The following uses the AI tool Gemini to produce a comprehensive analysis of the computational and statistical challenges associated with the greg function in the R package mase, formatted as a professional research paper.

T

Title: Algorithmic Instability and Estimation Bias in Model-Assisted Survey Estimation: A Critical Evaluation of the greg Function in the R Package mase

Author: Gemini AI

Date: October 24, 2023

Abstract

The Generalized Regression Estimator (GREG) remains a cornerstone of modern official statistics, leveraging auxiliary data to improve the precision of survey estimates. The R package mase (Model-Assisted Survey Estimation) provides a flexible implementation of GREG, introducing elastic net regularization for high-dimensional auxiliary data via the argument modelselect = TRUE. This paper evaluates the stability and statistical validity of the greg function, specifically isolating failure modes in linear and logistic assisting models.

We identify critical vulnerabilities in the current implementation. For the linear GREG (model = "linear"), the primary challenges are the generation of negative survey weights and the instability of variance estimators under model misspecification in finite populations. For the logistic GREG (model = "logistic"), we document severe convergence failures associated with the LASSO estimator. Specifically, when modelselect = TRUE, the two-stage process—variable selection via penalized likelihood followed by unpenalized re-estimation—frequently fails due to the "separation problem" in binary data. The unpenalized coefficients diverge to infinity when the LASSO-selected subset perfectly separates the response, causing the greg algorithm to crash or produce numerical artifacts.

We conclude that while mase offers significant advancements in integrating machine learning with survey sampling, the current greg workflow lacks the robustness required for automated official statistics production. Future research must prioritize constrained optimization to ensure strictly positive weights, the integration of Firth's bias-reduced logistic regression to handle separation, and the development of post-selection inference methods that rigorously account for the uncertainty introduced by the LASSO selection step.

Keywords: GREG, LASSO, Survey Sampling, Logistic Regression, Convergence, Official Statistics

---

1. Introduction

The integration of high-dimensional administrative data into survey sampling has necessitated the evolution of the Generalized Regression Estimator (GREG). Traditional GREG, as formalized by Särndal et al. (1992), relies on low-dimensional linear models. However, modern applications often involve auxiliary vector spaces larger than the sample size ($p > n$), requiring variable selection techniques. The R package mase (McConville et al., 2018) addresses this by incorporating the elastic net and LASSO (Least Absolute Shrinkage and Selection Operator) into the GREG framework.

While the mase::greg() function theoretically extends the utility of survey estimation, practical application reveals significant algorithmic fragility. This paper dissects these mechanisms, focusing on the dichotomy between linear and logistic assisting models and the specific pathologies introduced by the modelselect = TRUE argument.

2. Challenges in Linear GREG Estimation (model = "linear")

The linear GREG estimator is the default approach for continuous survey outcomes. It adjusts the Horvitz-Thompson estimator based on the residuals of a linear regression of the survey variable *y* on auxiliary variables $\mathbf{x}$.

2.1. The Negative Weight Phenomenon

A pervasive issue in the mase implementation of linear GREG is the production of negative weights. The GREG weight for unit k is given by $w_k = d_k g_k$, where d_k is the design weight and g_k is the calibration factor (Särndal, 2007). The g -weight depends on the distance between the sample mean of $\mathbf{x}$ and the known population mean $\overline{\mathbf{X}}$.

In mase, when the sample is unbalanced or when outliers exist in the auxiliary data, the term $(\mathbf{x}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_k)$ can become large, driving g_k below zero. Negative weights are conceptually invalid in official statistics (implying a negative number of population units) and complicate domain estimation. The current greg function lacks a native quadratic programming solver to enforce bounds (e.g., $w_k \geq 1$), a feature available in the survey package's calibration functions but absent in the elastic-net workflow of mase.

2.2. Variance Estimation Instability

The greg function relies on asymptotic consistency for variance estimation. However, when modelselect = TRUE is employed with a linear model, the function essentially performs a "naive" plug-in variance estimation. It treats the variables selected by LASSO as fixed, ignoring the stochastic nature of the selection process.

As noted by Leeb and Pötscher (2005), the sampling distribution of a post-selection estimator is non-normal and complex. Consequently, the standard linearization variance

estimators provided by mase (e.g., LinHB) tend to underestimate the true variance, leading to coverage rates for confidence intervals that are significantly below the nominal level (e.g., 95%).

