

# A comparison of stratified simple random sampling and sampling with probability proportional to size

**Edgar Bueno**  
**Dan Hedlin**  
**Per Gösta Andersson**

Department of Statistics  
Stockholm University

# Introduction

## Objective:

To find an **efficient strategy** (in terms of variance) for estimating the **total** of a study variable,  $y$ .

$y$  is known to be **right-skewed**.

One quantitative **auxiliary variable**,  $x$ , is available.

We will work under the model assisted approach.

# General Regression Estimator

$$\hat{t}_{GREG} = \sum_U \hat{y}_k + \sum_s \frac{e_{ks}}{\pi_k}$$

with  $\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}$  and  $e_{ks} = y_k - \hat{y}_k$ , where

$$\hat{\mathbf{B}} = \left( \sum_s \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k \pi_k} \right)^{-} \sum_s \frac{\mathbf{x}_k y_k}{a_k \pi_k}.$$

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}.$

# General Regression Estimator, Case 1

Let  $\mathbf{x}_k = 0$  for all  $k \in U$ , we have

$$\hat{\mathbf{B}} = \left( \sum_s \frac{\mathbf{x}'_k \mathbf{x}_k}{a_k \pi_k} \right)^{-} \sum_s \frac{\mathbf{x}'_k y_k}{a_k \pi_k} = 0$$

Then  $\hat{y}_k = \mathbf{x}_k \hat{\mathbf{B}} = 0$  and  $e_{ks} = y_k - \hat{y}_k = y_k - 0 = y_k$ .

The GREG-estimator becomes

$$\hat{t}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s \frac{e_{ks}}{\pi_k} = \sum_U 0 + \sum_s \frac{y_k}{\pi_k} = \hat{t}_{\pi}$$

The **HT-estimator** can be seen as the case where no auxiliary information is used into the GREG-estimator.

## General Regression Estimator, Case 2

Let  $a_k = c_j$  and  $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{Jk})$  with  $x_{jk}$  defined as

$$x_{jk} = \begin{cases} 1 & \text{if } k \in U'_j \\ 0 & \text{if not} \end{cases}$$

where the  $U'_j$  ( $j = 1, \dots, J$ ) form a partition of  $U$ .

The **post-stratified estimator** is obtained when this type of auxiliary information is used in the GREG-estimator.

The residuals become  $E_k = y_k - \bar{y}_{U'_j}$  ( $k \in U'_j$ ).

## General Regression Estimator, Case 3

Let  $a_k = c$  and  $\mathbf{x}_k = (1, z_k)$ , with  $z_k = f(x_k)$  and  $f$  known.

The **regression estimator** is obtained when this  $\mathbf{x}_k$  is used in the GREG-estimator.

The residuals become

$$E_k = y_k + B_2 \frac{t_z}{N} - \frac{t_y}{N} - B_2 z_k \quad \text{with} \quad B_2 = \frac{N t_{zy} - t_z t_y}{N t_{z^2} - t_z^2}$$

where  $t_y = \sum_U y_k$ ,  $t_z = \sum_U z_k$ ,  $t_{z^2} = \sum_U z_k^2$  and  $t_{zy} = \sum_U z_k y_k$ .

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②
- ③

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②  $\pi_k \propto E_k$ ;
- ③



# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②  $\pi_k \propto E_k$ ;
- ③

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^- \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②  $\pi_k \propto E_k$ ;
- ③  $\pi_k \propto |E_k|$  together with  $\pi_{kl} = \pi_k \pi_l$  if  $k \in U^+$  and  $l \in U^-$ ;

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②  $\pi_k \propto E_k$ ;
- ③  $\pi_k \propto |E_k|$  together with  $\pi_{kl} = \pi_k \pi_l$  if  $k \in U^+$  and  $l \in U^-$ ;

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ;
- ②  $\pi_k \propto E_k$ ;
- ③  $\pi_k \propto |E_k|$  together with  $\pi_{kl} = \pi_k \pi_l$  if  $k \in U^+$  and  $l \in U^-$ ;

Although not leading to a zero-variance, we can consider

- ②  $\pi_k \propto |E_k|$ .

# General Regression Estimator

$$AV_p(\hat{t}_{GREG}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left( \frac{E_k}{\pi_k} - \frac{E_l}{\pi_l} \right)^2$$

with  $E_k = y_k - \mathbf{x}'_k \mathbf{B}$  where  $\mathbf{B} = \left( \sum_U \frac{\mathbf{x}_k \mathbf{x}'_k}{a_k} \right)^{-} \sum_U \frac{\mathbf{x}_k y_k}{a_k}$ .

