

optimStrat: An R package for assisting the choice of the sampling strategy

Edgar Bueno

Department of Statistics
Stockholm University

Introduction/Summary

- The strategy that couples a π ps design with the regression estimator is sometimes called optimal.
- It can be shown that even under simple misspecifications of the model, this optimality breaks down.
- We propose a method for assisting the choice of the sampling strategy (in particular, the design) by taking into account misspecifications in the model.
- The method is implemented in an R package, `optimStrat`.

Super-population model

The statistician is willing to admit that the following model *adequately describes* the relation between \mathbf{y} and \mathbf{x} .

The values of \mathbf{y} are realizations of the model ξ_0

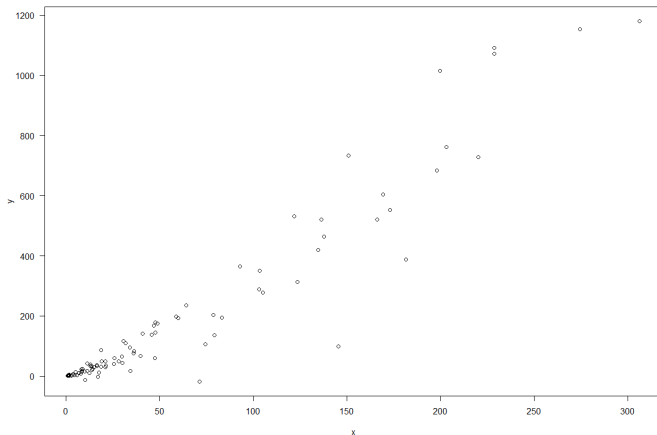
$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k$$

$$E_{\xi_0}(\epsilon_k) = 0 \quad V_{\xi_0}(\epsilon_k) = \delta_3 x_k^{2\delta_4} \quad E_{\xi_0}(\epsilon_k \epsilon_l) = 0 \quad (k \neq l)$$

where moments are taken with respect to the model ξ_0 and δ_i are constant parameters.

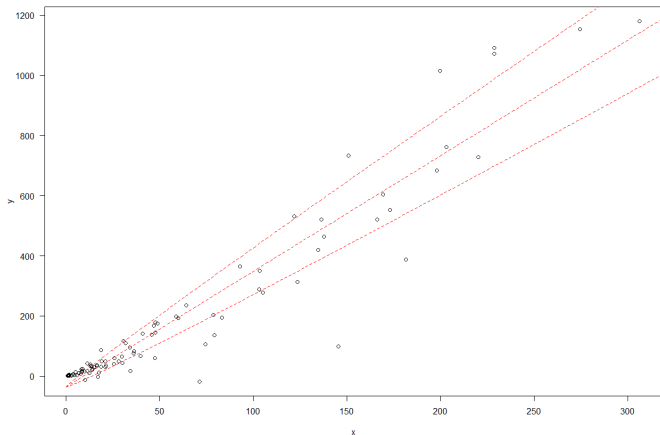
$\delta_0 + \delta_1 x_k^{\delta_2}$ will be called *trend* and $\delta_3 x_k^{2\delta_4}$ will be called *spread*.

Super-population model



$$\delta_0 + \delta_1 x_k^{\delta_2}$$
$$\delta_3 x_k^{2\delta_4}$$

Super-population model



$$\delta_0 + \delta_1 x_k^{\delta_2}$$
$$\delta_3 x_k^{2\delta_4}$$

The optimal strategy

The sampling strategy

- π ps with $\pi_k \propto x_k^{\delta_4}$,
- regression estimator with $\mathbf{x}_k = (1, x_k^{\delta_2})$

minimizes the anticipated Mean Squared Error

$$\text{MSE}_{\xi_0 p}(\hat{t}_y) = E_{\xi_0} \text{MSE}_p(\hat{t}_y) = E_{\xi_0} E_p ((\hat{t}_y - t_y)^2)$$

This optimality is model-dependent.

Model-based stratification is an alternative that has shown to be sometimes more efficient.

The misspecified model

The model

$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k$$

$$E_{\xi_0}(\epsilon_k) = 0 \quad V_{\xi_0}(\epsilon_k) = \delta_3 x_k^{2\delta_4} \quad E_{\xi_0}(\epsilon_k \epsilon_l) = 0 \quad (k \neq l)$$

is assumed. But the true model is of the form

$$Y_k = \beta_0 + \beta_1 x_k^{\beta_2} + \epsilon_k$$

$$E_{\xi_0}(\epsilon_k) = 0 \quad V_{\xi_0}(\epsilon_k) = \beta_3 x_k^{2\beta_4} \quad E_{\xi_0}(\epsilon_k \epsilon_l) = 0 \quad (k \neq l)$$

with $\beta_2 \neq \delta_2$ or $\beta_4 \neq \delta_4$.

It can be shown that π ps-reg is not necessarily optimal anymore.

