Robust Generalized Regression Predictor

Tobias Schoch

University of Applied Sciences Northwestern Switzerland FHNW
School of Business, Riggenbachstrasse 16, CH-4600 Olten
tobias.schoch@fhnw.ch

September 14, 2022

1 Introduction

The *population* regression model is given by

$$\xi: \quad Y_i = \mathbf{x}_i^T \boldsymbol{\theta} + \sigma\sqrt{v_i}E_i, \qquad \boldsymbol{\theta} \in \mathbb{R}^p, \quad \sigma > 0, \quad i \in U, \tag{1}$$

where

- the population U is of size N,
- the parameters $\boldsymbol{\theta}$ and σ are unknown,
- the $\mathbf{x}_i$'s are known values (possibly containing outliers), $\mathbf{x}_i \in \mathbb{R}^p$, $1 \leq p < N$; we denote the design matrix by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$,
- the v_i's are known positive (heteroscedasticity) constants,
- the errors E_i are independent and identically distributed (i.i.d.) random variables (r.v.) with zero expectation and unit variance,
- it is assumed that $\sum_{i \in U} \mathbf{x}_i \mathbf{x}_i^T / v_i$ is a non-singular $(p \times p)$ matrix.

Remarks. The i.i.d. assumption on the errors E_i is rather strict. This assumption can be replaced by the assumption that the E_i are identically distributed r.v. such that $\mathbb{E}_\xi(E_i \mid \mathbf{x}_i, \ldots, \mathbf{x}_N) = 0$ and $\mathbb{E}_\xi(E_i E_j \mid \mathbf{x}_i, \ldots, \mathbf{x}_N) = 1$ if $i = j$ and zero otherwise for all $i, j \in U$, where $\mathbb{E}_\xi$ denotes expectation w.r.t. model ξ in (1). Another generalization obtains by requiring that $\mathbb{E}_\xi(E_i \mathbf{x}_i) = \mathbb{E}_\xi(\mathbf{x}_i E_i) = \mathbf{0}$ in place of the conditional expectation. If the distribution of the errors E_i is asymmetric with non-zero mean, the regression intercept and the errors are confounded. The slope parameters, however, are identifiable with asymmetric distributions (Carroll and Welsh, 1988). In the context of GREG prediction, however, we deal with prediction under the model. Thus, identifiability is not an issue.

It is assumed that a sample s is drawn from U with sampling design $p(s)$ such that the independence (orthogonality) structure of the model errors in (1) is maintained. The sample regression

M- and GM-estimator of $\boldsymbol{\theta}$ are defined as the root to the following estimating equations (cf. Hampel, Ronchetti, Rousseeuw, and Stahel, 1986, Chapter 6.3)

$$\sum_{i \in s} \frac{w_i}{\sqrt{v_i}} \psi_k(r_i) \mathbf{x}_i = \mathbf{0} \qquad (M\text{-estimator}),$$

$$\sum_{i \in s} \frac{w_i}{\sqrt{v_i}} h(\mathbf{x}_i) \psi_k(r_i) \mathbf{x}_i = \mathbf{0} \qquad (\text{Mallows } GM\text{-estimator}),$$

$$\sum_{i \in s} \frac{w_i}{\sqrt{v_i}} h(\mathbf{x}_i) \psi_k \left(\frac{r_i}{h(\mathbf{x}_i)} \right) \mathbf{x}_i = \mathbf{0} \qquad (\text{Schweppe } GM\text{-estimator}),$$

where

- w_i is the sampling weight,
- ψ_k is a *generic* ψ-function indexed by the robustness tuning constant k,
- r_i is the standardized residual, defined as

$$r_i = \frac{y_i - \mathbf{x}_i^T \boldsymbol{\theta}}{\sigma \sqrt{v_i}}, \tag{2}$$

- $h : \mathbb{R}^p \to \mathbb{R}_+$ is a weight function,
- σ is the regression scale which is estimated by the (normalized) weighted median of the absolute deviations from the weighted median of the residuals.