The following are sufficient conditions for a zero-variance:

- ①  $E_k = 0$  for all  $k \in U$ ; **Estimator**
- ②  $\pi_k \propto E_k$
- ③  $\pi_k \propto |E_k|$  together with  $\pi_{kl} = \pi_k \pi_l$  if  $k \in U^+$  and  $l \in U^-$ ;

Although not leading to a zero-variance, we can consider

- ②  $\pi_k \propto |E_k|$ . **Design**

# 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  $\xi_0$

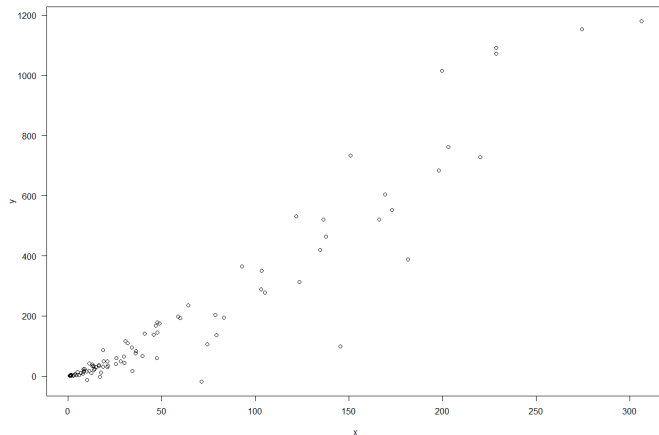
$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k$$

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

where moments are taken with respect to the model  $\xi_0$  and  $\delta_i$  are constant parameters.

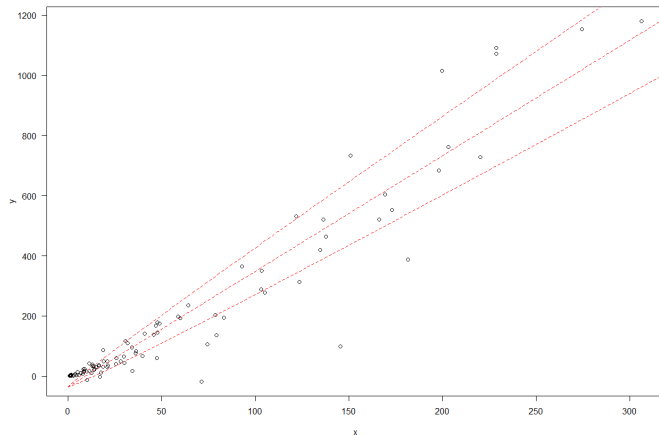
$\delta_0 + \delta_1 x_k^{\delta_2}$  will be called *trend* and  $\delta_3 x_k^{2\delta_4}$  will be called *spread*.

# Super-population model



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

# Super-population model



$$\delta_0 + \delta_1 x_k^{\delta_2}$$

$$\delta_3 x_k^{2\delta_4}$$



# Strategy $\pi_{ps} \text{---reg}$

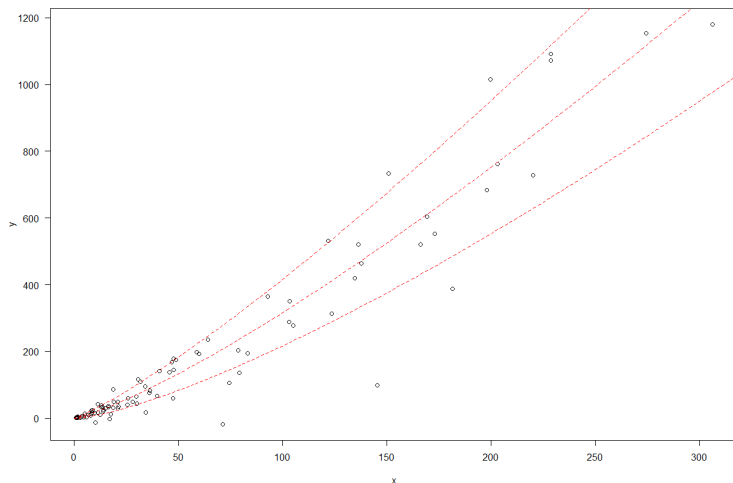
$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k \quad \text{with} \quad V_{\xi_0}(\epsilon_k) = \delta_3 x_k^{2\delta_4}$$

If  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, it is natural to use  $\mathbf{x}_k = (1, x_k^{\delta_2})$  in the GREG-estimator.

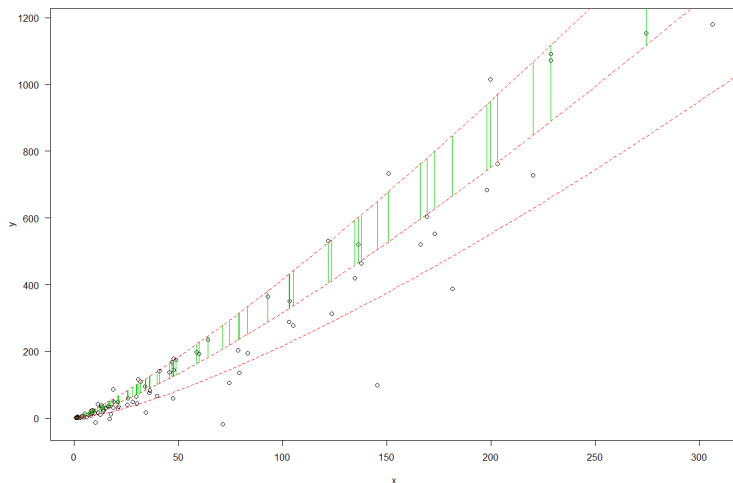
And a proxy for  $|E_k|$  is  $\tilde{E}_k = \delta_3^{1/2} x_k^{\delta_4}$ .

