The Linear Model in the Generalized Regression (GREG) Estimator

In model-assisted survey sampling, the **Generalized Regression Estimator (GREG)** is constructed using a "working model" (or assisting model) that describes the relationship between the variable of interest, y, and a vector of auxiliary variables, $\mathbf{x}$.

While the properties of the GREG (such as asymptotic unbiasedness) rely on the randomization distribution of the sampling design (the p-distribution), its **efficiency** (variance reduction) relies on how well this working model (ξ-model) explains the variability in the population.

1. The Finite Population and Auxiliary Data

Consider a finite population $U = \{1, \dots, N\}$.

- y_k: The value of the study variable for unit k.
- $\mathbf{x}_k$: A vector of known auxiliary variables for unit k, $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^\top$.
- T_x: The known population totals of the auxiliary variables, $T_x = \sum_U \mathbf{x}_k$.

We draw a probability sample s from U with inclusion probabilities $\pi_k = P(k \in s)$ and design weights $d_k = 1/\pi_k$.

2. The Linear "Working" Model (ξ)

The standard GREG estimator assumes a linear superpopulation model ξ. This model postulates that the population values y_k are generated as follows:

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k$$

The model assumptions for the error terms ε_k are:

1. **Zero Mean:** $E_\xi(\varepsilon_k) = 0$ (implies $E_\xi(y_k) = \mathbf{x}_k^\top \boldsymbol{\beta}$)
2. **Independence:** $E_\xi(\varepsilon_k \varepsilon_l) = 0$ for $k \neq l$
3. **Variance Structure (Heteroscedasticity):** $V_\xi(\varepsilon_k) = E_\xi(\varepsilon_k^2) = \sigma^2 v_k$

Key Parameters:

- $\boldsymbol{\beta}$: The vector of unknown regression coefficients.
- v_k: A known variance structure parameter associated with unit k. This allows the model to account for heteroscedasticity.
  - If $v_k =$, the model assumes homoscedasticity (constant variance).
  - If $v_k \propto x_k$, the variance increases with the size of the unit (common in business surveys).

3. Estimation of Model Parameters ($\hat{\mathbf{B}}$)

Since the true population parameter $\boldsymbol{\beta}$ is unknown, it must be estimated using the sample data. In the design-based GREG context, we use a **probability-weighted least squares** estimator (often denoted as $\hat{\mathbf{B}}$).

The estimator $\hat{\mathbf{B}}$ minimizes the weighted sum of squared residuals:

$$\hat{\mathbf{B}} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{k \in s} \frac{d_k (y_k - \mathbf{x}_k^\top \mathbf{b})^2}{v_k}$$

Solving this yields the explicit formula:

$$\hat{\mathbf{B}} = \left(\sum_{k \in s} \frac{d_k \mathbf{x}_k \mathbf{x}_k^\top}{v_k} \right)^{-1} \sum_{k \in s} \frac{d_k \mathbf{x}_k y_k}{v_k}$$

Note: The inclusion of the design weights d_k ensures that $\hat{\mathbf{B}}$ is a design-consistent estimator of the census regression coefficient $\mathbf{B}$ (the coefficient we would get if we ran the regression on the entire population).

4. The GREG Estimator Formula

The GREG estimator for the population total $T_y = \sum_U y_k$ combines the Horvitz-Thompson estimator with a model-based adjustment:

$$\hat{t}_{GREG} = \hat{t}_{y\pi} + (\mathbf{T}_x - \hat{\mathbf{t}}_{x\pi})^\top \hat{\mathbf{B}}$$

Where:

- $\hat{t}_{y\pi} = \sum_{k \in s} d_k y_k$: The Horvitz-Thompson estimator of y.
- $\mathbf{T}_x$: The *true* known population totals of the auxiliary variables.
- $\hat{\mathbf{t}}_{x\pi} = \sum_{k \in s} d_k \mathbf{x}_k$: The Horvitz-Thompson estimator of the auxiliary totals.

Alternative Prediction Form

The estimator can be intuitively rewritten as the sum of predicted values plus a bias correction for the residuals:

$$\hat{t}_{GREG} = \underbrace{\sum_{k \in U} \hat{y}_k}_{\text{Synthetic Term}} + \underbrace{\sum_{k \in s} d_k (y_k - \hat{y}_k)}_{\text{Bias Correction}}$$

- $\hat{y}_k = \mathbf{x}_k^\top \hat{\mathbf{B}}$: The predicted value for unit k based on the model.
- The first term projects the model over the entire population (using the known $\mathbf{T}_x$).
- The second term checks the error of the model on the sample (observed y vs. predicted $\hat{y}$) and adjusts the total up or down accordingly.

5. Common Special Cases

Different choices of the auxiliary vector $\mathbf{x}_k$ and the variance structure v_k lead to well-known estimators:

Estimator	Auxiliary Vector (xk)	Variance Structure (vk)

Ratio Estimator	Scalar x_k	$v_k =$ (Variance $\propto$ size)
Post-stratification	Dummy indicators for strata	$v_k =$ (Homoscedastic)
Simple Regression	$(1, x_k)$ (intercept + slope)	$v_k =$