The Huber and Tukey bisquare (biweight) ψ-functions are denoted by, respectively, ψ_k^{hub} and ψ_k^{tuk}. The sample-based estimators of $\boldsymbol{\theta}$ can be written as a weighted least squares problem

$$\sum_{i \in s} \frac{w_i}{v_i} u_i(r_i, k)(y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}_n) \mathbf{x}_i = \mathbf{0},$$

where

$$u_i(r_i, k) = \begin{cases} \dfrac{\psi_k(r_i)}{r_i} & M\text{-estimator}, \\ h(\mathbf{x}_i) \dfrac{\psi_k(r_i)}{r_i} & \text{Mallows } GM\text{-estimator}, \\ \dfrac{\psi_k(r_i^*)}{r_i^*}, \quad \text{where} \quad r_i^* = \dfrac{r_i}{h(\mathbf{x}_i)} & \text{Schweppe } GM\text{-estimator}, \end{cases} \tag{3}$$

and k denotes the robustness tuning constant.

2 Representation of the robust GREG as a QR-predictor

The robust GREG predictor of the population y-total can be written in terms of the g-weights (see e.g. Särndal, Swensson, and Wretman, 1992, Chapter 6) as

$$\widehat{t}_y^{rob} = \sum_{i \in s} g_i y_i, \tag{4}$$

where the g-weights are defined as (Duchesne, 1999)

$$g_i = b_i + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{b\mathbf{x}}\right)^T \left(\sum_{i \in s} q_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} q_i \mathbf{x}_i, \tag{5}$$

where $\widehat{\mathbf{t}}_{b\mathbf{x}} = \sum_{i \in s} b_i \mathbf{x}_i$ and $\mathbf{t_x} = \sum_{i \in U} \mathbf{x}_i$. The sampling weights, w_i, are "embedded" into the g-weights in (4).

In contrast to the non-robust "standard" GREG predictor, the g-weights in (5) depend on the study variable, y_i, through the choice of the constants $(q_i, b_i) = \{(q_i, b_i) : i \in s\}$. This will be easily recognized once we define the set of constants. The predictors of the population y-total that are defined in terms of the constants (q_i, r_i) form the class of *QR-predictor* due to (Wright, 1983). In passing we note that $\widehat{t}_y^{rob}$ can be expressed in a "standard" GREG representation. Let

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{i \in s} q_i \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i \in s} q_i \mathbf{x}_i y_i,$$

then $\widehat{t}_y^{rob}$ in (4) can be written as

$$\widehat{t}_y^{rob} = \sum_{i \in s} b_i y_i + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{b\mathbf{x}}\right)^T \widehat{\boldsymbol{\theta}} = \mathbf{t}_\mathbf{x}^T \widehat{\boldsymbol{\theta}} + \sum_{i \in s} b_i (y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}).$$

In the next two sections, we define the constants (q_i, b_i) of the QR-predictor.

2.1 Constants q_i of the QR-predictor

The set of constants $\{q_i\}$ is defined as

$$q_i = \frac{w_i \cdot u_i(r_i, k)}{v_i}, \qquad i = 1, \dots, n, \tag{6}$$

where v_i is given in (1) and $u_i(r_i, k)$ is defined in (3). The tuning constant k in $u_i(r_i, k)$ is the one that is used to estimate $\boldsymbol{\theta}$.

2.2 Constants b_i of the QR-predictor

The constants $\{b_i\}$ are predictor-specific. They depend on the argument `type`. Moreover, the b_i's depend on the robustness tuning constant `k` – which is an argument of `svymean_reg()` and `svytotal_reg()` – to control the robustness of the prediction. To distinguish it from the tuning constant k, which is used in fitting model ξ in (1), it will be denoted by κ. Seven sets $\{b_i\}$ are available.