This suggest the strategy  $\pi_{ps} \text{---reg}$  with  $\pi_k = n \frac{x_k^{\delta_4}}{t_x^{\delta_4}}$ , which is sometimes referred as “optimal”.

# Strategy $\pi_{ps}$ —reg



# Strategy $\pi_{ps}$ —reg



# Research questions

Our hypothesis is that, as it strongly relies on the model, the strategy above is not robust.

We will compare  $\pi_{ps-reg}$  with other four strategies.

- 1 When  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, is, in fact,  $\pi_{ps-reg}$  the “best” strategy?
- 2 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?
- 3 When  $\xi_0$  does not hold, is  $\pi_{ps-reg}$  the “best” strategy?
- 4 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?

# Strategy STSI—reg

We use again  $\mathbf{x}_k = (1, x_k^{\delta_2})$  in the GREG-estimator.

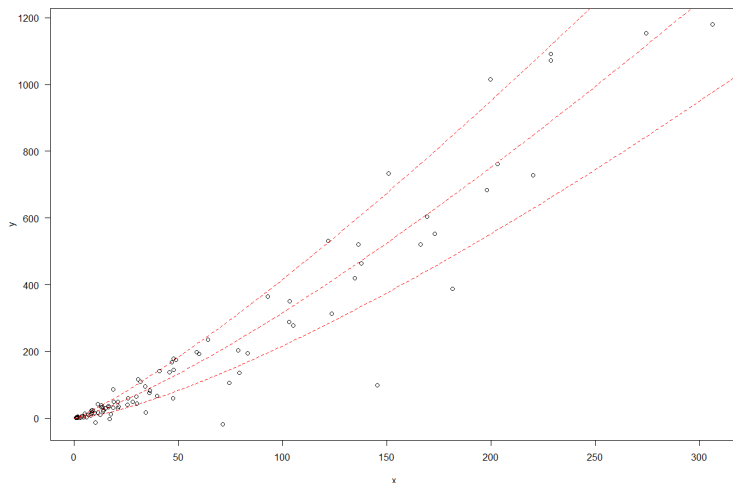
The proxies  $\tilde{E}_k \propto x_k^{\delta_4}$  are now partitioned, creating  $H$  strata. A Simple Random Sample of elements is selected in each stratum.

This strategy, STSI—reg, is often called model-based stratification.

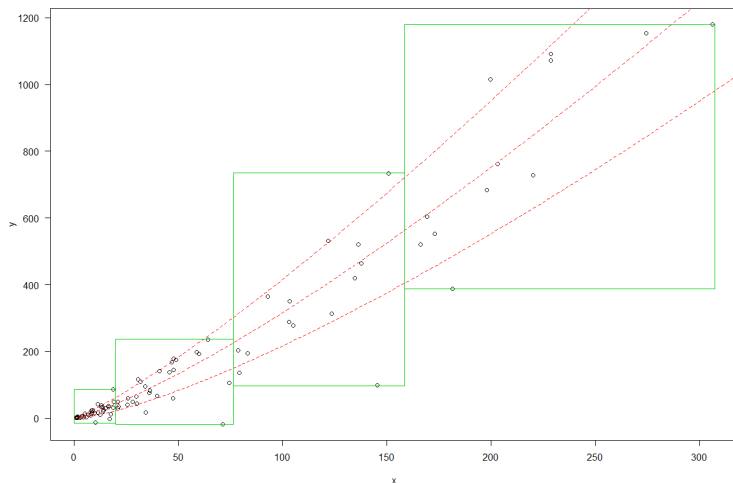
- The stratum boundaries are obtained using the cum  $\sqrt{f}$  rule on  $x_k^{\delta_4}$ ;
- The sample is allocated using Neyman allocation, i.e.

$$n_h = n \frac{N_h S_{x^{\delta_4}, U_h}}{\sum_j N_j S_{x^{\delta_4}, U_j}}$$

# Strategy STSI—reg



# Strategy STSI—reg



# Strategy STSI—HT

We use  $\mathbf{x}_k = 0$  in the GREG-estimator (i.e. the HT-estimator).

The population is stratified with respect to  $x_k^{\delta_2}$  and a Simple Random Sample of elements is selected in each stratum.

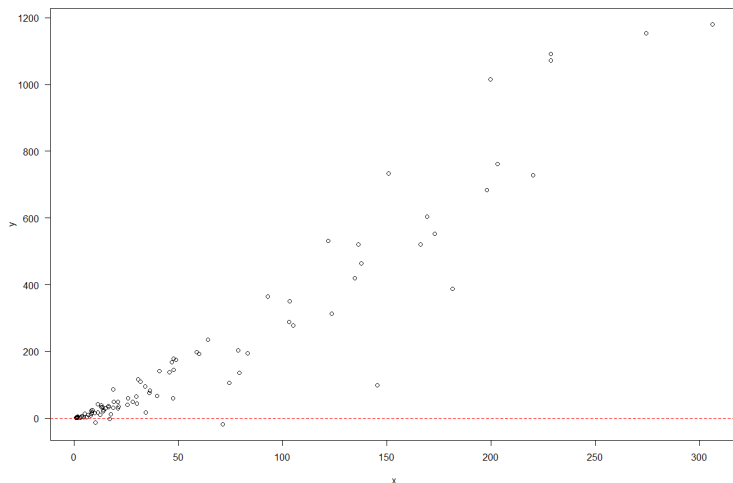
This strategy, STSI—HT, uses the auxiliary information only at the design stage. It will be considered as a benchmark.