Expected MSE under the misspecified model

For any design $p(\cdot)$, the expected MSE can be approximated by

$$\text{MSE}_{\xi p}(\hat{t}_y) \approx \beta_1^2 \left[\text{MSE}_p \left(\sum_s \frac{v_k}{\pi_k} \right) + F_0 \left(\sum_U \frac{x_k^{2\beta_4}}{\pi_k} - \sum_U x_k^{2\beta_4} \right) \right]$$

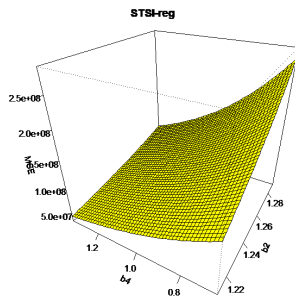
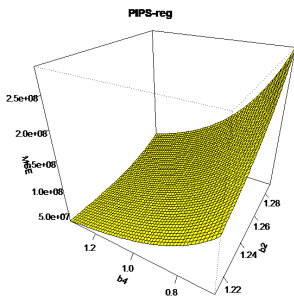
with

$$F_0 = \frac{1}{\overline{x^{2\beta_4}}} \frac{S_{1,\beta}^2}{S_{1,1}} \left(\frac{1}{R_{x,y}^2} - \frac{1}{R_{1,\beta}^2} \right)$$
$$v_k = \left(x_k^{\beta_2} - \overline{x^{\beta_2}} \right) - \left(x_k^{\delta_2} - \overline{x^{\delta_2}} \right) \frac{S_{\beta,\delta}}{S_{\delta,\delta}}$$

which can be seen a function of β_2 and β_4 (and $R_{x,y}$).

Choosing the sampling design

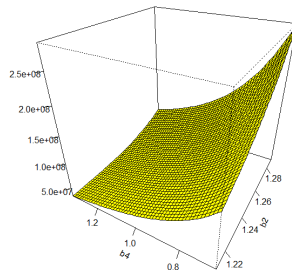
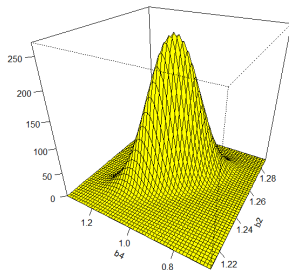
For any design, $\text{MSE}_{\xi_p}(\hat{t}_y)$ can be evaluated in a neighborhood of (δ_2, δ_4) and it indicates the MSE that is expected if (δ_2, δ_4) is assumed when the true parameter is (β_2, β_4) .



Choosing the sampling design

A prior distribution on (β_2, β_4) , $h(\beta)$, can be proposed and the expected value of $\text{MSE}_{\xi p}(\hat{t}_y)$ under this prior (the risk) obtained.

$$R(p) = E_h(\text{MSE}_{\xi p}(\beta|x, \delta, \sigma)) = \int_{\Theta} h(\beta) \cdot \text{MSE}_{\xi p}(\beta|x, \delta, \sigma) d\beta$$



Choosing the sampling design

The package allows for calculating $\text{MSE}_{\xi p}(\hat{t}_y)$ for the more general model misspecification:

Working model:

$$Y_k = \sum_{j=1}^J \delta_{1,j} x_{jk}^{\delta_{1,J+j}} + \epsilon_k \quad \text{with} \quad V_{\xi_0}(\epsilon_k) = \sum_{j=1}^J \delta_{2,j} x_{jk}^{\delta_{2,J+j}}$$

True model:






$$Y_k = \sum_{j=1}^J \beta_{1,j} x_{jk}^{\beta_{1,J+j}} + \epsilon_k \quad \text{with} \quad V_{\xi_0}(\epsilon_k) = \sum_{j=1}^J \beta_{2,j} x_{jk}^{\beta_{2,J+j}}$$

Choosing the sampling design






It might be argued that by introducing $h(\beta)$ an additional source of subjectivity has been added to the choice of the sampling design. However, our view is that it is more subjective to completely rely on some assumptions without any assessment of them and that if even under a simple misspecification, as the one represented by ξ , π_{ps} fails in minimizing the MSE, this should be evidence for not using it.

optimStrat

Bibliography I

-  Godambe, V.P. (1955). *A unified theory of sampling from finite populations*. Journal of the Royal Statistical Society, Series B **17**, 269-278.
-  Hájek, J. (1959) *Optimal Strategy and Other Problems in Probability Sampling* Casopis pro pestování matematiky, Vol. 84, No. 4, 387-423.
-  Holmberg, A. and Swensson, B. (2001). *On Pareto πps Sampling: Reflections on Unequal Probability Sampling Strategies*. Theory of Stochastic Processes, **7(23)**, 142-155.
-  Isaki, C.T. and Fuller, W.A. (1982) *Survey design under the regression superpopulation model*. Journal of the American Statistical Association **77**, 89-96.
-  Lanke, J. (1973). *On UMV-estimators in Survey Sampling*. Metrika **20**, 196-202.

Bibliography II

-  Nedyalkova, D. and Tillé, Y. (2008) *Optimal Sampling and Estimation Strategies under the Linear Model*. Biometrika, **95**, 3, pp. 521–537.
-  Rosén, B. (1997). *On sampling with probability proportional to size*. Journal of statistical planning and inference **62**, 159-191.
-  Särndal, C.E., Thomsen, Ib., Hoem, J., Lindley, D., Barndorff-Nielsen, O. and Dalenius, T. (1978). *Design-Based and Model-Based Inference in Survey Sampling*. Scandinavian Journal of Statistics, Vol. 5, No. 1, pp. 27–52.
-  Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
-  Wright, R.L. (1983). *Finite Population Sampling with Multivariate Auxiliary Information*. Journal of the American Statistical Association, **78**, 879—884.

Thanks for your attention!

Edgar Bueno, embuenoc@stat.su.se