`type = "projective":` $b_i \equiv 0$ (Särndal and Wright, 1984),

`type = "ADU":` $b_i \equiv w_i$ (Särndal et al., 1992, Chapter 6),

`type = "huber":` $b_i \equiv w_i \cdot u_i(r_i, \kappa)$, where u_i is defined in (3) with $\psi_k \equiv \psi_k^{hub}$ (Lee, 1995; Hulliger, 1995; Beaumont and Alavi, 2004),

`type = "tukey":` $b_i \equiv w_i \cdot u_i(r_i, \kappa)$, where u_i is defined in (3) with $\psi_k \equiv \psi_k^{tuk}$ (Lee, 1995; Hulliger, 1995; Beaumont and Alavi, 2004),

`type = "lee":` $b_i \equiv \kappa \cdot w_i$, where $0 \leq \kappa \leq 1$ (Lee, 1991, 1995),

`type = "BR":` $b_i \equiv w_i \cdot u_i(r_i, \kappa)$, where u_i is defined in (3) with ψ_k replaced by (Beaumont and Rivest, 2009)

$$\psi_k^{mod}(x) = \frac{x}{w_i} + \frac{w_i - 1}{w_i}\psi_k^{hub}(x),$$

`type = "duchesne":` $b_i \equiv w_i \cdot u_i(r_i; a, b)$, where u_i is defined in (3) with ψ_k replaced by (Duchesne, 1999)

$$\psi_{a,b}^{hub}(x) = \begin{cases} x & \text{if} \quad |x| \leq a, \\ a \cdot \mathrm{sign}(x) & \text{if} \quad |x| > a \quad \text{and} \quad |x| < a/b, \\ b \cdot x & \text{if} \quad |x| > a/b, \end{cases}$$

where $\psi_{a,b}^{hub}$ is a modified Huber ψ-function with tuning constants a and b (in place of κ). Duchesne (1999) suggested the default parametrization $a = 9$ and $b = 0.25$.

2.3 Implementation

Let $\mathbf{q} = (q_1, \ldots, q_n)^T$ and $\mathbf{b} = (b_1, \ldots, b_n)^T$, where q_i and b_i are defined in, respectively, (6) and Section 2.2. Put $\mathbf{Z} = \sqrt{\mathbf{q}} \circ \mathbf{X}$, where $\circ$ denotes Hadamard multiplication and the square root is applied element by element. The vector-valued g-weights, $\mathbf{g} = (g_1, \ldots, g_n)^T$, in (5) can be written as

$$\mathbf{g}^T = \mathbf{b}^T + \left(\mathbf{t_x} - \widehat{\mathbf{t}}_{b\mathbf{x}}\right)^T \underbrace{\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T}_{=\mathbf{H}, \text{ say}} \circ (\sqrt{\mathbf{q}})^T.$$

Define the QR factorization $\mathbf{Z} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is an upper triangular matrix (both of conformable size). Note that the matrix QR-factorization and Wright's QR-estimators have nothing in common besides the name; in particular, $\mathbf{q}$ and $\mathbf{Q}$ are unrelated. With this we have

$$\mathbf{H} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T = \mathbf{R}^{-1}\mathbf{Q}^T$$

and multiplying both sides by $\mathbf{R}$, we get $\mathbf{RH} = \mathbf{Q}^T$ which can be solved easily for $\mathbf{H}$ since $\mathbf{R}$ is an upper triangular matrix (see `base::backsolve()`). Thus, the g-weights can be computed as

$$\mathbf{g} = \mathbf{b} + \mathbf{H}^T\big(\mathbf{t_x} - \widehat{\mathbf{t}}_{b\mathbf{x}}\big) \circ \sqrt{\mathbf{q}},$$

where the $(p \times n)$ matrix $\mathbf{H}$ need not be explicitly transposed when using `base::crossprod()`. The terms $\mathbf{b}$ and $\widehat{\mathbf{t}}_{b\mathbf{x}}$ are easy to compute. Thus,

$$\widehat{t}_y^{rob} = \mathbf{g}^T\mathbf{y}, \qquad \text{where} \quad \mathbf{y} = (y_1, \ldots, y_n)^T.$$

3 Variance estimation

Remark. Inference of the regression estimator is only implemented under the assumption of representative outliers (in the sense of Chambers, 1986). We do not cover inference in presence of nonrepresentative outliers.