- The stratum boundaries are obtained using the cum  $\sqrt{f}$  rule on  $x_k^{\delta_2}$ ;
- The sample is allocated using Neyman allocation, i.e.

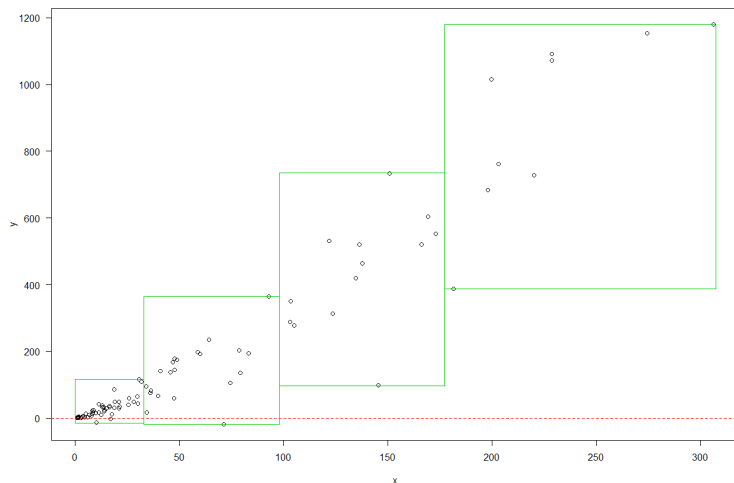
$$n_h = n \frac{N_h S_{x^{\delta_2}, U_h}}{\sum_j N_j S_{x^{\delta_2}, U_j}}$$



# Strategy STSI—HT



# Strategy STSI—HT



# Strategy $\pi$ ps—pos

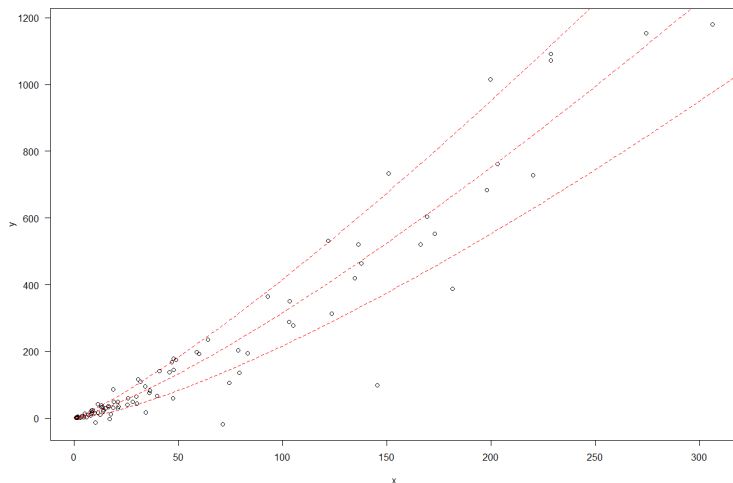
Let's assume that  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, but still we plan to use the post-stratified estimator.

As the estimator must explain the trend, the population is post-stratified with respect to  $x_k^{\delta_2}$  in the same way as in STSI—HT.

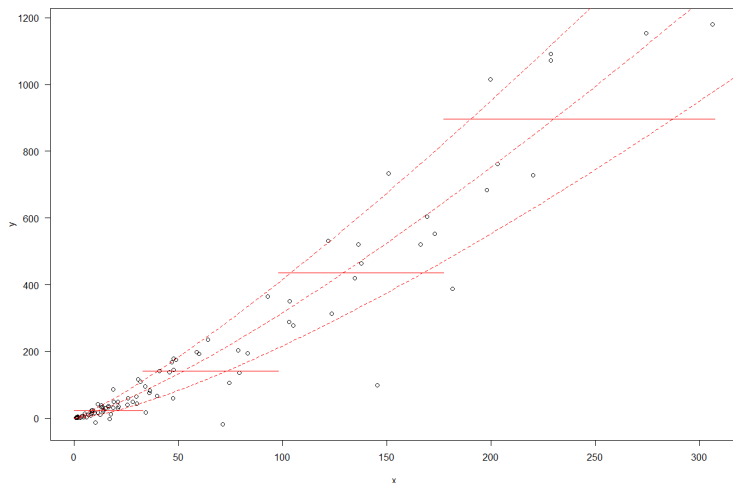
A proxy for  $|E_k|$  is  $\tilde{E}_k = \delta_3^{1/2} \sqrt{\left(1 + \frac{2}{N_j}\right) x_k^{2\delta_4} + \frac{t_{x^{2\delta_4}, U'_j}}{N_j^2}} = \delta_3^{1/2} v_k$ .

The design is a  $\pi$ ps with  $\pi_k \propto \tilde{E}_k$ .