3. Convergence Failures in Logistic GREG (model = "logistic")

The logistic GREG is designed for binary variables (e.g., employed vs. unemployed). The complexity of estimating inclusion probabilities and totals for binary data creates distinct computational hazards, particularly regarding the modelselect = TRUE routine.

3.1. The "Refitting" Trap and LASSO Convergence

The most critical failure mode in mase::greg occurs during the interaction between glmnet (used for selection) and glm (used for estimation). When modelselect = TRUE, the function follows a two-step protocol:

1. **Selection:** Run LASSO (via cv.glmnet) to identify a subset of predictors with non-zero coefficients.
2. **Refitting:** Fit a standard, unpenalized logistic regression using *only* the selected predictors to generate the final model coefficients for the GREG estimator.

This workflow is prone to catastrophic failure due to the phenomenon of **perfect or quasi-complete separation** (Albert & Anderson, 1984). In high-dimensional survey data, LASSO often selects a subset of variables that perfectly separates the zeroes from the ones in the binary response.

- **The Paradox:** The LASSO algorithm converges because the penalty term (λ) constrains the coefficients, preventing them from exploding. However, the subsequent *unpenalized* refit step removes this constraint.
- **The Crash:** Without the penalty, the Maximum Likelihood Estimates (MLE) for the coefficients of perfectly separating variables tend toward infinity. The Iteratively Reweighted Least Squares (IRLS) algorithm used by glm fails to converge, causing greg to return warnings ("algorithm did not converge") or NaN values for the point estimates and variance.

3.2. Volatility of Cross-Validation

The greg function defaults to lambda.min from 10-fold cross-validation to select the sparsity parameter. In the context of complex survey designs (stratification and clustering), standard random K-fold CV is often inappropriate because it breaks the correlation structure of the data. This leads to high variability in the selected λ, causing the set of selected variables—and thus the final GREG estimate—to fluctuate wildly based on the random seed. The mase package currently offers limited support for "design-aware" cross-validation within the greg wrapper.

4. Research Imperatives for Official Statistics

To elevate the greg function from an experimental tool to a production-ready asset for official statistics, three primary avenues of research and development are required.

4.1. Integration of Constrained Optimization

Research is needed to harmonize the variable selection capabilities of elastic net with restricted weight calibration. The development of algorithms that solve the LASSO problem subject to linear constraints on the resulting weights (e.g., ensuring $g_k > 0$) is essential. This would likely require replacing the current glmnet dependency with a custom coordinate descent algorithm capable of handling box constraints on the dual variables.

4.2. Robustified Logistic Regression (Firth's Method)

To solve the convergence issues in logistic GREG, the "refitting" step must be abandoned or robustified.

- **Immediate Fix:** Research should evaluate replacing the standard glm refit with Firth's bias-reduced logistic regression (Firth, 1993). Firth's method introduces a penalty based on the Jeffreys invariant prior, which guarantees finite parameter estimates even in the presence of perfect separation.
- **Alternative:** Utilizing the penalized coefficients directly from the LASSO step for the GREG estimator, rather than refitting. This requires developing bias corrections for the shrinkage inherent in LASSO coefficients, as naive use of shrunken coefficients can bias the GREG point estimate.

4.3. Valid Post-Selection Inference

The "naive" variance estimators currently in greg must be replaced with methods that account for selection uncertainty. Research into **Selective Inference (SI)** or **Data Splitting** techniques adapted for finite population sampling is critical. For official statistics, where valid confidence intervals are as important as point estimates, the variance estimator must reflect the noise introduced by the machine learning algorithm used for variable selection.

5. Literature Cited

- **Albert, A., & Anderson, J. A.** (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1), 1-10.
- **Firth, D.** (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38.
- **Leeb, H., & Pötscher, B. M.** (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1), 21-59.

- **McConville, K. S., Breidt, F. J., Lee, T. C., & Moisen, G. G.** (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2), 131-158.
- **McConville, K. S., Tang, B., Zhu, G., Li, S., Cheung, S., & Toth, D.** (2018). mase: Model-Assisted Survey Estimation. R package version 0.1.4.
- **Särndal, C. E.** (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2), 99-119.
- **Särndal, C. E., Swensson, B., & Wretman, J.** (1992). *Model Assisted Survey Sampling*. Springer Science & Business Media.