Our discussion for variance estimation follows the line of reasoning in Särndal et al. (1992, p. 233–234) on the variance of the non-robust GREG estimator. To this end, denote by $E_i = y_i - \mathbf{x}_i^T\boldsymbol{\theta}_N$, $i \in U$, the census residuals, where $\boldsymbol{\theta}_N$ is the census parameter. With this, any g-weighted predictor can be written as

$$\widehat{t}_y^{rob} = \sum_{i \in s} g_i y_i = \sum_{i \in s} g_i(\mathbf{x}_i^T\boldsymbol{\theta}_N + E_i) = \sum_{i \in U} \mathbf{x}_i^T\boldsymbol{\theta}_N + \sum_{i \in s} g_i E_i, \tag{7}$$

where we have used the fact that the g-weights in (5) satisfy the calibration property

$$\sum_{i \in s} g_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i.$$

The first term on the r.h.s. of the last equality in (7) is a population quantity and does therefore not contribute to the variance of $\widehat{t}_y^{rob}$. Thus, we can calculate the variance of the robust GREG predictor by

$$\mathrm{var}\left(\widehat{t}_y^{rob}\right) = \mathrm{var}\left(\sum_{i \in s} g_i E_i\right) \tag{8}$$

under the assumptions that (1) the E_i are known quantities and (2) the g_i do not depend on the y_i.

Disregarding the fact that the g-weights are sample dependent and substituting the sample residual r_i for E_i in (8), Särndal et al. (1992, p. 233-234 and Result 6.6.1) propose to estimate the variance of the GREG predictor by the g-weighted variance of the total $\sum_{i \in s} g_i r_i$. Following the same train of thought and disregarding in addition that the g_i depend on y_i, the variance of $\widehat{t}_y^{rob}$ can be approximated by

$$\widehat{\mathrm{var}}\big(\widehat{t}_y^{rob}\big) \approx \widehat{\mathrm{var}}\left(\sum_{i \in s} g_i r_i\right),$$

where $\widehat{\mathrm{var}}(\cdot)$ denotes a variance estimator of a total for the sampling design $p(s)$.

References

BEAUMONT, J.-F. AND A. ALAVI (2004). Robust Generalized Regression Estimation, *Survey Methodology* **30**, 195–208.

BEAUMONT, J.-F. AND L.-P. RIVEST (2009). Dealing with outliers in survey data, in *Sample Surveys: Theory, Methods and Inference*, ed. by D. Pfeffermann and C. R. Rao, Amsterdam: Elsevier, vol. 29A of *Handbook of Statistics*, chap. 11, 247–280.

CARROLL, R. J. AND A. H. WELSH (1988). A Note on Asymmetry and Robustness in Linear Regression, *The American Statistician* **42**, 285–287.

DUCHESNE, P. (1999). Robust calibration estimators, *Survey Methodology* **25**, 43–56.

HAMPEL, F. R., E. M. RONCHETTI, P. J. ROUSSEEUW, AND W. A. STAHEL (1986). *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley and Sons.

HULLIGER, B. (1995). Outlier Robust Horvitz–Thompson Estimators, *Survey Methodology* **21**, 79–87.

LEE, H. (1991). Model-Based Estimators That Are Robust to Outliers, in *Proceedings of the 1991 Annual Research Conference, Bureau of the Census 178-202. Washington, DC, Department of Commerce.*

LEE, H. (1995). Outliers in business surveys, in *Business survey methods*, ed. by B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: John Wiley and Sons, chap. 26, 503–526.

SÄRNDAL, C.-E., B. SWENSSON, AND J. WRETMAN (1992). *Model Assisted Survey Sampling*, New York: Springer.

SÄRNDAL, C.-E. AND R. L. WRIGHT (1984). Cosmetic Form of Estimators in Survey Sampling, *Scandinavian Journal of Statistics* **11**, 146–156.

WRIGHT, R. L. (1983). Finite population sampling with multivariate auxiliary information, *Journal of the American Statistical Association* **78**, 879–884.