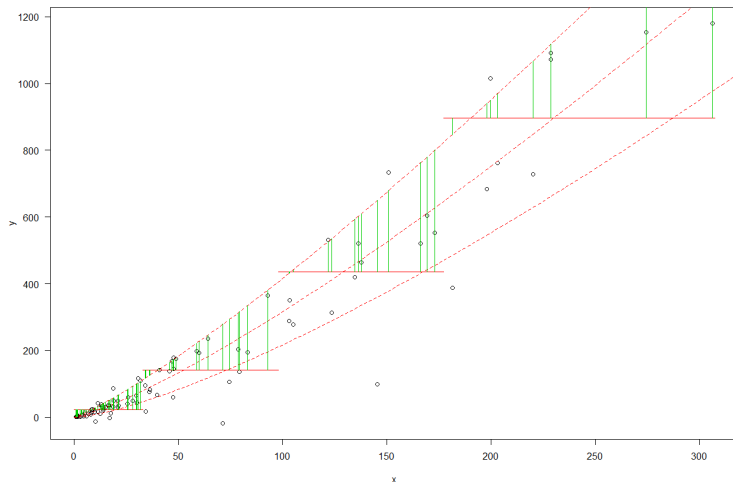
# Strategy $\pi_{ps}$ —pos



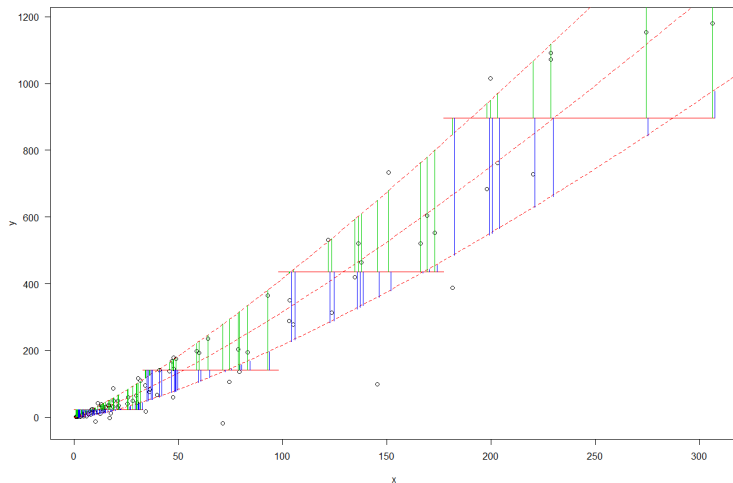
# Strategy $\pi_{ps}$ —pos



# Strategy $\pi_{ps} - \pi_{pos}$



# Strategy $\pi_{ps} - \pi_{pos}$



# Strategy STSI—pos

We decide to use the post-stratified estimator again in the same way as above.

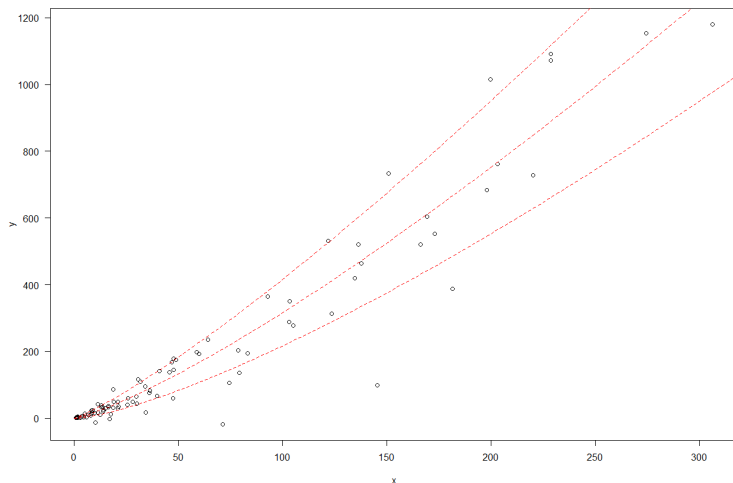
The proxies  $\tilde{E}_k \propto v_k$  are now partitioned, creating  $H$  strata. A Simple Random Sample of elements is selected in each stratum:

- The stratum boundaries are obtained using the cum  $\sqrt{f}$  rule on  $v_k$ ;
- The sample is allocated using Neymal allocation, i.e.

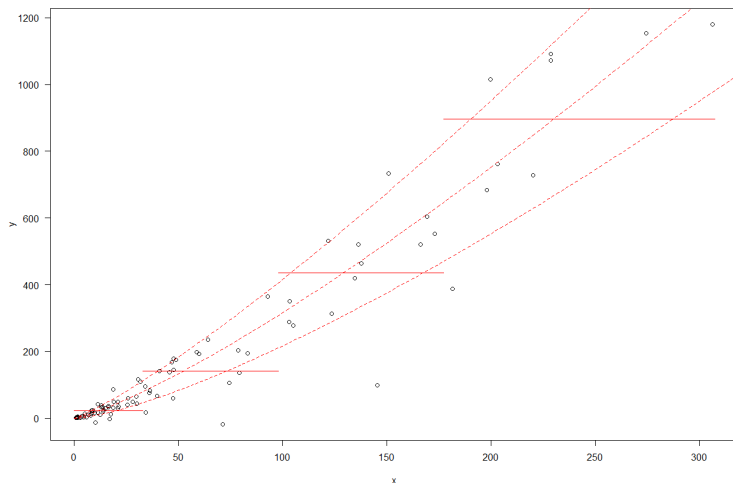
$$n_h = n \frac{N_h S_{v, U_h}}{\sum_j N_j S_{v, U_j}}$$



