Var}(\hat{\beta}_2)$.

	<i>t</i> -ratio	Variance Ratio
Security	23.35	0.55
Weather	39.00	0.57
LateAircraft	113.34	0.67
Carrier	126.47	0.70
NAS	170.76	0.79

Table 7

In the logistic regression where y is the arrival delay (binary), x is the distance, types of delay are added into the model one by one as new variables. The variance ratio is defined as $\text{Var}(\tilde{\beta}_2) / \text{Var}(\hat{\beta}_2)$.

New Variables (Type of Delay)	Variance Ratio
Security	0.55
Security, Weather	0.58
Security, Weather, LateAircraft	0.69
Security, Weather, LateAircraft, Carrier	0.82
Security, Weather, LateAircraft, Carrier, NAS	1.00