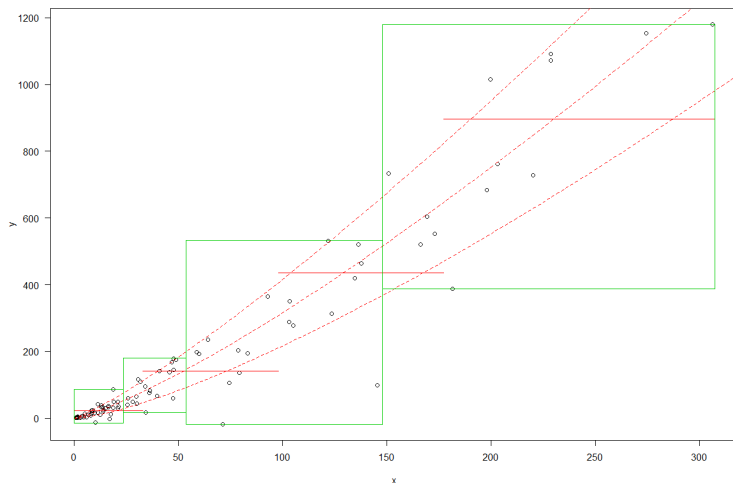
# Strategy STSI—pos



# Strategy STSI—pos



# Strategy STSI—pos



# Strategies

Design	Estimator		
	HT	Pos	Reg
STSI	1	2	4
$\pi$ ps		3	5

# Simulation study under the correct model

- A finite population of size  $N$  was generated as follows.
- The auxiliary variable,  $x$ , is obtained as  $N$  realizations from a  $\Gamma\left(\frac{4}{\gamma^2}, 12\gamma^2\right)$  plus one unit, where  $\gamma$  is the skewness.
- The study variable is generated as

$$Y_k = \delta_0 + \delta_1 x_k^{\delta_2} + \epsilon_k \quad \text{with} \quad \epsilon_k \sim N\left(0, \delta_3 x_k^{2\delta_4}\right)$$

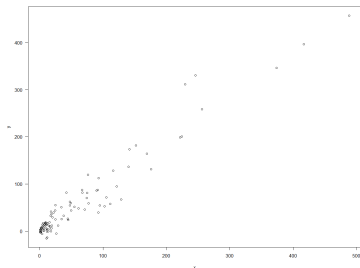
- For each strategy, the variance of sampling  $n$  elements is computed.
- The procedure is repeated  $R = 5000$  times.
- The number of strata/post-strata,  $H$ , is the same for every strategy.

# The simulation study

- $N = 5000$
- $n = 500$
- $\gamma = 3, 12$
- $H = 5$
- $\delta_0 = 0$
- $\delta_1 = 1$
- $\delta_2 = \frac{3}{4}, \frac{4}{4}, \frac{5}{4}$
- $\delta_3$  two levels in order to obtain  $\rho(X, Y) = 0.65, 0.95$
- $\delta_4 = \frac{2}{4}, \frac{3}{4}, \frac{4}{4}$

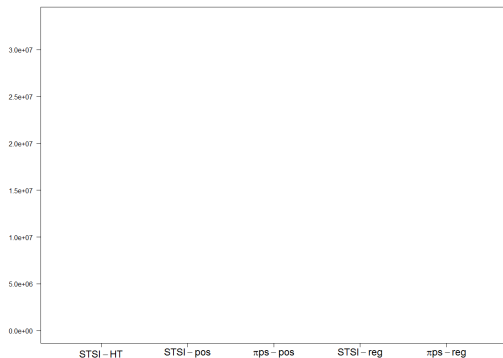
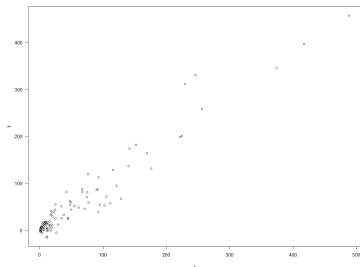
# Results

$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$



# Results

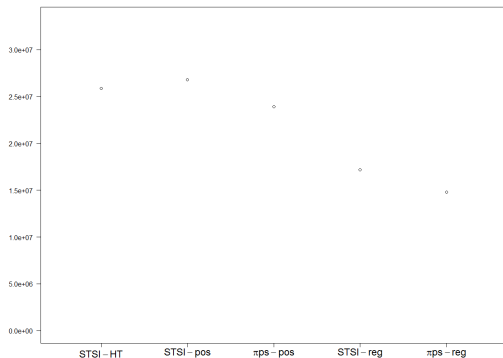
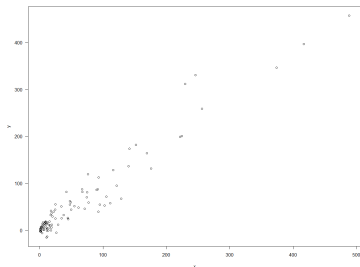
$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$





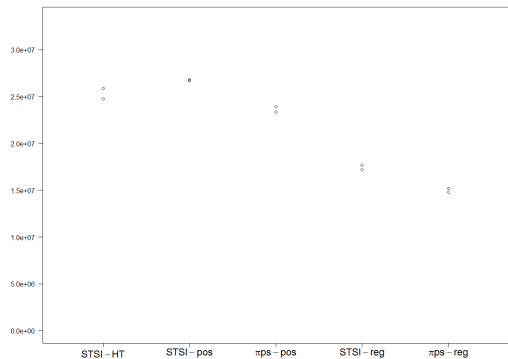
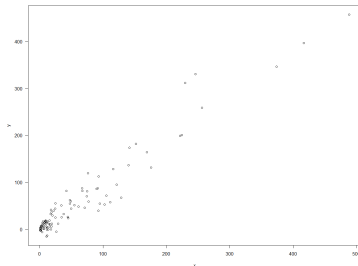
# Results

$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$



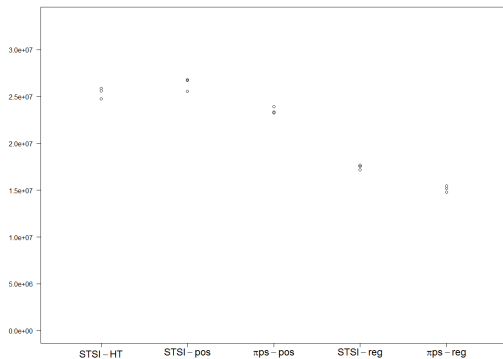
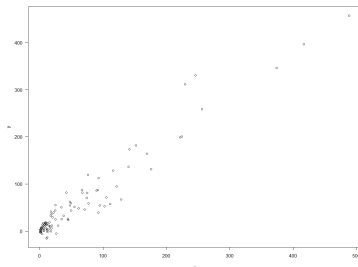
# Results

$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$



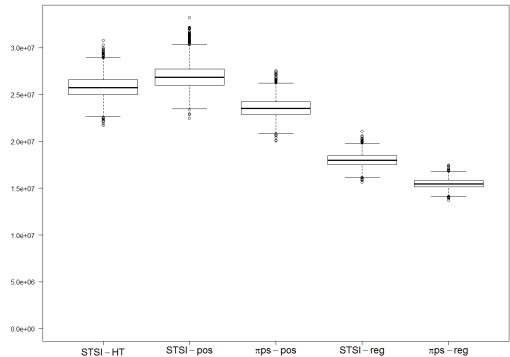
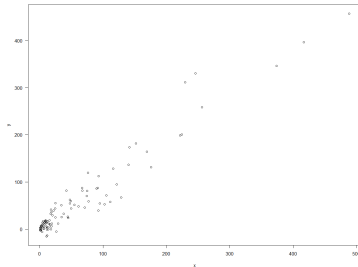
# Results

$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$



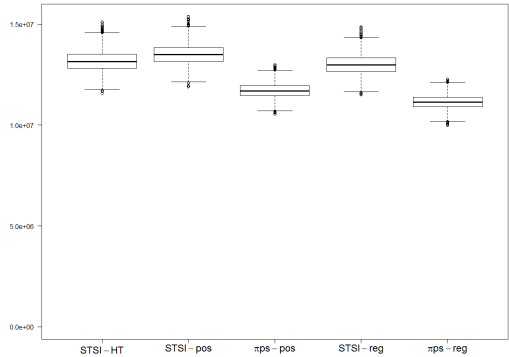
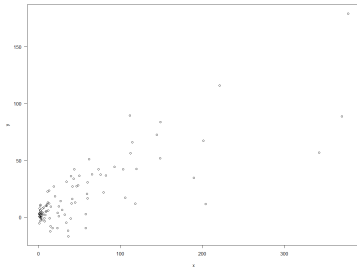
# Results

$$\gamma = 3, \delta_2 = 1, \delta_4 = 0.5, \rho = 0.95$$



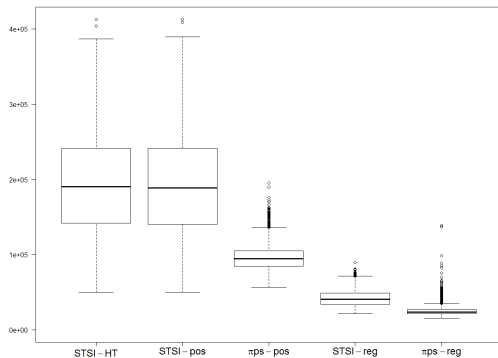
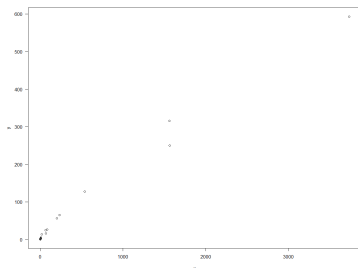
# Results

$$\gamma = 3, \delta_2 = 0.75, \delta_4 = 0.5, \rho = 0.65$$



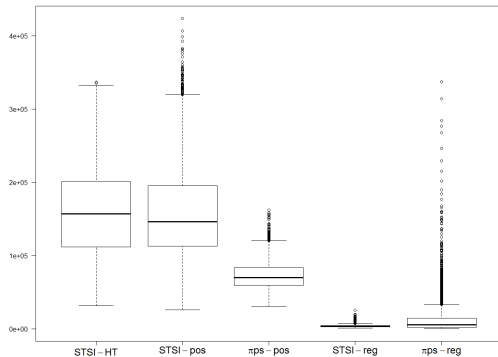
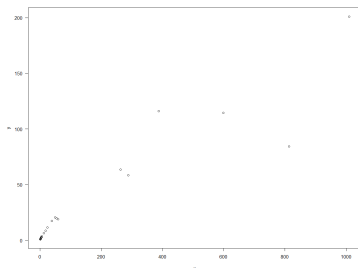
# Results

$$\gamma = 12, \delta_2 = 0.75, \delta_4 = 0.75, \rho = 0.95$$



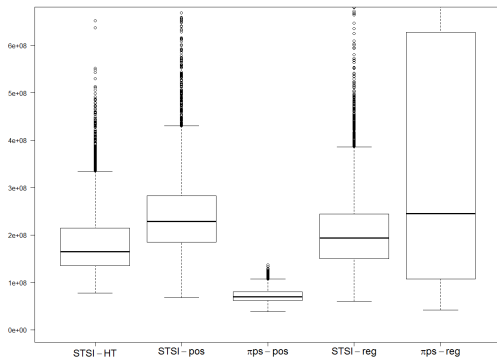
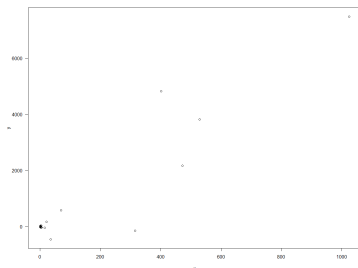
# Results

$$\gamma = 12, \delta_2 = 0.75, \delta_4 = 1.00, \rho = 0.95$$



# Results

$$\gamma = 12, \delta_2 = 1.25, \delta_4 = 1.00, \rho = 0.65$$





# Research questions

Our hypothesis is that, as it strongly relies on the model, the strategy above is not robust.

We will compare  $\pi_{ps-reg}$  with other four strategies.

- 1 When  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, is, in fact,  $\pi_{ps-reg}$  the “best” strategy?
- 2 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?
- 3 When  $\xi_0$  does not hold, is  $\pi_{ps-reg}$  the “best” strategy?
- 4 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?

# Research questions

Our hypothesis is that, as it strongly relies on the model, the strategy above is not robust.

We will compare  $\pi_{ps-reg}$  with other four strategies.

- ① When  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, is, in fact,  $\pi_{ps-reg}$  the “best” strategy? **Not always!**
- ② How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?
- ③ When  $\xi_0$  does not hold, is  $\pi_{ps-reg}$  the “best” strategy?
- ④ How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?





# Research questions

Our hypothesis is that, as it strongly relies on the model, the strategy above is not robust.





We will compare  $\pi_{ps-reg}$  with other four strategies.

- 1 When  $\xi_0$  holds and  $\delta_2$  and  $\delta_4$  are known, is, in fact,  $\pi_{ps-reg}$  the “best” strategy? **Not always!**
- 2 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?
- 3 When  $\xi_0$  does not hold, is  $\pi_{ps-reg}$  the “best” strategy? **Not always!**
- 4 How does  $\pi_{ps-reg}$  behave with respect to other strategies in terms of finite population characteristics?





# Bibliography I

-  Brewer, K.R.W. (1963). *A Model of Systematic Sampling with Unequal Probabilities*. Australian Journal of Statistics, **5**, 5-13.
-  Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. London: Arnold.
-  Cassel, C.M., Särndal, C. E. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley.
-  Dalenius, T. and Hodges, J.L. (1959) *Minimum variance stratification*. Journal of the American Statistical Association, **54**, 88-101.





# Bibliography II

-  Godambe, V.P. (1955). *A unified theory of sampling from finite populations*. Journal of the Royal Statistical Society, Series B **17**, 269-278.
-  Hanif, M. and Brewer K. R. W. (1980). *Sampling with Unequal Probabilities without Replacement: A Review*. International Statistical Review **48**, 317-335.
-  Holmberg, A. and Swensson, B. (2001). *On Pareto  $\pi ps$  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.

# Bibliography III

-  Kozak, M. and Wieczorkowski, R. (2005).  *$\pi$ ps Sampling versus Stratified Sampling ? Comparison of Efficiency in Agricultural Surveys*. Statistics in Transition, **7**, 5-12.
-  Lanke, J. (1973). *On UMV-estimators in Survey Sampling*. Metrika **20**, 196 202.
-  Rosén, B. (1997). *On sampling with probability proportional to size*. Journal of statistical planning and inference **62**, 159-191.
-  Rosén, B. (2000a). *Generalized Regression Estimation and Pareto  $\pi$ ps*. R&D Report 2000:5. Statistics Sweden.

# Bibliography IV

-  Rosén, B. (2000b). *On inclusion probabilities for order  $\pi ps$  sampling*. Journal of statistical planning and inference **90**, 117-143.
-  Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
-  Tillé, Y. (2006). *Sampling algorithms*. 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@stat.su.se